Simulating Many-body Quantum Dynamics on Classical and Quantum Computers

by

Dong An

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Applied Mathematics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Lin Lin, Chair
Professor Jon Wilkening
Professor Anil Aswani

Summer 2021

Simulating Many-body Quantum Dynamics on Classical and Quantum Computers

Abstract

Simulating Many-body Quantum Dynamics on Classical and Quantum Computers

by

Dong An

Doctor of Philosophy in Applied Mathematics

University of California, Berkeley

Professor Lin Lin, Chair

This dissertation concerns the numerical simulation of many-body quantum dynamics, which is fundamental for predicting physical and chemical properties at the atomic and sub-atomic scale. The problem appears ubiquitously in many areas, such as quantum chemistry, quantum controls, and quantum information theory. Simulating many-body quantum dynamics poses a variety of computational challenges, including high dimensionality and fast oscillations. While model reduction techniques can partially resolve the high dimensionality issue on classical computers, quantum computers give rise to new hopes to directly simulate the full many-body quantum dynamics. Nevertheless, both classical and quantum simulations still suffer from highly oscillatory solutions, limiting the time step sizes in time discretization and hindering the practical applications of quantum dynamics simulation. The broad goal of this dissertation is to investigate how classical and quantum computers can efficiently treat fast oscillatory solutions. As a notable application, such progress leads to new methods for solving linear system problems on quantum computers. This dissertation consists of three parts: adiabatic dynamics (Part II), classical simulation (Part III) and quantum simulation (Part IV).

Although the quantum dynamics are generally complicated, when the Hamiltonian varies slowly with time and satisfies certain spectrum gap conditions, the solution can approximately remain within some specific eigenspace of the Hamiltonian. This phenomenon is called the near adiabatic evolution, which has attracted much attention since the early days of quantum mechanics. It weaves together eigenvalue problems and differential equations. Adiabatic dynamics is also one of the underlying physical principles for building universal quantum computational devices. The near adiabatic evolution serves as a glue, explicitly and implicitly, throughout this dissertation. In Part II, we quantitatively study the adiabatic

error between the quantum dynamics and the exact eigenspace by proving a new version of the quantum linear adiabatic theorem. Under the gap condition and the vanishing boundary condition, we show that the adiabatic error can converge exponentially in terms of the inverse evolution time. Meanwhile, to control the adiabatic error at the desired level, the evolution time is sufficient to scale almost quadratically in terms of the magnitude of the inverse spectrum gap. This result is almost sharp in both the convergence order and the gap dependence, and appears for the first time beyond the two-level system.

Part III is devoted to designing a new approach to efficiently deal with highly oscillatory solutions of quantum dynamics on classical computers. The critical observation is that such fast oscillations in the wave functions are not physical and are solely due to the generally non-optimal gauge choice (*i.e.* degrees of freedom irrelevant to physical observables) of the Schrödinger equation. The optimal gauge choice is given by a parallel transport formulation, which can significantly flatten the wave functions and thus allow much larger time step sizes in time discretization. We establish the framework of the parallel transport dynamics for evolutions of pure states and mixed states, as well as the time-dependent density functional theory.

We start with the simplest single pure state evolution and derive the dynamics under the parallel transport gauge via two approaches: solving the optimization problem and evolving the dynamics under the parallel transport operator. We analyze the resulting parallel transport dynamics in the context of the singularly perturbed linear Schrödinger equation and demonstrate its superior performance in the near adiabatic regime. Then we derive the dynamics under parallel transport gauge for real-time time-dependent density functional theory and numerically test its performance using absorption spectrum, ultrashort laser pulse, and Ehrenfest dynamics calculations as examples. Our tests show that propagating parallel transport dynamics is more than 10 times faster in terms of the wall clock time when compared to the standard explicit fourth-order Runge-Kutta time integrator for the original Schrödinger equation. Finally, we generalize the parallel transport dynamics to the scenario of mixed state evolution. Going beyond the linear and near adiabatic regime, we find that the error of the parallel transport dynamics can be bounded by certain commutators between Hamiltonians, density matrices, and their derived quantities. Such a commutator structure is not present in the Schrödinger dynamics. The commutator structure of the error bound and numerical results in the nonlinear regimes further confirm the advantage of the parallel transport dynamics.

Part IV is about simulating linear quantum dynamics on quantum computers, as well as application to solving quantum linear system problems. For quantum simulation, we focus on the standard and generalized Trotter methods and study their performance on simulating unbounded time-dependent control Hamiltonian, where the cost of the simulation cannot

be well bounded by existing theoretical analysis for most quantum algorithms. We observe that nearly all existing analyses on quantum simulation focus on the difference between the exact evolution operator and the numerical evolution operator. This measures the worst-case error of the quantum simulation, which might not be of practical interest. By proving a new vector norm error bounds for the Trotter type methods, we demonstrate that if the quantum dynamics are smooth enough, the cost of quantum simulation using the Trotter type methods does not increase as the Hamiltonian norm increases. Our result extends that of [Jahnke, Lubich, BIT Numer. Math. 2000] to the time-dependent setting and outperforms all previous analyses in the quantum simulation literature for simulating unbounded time-dependent Hamiltonian. We also clarify the existence and the importance of commutator scalings of Trotter and generalized Trotter methods for time-dependent Hamiltonian simulations.

Linear system solvers are used ubiquitously in scientific computing. Quantum algorithms for solving large systems of linear equations have received much attention recently, but most existing algorithms either do not have optimal asymptotic complexity scalings or involve rather complicated subroutines. We study how simulating quantum dynamics and adiabatic theorem can be combined to construct a new near-optimal quantum linear system solver. Our approach first transforms the linear system problem to an eigenvalue problem, then constructs a near adiabatic dynamics with the final solution solving this eigenvalue problem, finally simulating this near adiabatic dynamics by existing quantum simulation algorithms. We demonstrate that with an optimally tuned scheduling function, the new adiabatic-based solver can readily solve a quantum linear system problem with $\mathcal{O}(\kappa \text{ poly}(\log(\kappa/\epsilon)))$ runtime, where κ is the condition number, and ϵ is the target accuracy. This is near-optimal in both κ and ϵ. The complexity estimate of the adiabatic-based solver is derived from an improved adiabatic theorem in which the constant and gap dependence are carefully and explicitly tracked. We also investigate the possibility of solving quantum linear system problems using the related quantum approximate optimization algorithm with an optimal control protocol.

Acknowledgments

I would first like to express my deepest appreciation to my advisor Lin Lin. I have always been learning from his extensive knowledge and ingenious suggestions, and his guidance and encouragement cannot be overestimated. Furthermore, he introduced me to the exciting research areas of adiabatic theorems, quantum computing, and quantum algorithms, which form an essential part of this dissertation and I regard as one of my research directions in my future academic career. None of the work in this dissertation could have happened without Lin's help.

For insightful discussions and enjoyable collaborations, I want to extend my sincere thanks to my collaborators: Sara Cheng, Di Fang, Teresa Head-Gordon, Weile Jia, Itai Leven, Lin Lin, Noah Linden, Michael Lindsey, Jin-Peng Liu, Jianfeng Lu, Ashley Montanaro, Changpeng Shao, Songchen Tan, Yu Tong, Jiasu Wang, Lin-Wang Wang, Nathan Wiebe, and Ze Xu, with whom I feel very honored and fortunate to have the opportunity to work.

Among those who have offered me unwavering support and encouragement during my graduate study, I would like to thank Per-Olof Persson, Daniel Tataru, and Jon Wilkening, from whom I learned a lot of valuable knowledge at the starting of my research. Many thanks to Ryan Babbush and Jarrod Mcclean for the warm host of my summer research internship at Google, and to Dominic Berry, Pedro Costa, Yuval Sanders, Yuan Su, and Nathan Wiebe for patient explanations and engaging discussions on fascinating projects on quantum algorithms.

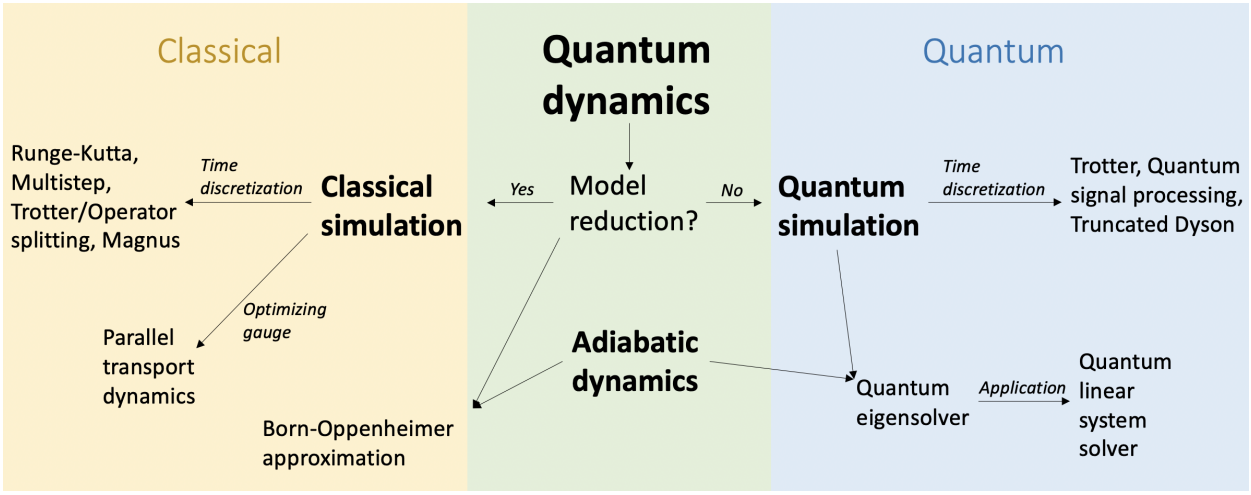
I am also grateful to the members of my qualification exam and dissertation committee: Anil Aswani, Lin Lin, John Strain, and Jon Wilkening, for the time and effort in holding my qualification exam, and the reading and assistance on this dissertation.

Finally, I wish to express my special thanks to my beloved parents, my mother Mingfen Wei and my father Shengbing An, for their endless love and support, and my fiancée Bing Li for her passion and accompanying in every single day.

Contents

Part I

Introduction



A map of ideas and concepts, to be further interpreted and expanded in the introduction and throughout this dissertation. The yellow and blue regions correspond to classical and quantum simulations, respectively.

This dissertation focuses on simulating quantum dynamics, *i.e.* numerically solving the time-dependent Schrödinger equation

$$\mathrm{i}\partial_t \psi(t) = H(t, \psi(t))\psi(t), \quad t \in \mathbb{R}.$$

Here t is the time variable, i is the imaginary unit, the complex-valued vector $\psi(t)$ is called a *wave function*, and the *Hamiltonian* $H(t, \psi)$ is a finite-dimensional Hermitian matrix for all t, ψ which describes a discrete quantum system or a discretized quantum system originating from a continuous system. The time-dependent Schrödinger equation is a fundamental model to describe physical phenomenon at the atomic and sub-atomic scale and appears ubiquitously in the context of quantum chemistry, quantum controls, and quantum information science, to name a few. The solution of the time-dependent Schrödinger equation is in general complicated, but when the Hamiltonian varies slowly with time and satisfies certain spectrum gap conditions, the solution can be simpler in the sense that it approximately remains in the eigenspace of the Hamiltonian. This phenomenon is called the near adiabatic evolution and serves as the second glue, besides simulating quantum dynamics, throughout this dissertation.

Simulating many-body quantum dynamics faces several computational challenges despite its significant importance, including high dimensionality, multiple scales, and highly oscillatory solution. High dimensionality is the most severe challenge in simulating many-body quantum dynamics. We take the transverse field Ising model (*i.e.* quantum version of the classical Ising model) with n sites [142] as the first example. In this model, the state of a

single site can be described using a vector in a two-dimensional complex Hilbert space, and the state space of the n-site system is the tensor product of n multiple two-dimensional complex Hilbert spaces, which is of dimension 2^n. This rapidly becomes prohibitively expensive as the number of the sites n increases. Another important example is the many-body quantum dynamics for molecular systems. In this example, the Hamiltonian is the sum of the kinetic operator of the nuclei and the electrons, the interaction operator between each pair of the particles, and the possible external force field operator. In particular, for a many-body molecular system consisting of n_a nuclei and n_e electrons, the Hamiltonian[1] can be written as [111]

$$H = -\sum_{j=1}^{n_a} \frac{1}{2M_j}\Delta_{R_j} - \sum_{j=1}^{n_e} \frac{1}{2m}\Delta_{r_j} + V_I(r, R).$$

Here R_j's and r_j's are three-dimensional vectors representing the position coordinates of the corresponding nuclei and electrons, respectively. M_j's and m are the mass of nuclei and electrons, respectively. $V_I(r, R)$ represents the interaction which can be written in the general form as

$$V_I(r, R) = \sum_{1\le j<k\le n_a} v_{NN}(R_j, R_k) + \sum_{1\le j<k\le n_e} v_{ee}(r_j, r_k) + \sum_{j=1}^{n_a}\sum_{k=1}^{n_e} v_{Ne}(R_j, r_k),$$

where v's are the interaction operators between two particles, such as the Coulomb potential $v(x, y) \sim 1/|x - y|$. In the quantum many-body molecular system, the spatial variable is the set of all R_j's and r_j's, and thus in the dimension of $(3n_a + 3n_e)$. After spatial discretization, the dimension of the discretized Schrödinger equation scales exponentially in $(3n_a + 3n_e)$, in which the base depends on the degree of freedom in the spatial discretization along each dimension.

The treatment of many-body quantum dynamics on classical and quantum devices diverges from here. Due to the exponentially large cost, directly simulating many-body quantum dynamics via standard discretization is generally intractable on classical computers nowadays, even if the simulation is only for a small molecule[2]. To obtain yet reasonable simulation results on classical computers, *model reduction* before the simulation is unavoidable. The model reduction reduces the degree of freedom of the full quantum models to a classically amenable scale. However, it introduces extra systematic approximation errors, and the practical applications of the reduced models are limited. On the other hand, quantum computers are based on the law of quantum mechanics and can potentially handle specific

[1]For simplicity, here the Hamiltonian is before spatial discretization.

[2]For example, for a single water molecule, there are 3 nuclei and 10 electrons. Even if only 8 spatial grids are used along each coordinate, the size of the corresponding discretized Hamiltonian becomes as large as 8^{39}.

exponentially large problems with only polynomial cost, including simulating quantum dynamics [59]. As a result, model reduction is no longer necessary for quantum simulation, and thus the range of applications of quantum simulation can potentially be much broader than that of classical simulation.

Classical simulation

The first step is to perform a model reduction technique to overcome the high dimensionality of simulating quantum dynamics on classical computers. This model reduction is based on physical insight and mathematical analysis and typically results in intermediate models between classical and quantum dynamics. For many-body systems, there exist numerous approaches for model reduction, such as Born-Oppenheimer approximation [23], time-dependent Hartree-Fock method [51], multi-configuration method [117], dynamical low-rank approximation [97], time-dependent density functional theory (TDDFT) [133], or a combination of several approaches[3]. Here we only briefly present the Born-Oppenheimer approximation and TDDFT, on which we will design and test efficient time propagation algorithms later in this dissertation.

The Born-Oppenheimer approximation makes two key assumptions. The first assumption is the separation of the nuclear dynamics and the electronic dynamics, in the sense that the electronic Hamiltonian with fixed nuclei position is first simulated, then the nuclear dynamics fed back from the electronic wave function is considered. The reason behind such a separation is that the nuclear mass is much larger than the electronic mass, and thus the electrons move much faster than the nuclei. However, the size of the electronic Hamiltonian with fixed nuclei still depends exponentially on the number of electrons, and further simplification is required. The second assumption is that the electrons are stable and slaved in the low energy levels instead of moving dynamically. This means that the electronic wave functions are (at least approximately) the eigenvector of the electronic Hamiltonian corresponding to the smallest eigenvalue, where the resulting eigenvalue problems can be further solved via relatively better studied time-independent model reduction techniques [27]. The second assumption can be reasonable when the system is not influenced by any external force field, and there exists a spectrum gap in the electronic Hamiltonian. This is indeed related to the adiabatic evolution and will be discussed later in the introduction.

The mathematical formalism of the Born-Oppenheimer approximation proceeds as follows. By fixing the nuclei positions, the electronic Hamiltonian becomes

$$H_e(R) = -\sum_{j=1}^{n_e} \frac{1}{2m}\Delta_{r_j} + V_I(\cdot, R),$$

[3]We refer interested readers to [111] for a comprehensive review.

and the electronic wave functions solve the eigenvalue problem

$$H_e(R)\psi(\cdot, R) = E(R)\psi(\cdot, R).$$

After we plug this back into the original full Schrödinger equation, the nuclear dynamics is given by another Schrödinger equation governed by the Born-Oppenheimer Hamiltonian

$$H_{\mathrm{BO}} = -\sum_{j=1}^{n_a} \frac{1}{2M_j}\Delta_{R_j} + E(R).$$

Therefore the complexity of simulating Born-Oppenheimer molecular dynamics depends on the complexity of two sub-problems: the time-independent eigenvalue problem, which can be satisfactorily solved via time-independent model reduction techniques, and the simulation of H_{BO}, in which the degree of freedom only depends on the number of nuclei.

Despite its great success and wide application, the Born-Oppenheimer approximation does not take into consideration electron excitation during the dynamics, which cannot be ignored in the ultrafast interacting systems. TDDFT is an alternate theory to model this scenario satisfactorily. In TDDFT, the electronic dynamics is involved through the electron density in far less dimension than the original set of particle coordinates. Without further investigation of establishing TDDFT[4], the many-body quantum system is described by a set of three-dimensional wave functions $\Psi(t) = \{\psi_j(t)\}_{j=1}^{n_e}$ (which are also called electron orbitals), and the TDDFT Hamiltonian takes the form[5]

$$H_e = -\frac{1}{2m}\Delta + V(R, \rho(t)),$$

where $\rho(t) = \sum |\psi_j(t)|^2$ denotes the electron density. Since the electrons move much faster than the nuclei, the corresponding nuclear dynamics can be approximated by averaging over the electronic wave functions with fixed nuclei position and taking the classical limit, leading to the Ehrenfest dynamics [103, 111] as

$$M_j \frac{d^2}{dt^2} R_j(t) = -\Psi^*(t)(\nabla_{R_j} H_e)\Psi(t).$$

The Ehrenfest dynamics is a set of Newton-like equations. The complexity of simulating TDDFT is thereby dominated by the degree of freedom in the electron density, which is less relevant to the number of the electrons.

[4]The rigorous foundation of TDDFT is much less clear than Born-Oppenheimer approximation, and we do not provide the original physical argument either.

[5]Here we omit the external force field.

After model reduction, simulating many-body quantum dynamics is reduced to solving another time-dependent Schrödinger equation with feasible system size. This can readily be numerically solved using standard time propagators for ordinary differential equations [69], such as explicit Runge-Kutta methods [137] and implicit Runge-Kutta methods [36]. Another commonly used class of time propagators is the class of operator splitting methods [12, 112], especially when the Hamiltonian can be decomposed as $H = A+B$ where the Hamiltonians A and B are easy to simulate separately. An example is that A and B are the Laplacian operator and the potential operator, diagonalizable under different but efficiently transformable sets of basis. The first order splitting method is based on the Lie-Trotter formula

$$\exp\left(-\mathrm{i}Ht\right) \approx \left(\exp\left(-\mathrm{i}Bt/m\right)\exp\left(-\mathrm{i}At/m\right)\right)^m,$$

and hereby the operator splitting methods are also referred to as Trotter formulae, especially in physics context. Besides standard Runge-Kutta methods and operator splitting methods, a wide range of other numerical discretization methods have also been thoroughly studied, such as Magnus expansion methods [36, 39], exponential time differencing methods [92], spectral deferred correction methods [84], dynamical low rank approximation [97], adiabatic state expansion [80, 152], to name a few.

Despite various model reduction approaches to overcome the challenge of high dimensionality and comprehensive studies on numerical discretization methods, simulating quantum dynamics still suffers from the highly oscillatory solution. As an illustration, we consider a trivial Schrödinger equation where the Hamiltonian $H = \lambda$ is just a constant scalar. The solution of this trivial example can be explicitly written as

$$\psi(t) = e^{-\mathrm{i}\lambda t}\psi(0), \tag{0.0.1}$$

which oscillates on the time scale $\sim 1/\lambda$. In general, the possibly fastest component of the wave function oscillates on the scale of $1/\|H\|$[6], and $\|H\|$ can potentially be very large when, for instance, part of H comes from the spatial discretization of the unbounded Laplacian operator. To accurately resolve such fast oscillations, the time step sizes required in a broad subset of numerical discretization methods (including Runge-Kutta methods) are required to be small enough such that $\Delta t\|H\| \lesssim 1$, which is also the stable condition for widely used explicit Runge-Kutta time propagators. Unfortunately, the small time step sizes hinder the practical applications of quantum dynamics simulation, especially up to a relatively long time.

In Part III of this dissertation, we propose an efficient approach of simulating fast oscillatory dynamics by deriving a new representation of the time-dependent Schrödinger equation which allows much larger time step sizes in numerical propagators. The key of the improvement is based on the observation that most oscillations in the wave functions are indeed not

[6]Here $\|\cdot\|$ denotes the matrix 2-norm.

physically intrinsic. Specifically, quantities of physical interest are called the observables. An observable can be defined as $\mathrm{Tr}(OP)$ where O is a Hermitian matrix and $P = \psi\psi^*$ is called the density matrix. Therefore, the observable depends on the density matrix, and the oscillations in the density matrix can be much slower than those in the wave function. Retake the trivial example where $H = \lambda$ is a constant scalar, then the wave function oscillates on the time scale $\sim 1/\lambda$, but the density matrix is just a constant matrix. Such a gap between the wave function and the density matrix inspires us to construct another wave function ϕ which forms the same density matrix as the original Schrödinger wave function ψ but oscillates on a much larger time scale. To reconstruct the same density matrix, the newly transformed wave function $\phi(t)$ should satisfy $\phi(t) = \psi(t)U(t)$ for a unitary matrix $U(t)$, called gauge matrix. Optimizing $U(t)$ such that the oscillation within the new wave function $\phi(t)$ is minimized gives rise to the *parallel transport gauge*. Under the parallel transport gauge, another form of the Schrödinger equation can be derived for the transformed wave functions, which is driven by the residue and only adds one additional term to the original Schrödinger equation. Therefore the Schrödinger equation under the parallel transport gauge can readily be solved by any existing numerical propagators. Due to the minimized oscillation in the parallel transport wave function, much larger time step sizes are allowed in numerical propagators, and the simulation can be performed more efficiently.

Quantum simulation

Roughly speaking, quantum computes are computational devices that explicitly take advantage of the law of quantum mechanics. Due to different underlying principles, quantum computers can perform computations differently from classical computers. Several classes of operations can be performed much more efficiently on quantum computers than on classical computers, and vice versa. For example, on quantum computers, computing a generic matrix-vector multiplication is very hard, and copying an unknown state is explicitly forbidden (called the no-cloning theorem [156]). However, it is very efficient to perform certain unitary transformations[7]. This different architecture gives rise to the possibility for quantum algorithms to achieve speedups over classical algorithms for certain problems, such as factoring [140] and unstructured search [66].

In some applications, quantum algorithms can even achieve an exponential speedup over classical algorithms, in the sense that the complexity of the best existing classical algorithms scale exponentially in some parameters, but there exists a quantum algorithm that can solve the same problem or a quantum analog in polynomial cost. An overly simplified mathematical intuition behind such a possible exponential speedup is as follows. Take as an

[7]Indeed, quantum computers can only perform linear unitary transformations since quantum states are described by normalized vectors, among which the generic transformations are unitary.

example the quantum circuit model [124], which can be regarded as a quantum analog of the classical logic gate based computational model and is one of the most favorite quantum computing models nowadays. Unlike the classical bit, which can only take discrete values 0 or 1, a *qubit* (*i.e.* quantum bit) generally carries the information of the superposition of 0 and 1. Mathematically, let $|0\rangle$ and $|1\rangle$ denote two orthonormal vectors, then a qubit can be represented as $\alpha|0\rangle + \beta|1\rangle$ for complex numbers α, β such that $|\alpha|^2 + |\beta|^2 = 1$. Therefore, a qubit can be regarded as an element (α, β) in the Hilbert space $\mathbb{C}^2$ (module the norm), and thus carries much more information than a classical bit. Furthermore, if we consider an n-qubit system, the corresponding state space becomes the tensor product of n multiple $\mathbb{C}^2$, of which the dimension scales $\sim 2^n$. In other words, quantum computers can "store" a vector in a 2^n-dimensional linear space with only $\mathcal{O}(n)$ costs. Furthermore, linear operations on this 2^n-dimensional space can also be constructed via sequential simple operations on all or part of the n qubits. This implies that generic linear unitary transformations on a 2^n-dimensional space can also be efficiently performed on quantum computers with polynomial cost in n.

The challenge of high dimensionality in simulating quantum dynamics is naturally resolved on quantum computers by such an exponential speedup, and no model reduction technique is necessary for quantum simulation. This makes simulating quantum dynamics on quantum computers very attractive. The past few years have witnessed significant progress in the development of new quantum algorithms and the improvement of theoretical error bounds of existing quantum algorithms. Since numerous complicated Hamiltonians of practical interest can be decomposed as the sum of easily simulated Hamiltonians, Trotter methods become applicable and privilege for quantum simulation [78, 155, 42, 44].[8] To further improve the precision of the simulation, many post-Trotter algorithms have been proposed and analyzed, such as, for a time-independent Hamiltonian H, linear combination of unitaries [15], truncated Taylor series [16], quantum signal processing [109], quantum singular value transform [64], multi-product formula [110], and randomization product formula [30, 41, 38]; and for a time-dependent Hamiltonian $H(t)$, truncated Dyson series [19, 16, 108], and rescaled Dyson series [18]. We remark that most existing studies only consider simulating linear Schrödinger equations since the power of quantum computers to perform nonlinear mappings is believed to be severely limited [45].

Despite naturally overcoming the challenge of high dimensionality thanks to the architecture of quantum computers, quantum simulation still suffers from fast oscillatory solutions. Such an issue is displayed as the explicit linear dependence on $\|H\|$ in the complexity of quantum simulation algorithms. Specifically, even the best existing quantum algorithms take the cost $\mathcal{O}(\|H\|)$ to simulate quantum dynamics in the worst case, and the so-called

[8] Runge-Kutta and multistep methods seem not preferable in quantum simulation because the corresponding propagators are not unitary, though there exist efforts on building a quantum version of the multistep methods for solving generic ordinary differential equations [13].

"no-fast-forwarding" theorem [14, 17] demands at least a linear dependence in the norm of H for generic quantum simulation algorithms. This can be expensive when the norm of the Hamiltonian is huge, for example, when the Hamiltonian contains the discretized Laplacian operator. In this dissertation, we focus on the problem of simulating time-dependent Hamiltonian with large spectrum norm. While the parallel transport dynamics in Part III seems not suitable for quantum simulation due to its unavoidable strong nonlinearity, we instead study whether existing quantum algorithms can perform better and allow tighter theoretical complexity estimate for the dynamics which does not oscillate too fast. In Chapter 5, we answer this question positively for the simplest Trotter methods by deriving improved error bounds in vector norm scalings, which can explore the information of the initial vectors as well as the possibly better regularity in the wave functions.

Near adiabatic dynamics

The solution of the time-dependent Schrödinger equation is in general complicated. For example, it possibly involves the superposition of several quantum states at different energy levels as well as transitions among those during the evolution. However, when the Hamiltonian varies slowly with time and satisfies certain spectrum gap condition, the solution becomes relatively simple because it can approximately remain in the instantaneous eigenspace of the Hamiltonian. Such a phenomenon is called the near adiabatic evolution, which can be dated back to Born and Fock [22] and has attracted much attention since then [93, 120, 121, 9, 82, 53, 2, 57].

Mathematically, consider the singularly perturbed Schrödinger equation

$$\mathrm{i}\frac{1}{T}\partial_t\psi(t) = H(t,\psi(t))\psi(t), \quad 0 \le t \le 1.$$

Here T represents the physical time, and t is the rescaled dimensionless time, then increasing T is equivalent to slowing down the Hamiltonian over the entire physical evolution. The spectrum gap condition[9] says that there exists a continuous function $\lambda(t,v)$ such that $\lambda(t,v)$ is an eigenvalue of the Hamiltonian $H(t,v)$ and separated from the rest of the spectrum uniformly by a positive constant (called gap). The near adiabatic dynamics is that, if the dynamics starts from an eigenvector of $H(0,\psi(0))$ corresponding to the eigenvalue $\lambda(0,\psi(0))$, then the solution $\psi(t)$ will approximately remains in the eigenspace of $H(t,\psi(t))$ corresponding to the eigenvalue $\lambda(t,\psi(t))$.

Theoretical results on rigorously bounding the difference between the exact quantum dynamics and the eigenspace are usually referred to as the adiabatic theorems. Adiabatic

[9]For simplicity, here we only consider the gap condition for a single eigenvalue. Gap condition can also be generalized to the scenario of multiple eigenvalues, which will be specified later in Part II.

theorems also focus on how the physical evolution time T and the magnitude of the gap affect the approximation error. Most existing adiabatic theorems treat the linear regime where H is independent of ψ. The arguably first rigorous adiabatic theorem is by Kato [93], which was further extended and improved to have better time convergence [89, 82, 121, 62] and explicit gap dependence [88, 53], to the gapless scenario [9], to the discrete evolution [52], and to the extended many-body system [10]. In Part II of this dissertation, we prove a new version of linear adiabatic theorem with near-optimal dependence in both the evolution time T and the size of the spectrum gap. Compared to the linear regime, there are significantly fewer nonlinear adiabatic theorems available, partially due to the limited nonlinear spectrum theory. [35, 72, 141] prove the adiabatic theorem with weak nonlinearity, and [61] studies the small initial condition scenario, which is equivalent to the weak nonlinear formalism for the example of Gross-Pitaevskii equation. A remarkable nonlinear adiabatic theorem is obtained in [57] beyond the weak nonlinear and small initial condition regime.

We remark that near adiabatic dynamics serve as the second glue of this dissertation besides simulating quantum dynamics. Theoretical derivation and analysis of new algorithms in classical and quantum simulation in this dissertation are closely related to or can be better understood in the near adiabatic dynamics. In the classical simulation, the Born-Oppenheimer approximation can be justified by the adiabatic theorems because the eigenvalue problem for the electronic Hamiltonian becomes a good approximation of the electronic quantum dynamics when it is near the adiabatic regime. Therefore the Born-Oppenheimer approximation can only be applied to the system where the electronic Hamiltonian satisfies the spectrum gap condition, and its approximation errors can be analyzed using the adiabatic theorems. Another example is that the advantage of the parallel transport dynamics proposed in Part III can be more transparent and rigorously justified in the near adiabatic regime. This is because the parallel transport dynamics is driven by the residue terms, which becomes very small in the near adiabatic regime.

Near adiabatic dynamics plays an even more important role in quantum simulation. Adiabatic theorems guarantee that the solution of the quantum dynamics approximately remains in certain eigenspace of the gapped Hamiltonian. Therefore large-scale eigenvalue problems can be solved on quantum computers by encoding the target matrix into the final Hamiltonian and simulating quantum dynamics. This procedure is called *adiabatic quantum computing* (AQC). More precisely, to find an eigenvector of a Hermitian matrix A[10], we construct a time-dependent Hamiltonian $H(t) = (1-t)B + tA$ for another simple Hamiltonian B and solve the corresponding quantum dynamics. If the construction of B is adequate such that the eigenvector of B is given or easy to prepare, and $H(t)$ satisfies the gap condition, then, starting from the eigenvector of B, the final solution of the quantum dynamics

[10]It also works for non-Hermitian matrices using the dilation trick at the sacrifice of doubling the size of the linear space.

governed by $H(t)$ will reasonably approximate the eigenvector of A at which we aim. In Chapter 6 of this dissertation, we show how the linear system problems can be solved by the AQC approach. In particular, we discuss the procedure of transforming a linear system problem into an eigenvalue problem and encoding the corresponding eigenvalue problem into a quantum dynamics with sufficient gap condition. Specifically, our construction of the time-dependent Hamiltonian is to interpolate the initial and the final target Hamiltonian by optimized scheduling functions and beyond the linear interpolation. Such optimized scheduling functions allow us to sufficiently employ the instantaneous spectrum gap information and reduce the adiabatic error as much as possible. Our new AQC-based quantum linear system solver achieves exponential speedup over the classical linear system solvers and achieves near-optimal complexity in all other parameters of interest, including the precision and the condition number of the linear system problem.

Organization

The rest of this dissertation is organized as follows. In Part II, we prove a new version of the adiabatic theorem with exponential convergence in the inverse evolution time and almost quadratic dependence in the spectrum gap up to a logarithmic factor. Part III studies simulating quantum dynamics on classical computers with the focus on dealing with the computational challenge of fast oscillatory solutions. We propose a framework of propagating the dynamics under the parallel transport gauge, which can remove unnecessary oscillations and allow much larger time step steps in numerical integrators. Theoretical analysis and numerical experiments are carried out for evolution of pure states in Chapter 2, for time-dependent density functional theory in Chapter 3, and for evolution of mixed states in Chapter 4. Part IV is devoted to simulating quantum dynamics on quantum computers. We first show in Chapter 5 that Trotter and generalized Trotter methods can perform well for quantum dynamics with better regularity by establishing improved error bounds in vector norm scalings. Then in Chapter 6 we study how the adiabatic theorems and quantum simulation can be weaved together to design a quantum linear system solver with near-optimal scalings in the precision and the condition number.

Please note that Chapter 2 is based on [5] (joint work with Lin Lin), Chapter 3 is based on [85] (joint work with Weile Jia, Lin Lin, and Lin-Wang Wang), Chapter 4 is based on [6] (joint work with Di Fang and Lin Lin), Chapter 5 is based on [4] (joint work with Di Fang and Lin Lin), and Chapter 6 is based on [7] (joint work with Lin Lin).

Part II

Quantum linear adiabatic theorem

Chapter 1

Adiabatic theorem with exponential time convergence and quadratic gap dependence

1.1 Introduction

This chapter provides a rigorous description of the linear near adiabatic evolution and proves a new version of the adiabatic theorem, which is almost tight in both parameters: the inverse evolution time and the magnitude of the spectrum gap. Consider the following linear time-dependent Schödinger equation

$$\mathrm{i}\frac{1}{T}\partial_s\psi(s) = H(s)\psi(s), \quad 0 \le s \le 1 \tag{1.1.1}$$

with the initial condition $\psi(0)$ being a normalized eigenvector of the initial Hamiltonian $H(0)$. Here we have already rescaled the time such that s denotes the rescaled time and T represents the original physical evolution time. The implicit assumption that the Hamiltonian $H(s)$ only depends on the rescaled time s indicates that the Hamiltonian changes very slowly in the real physical evolution.

An important assumption for the near adiabatic dynamics is the so-called gap condition. The gap condition roughly says that there exist eigenpaths of the Hamiltonian $H(s)$ which can be separated from the rest of the spectrum uniformly by a positive constant (called gap). Here we follow [82] to provide a rigorous description of the gap condition:

Assumption 1. *There exist two real-valued, continuous function $b_+(s)$ and $b_-(s)$, and a number $\Delta_* > 0$, such that*

$$dist(\{b_+(s), b_-(s)\}, \sigma(H(s))) =: \Delta(s)/2 \ge \Delta_*/2. \tag{1.1.2}$$

A remark is that here distance between the functions b and the spectrum of the Hamiltonian is denoted to be half of the parameter Δ, because we wish to relate Δ to the distance between the two eigenvalues forming the spectrum gap, which is consistent with the physical energy gap.

Let $P(s)$ denote the spectral projection operator associated with the nonempty band $\sigma(H(s)) \cap [b_-(s), b_+(s)]$. Under Assumption 1 and some further regularization assumptions for $H(s)$ (which will be specified later), the adiabatic theorem says that for the quantum dynamics, if the initial vector $\psi(0)$ is within the range of $P(0)$, then the final solution $\psi(1)$ will be approximately within the range of $P(1)$. The leakage of the dynamics to the space $\text{span}(I - P(1))$, referred to as the adiabatic error, mainly depends on two parameters: the evolution time T and the spectrum gap Δ. Qualitatively speaking, the adiabatic error will be reduced when the evolution time T becomes larger and the gap Δ becomes larger.

The rest of this part is organized as follows. In Section 1.2 we provide a brief summary of existing rigorous versions of the linear adiabatic theorems. We focus on the explicit dependence of the adiabatic error on the evolution time T and the gap Δ, as well as the advantages and limitations of the existing results. We then prove a new version of the linear adiabatic theorem in Section 1.3. Our new adiabatic theorem overcomes the limitations of the existing results and becomes sharp in both the evolution time T and the gap Δ. Conclusions and possible further directions are given in Section 1.4.

1.2 Existing adiabatic theorems

There are numerous versions of the adiabatic theorems under different assumptions, leading to different bounds on the adiabatic error. Here we only summarize a few of the rigorously proved linear adiabatic theorems, which are representative for studying the time and gap dependence. Interested readers are referred to [2] for a nice review of linear adiabatic theorems, as well as [57] for a detailed review of recently developed nonlinear adiabatic theorems.

We start with the dependence of the adiabatic error on the evolution time T. Typical results reveal a linear convergence of adiabatic error in the inverse time $1/T$, under the (relatively not strong) assumption that the Hamiltonian $H(s)$ is twice continuously differentiable. Here we present one of these theorems established in [82].

Theorem 2. *Assume that $H(s)$ is twice continuously differentiable, and let $m(s)$ denote the number of different eigenvalues of which $P(s)$ consists. Then for any normalized vector $\psi(0) \in P(0)$ and any $s \in [0, 1]$, we have*

$$\|\psi(s)\psi^*(s) - P(s)\| \leq A(s) \tag{1.2.1}$$

where there exists an absolute constant $C > 0$ such that

$$A(s) \leq C \left[\frac{m(0)\|H'(0)\|}{T\Delta(0)^2} + \frac{m(s)\|H'(s)\|}{T\Delta(s)^2} + \frac{1}{T}\int_0^s \left(\frac{m(\tau)\|H''(\tau)\|}{\Delta(\tau)^2} + \frac{m(\tau)^{3/2}\|H'(\tau)\|^2}{\Delta(\tau)^3} \right) d\tau \right]. \tag{1.2.2}$$

The idea of proving Theorem 2 is first representing the adiabatic error as an integral over a slowly varying function on the time scale $\sim T$, then applying the integration by parts formula to obtain an extra order in $1/T$. Notice that such a procedure can be repeated to obtain extra orders in $1/T$. Roughly speaking, by applying integration by parts formula repeatedly, the adiabatic error can be reformulated as (where BC represents the boundary condition)

$$\begin{aligned} \text{adiabatic error} &= \frac{1}{T}\text{BC}_1 + \frac{1}{T}\int_0^s \\ &= \frac{1}{T}\text{BC}_1 + \frac{1}{T^2}\text{BC}_2 + \frac{1}{T^2}\int_0^s \\ &= \frac{1}{T}\text{BC}_1 + \frac{1}{T^2}\text{BC}_2 + \frac{1}{T^3}\text{BC}_3 + \frac{1}{T^3}\int_0^s \\ &= \cdots\cdots . \end{aligned}$$

Once all the boundary conditions vanish and the Hamiltonian $H(s)$ is smooth enough for legal repeated integration by parts procedure, the converge order of the adiabatic error in $1/T$ can be further improved to be arbitrarily high, and even exponential.

One of the rigorously established adiabatic theorems with exponentially small error and an explicit gap dependence is in [62], though via a different approach from repeated integration by parts. To introduce this result, we first state the technical assumptions, namely the vanishing boundary condition and a regularity assumption on the Hamiltonian $H(t)$. The vanishing boundary condition can be rigorously stated as follows.

Assumption 3. *$H(s)$ is smooth, and for any $k \geq 1$,*

$$H^{(k)}(0) = H^{(k)}(1) = 0. \tag{1.2.3}$$

As discussed before, the vanishing boundary condition allows the adiabatic error to cancel at the boundary and thus is the key to the exponential convergence order in $1/T$. However, the vanishing boundary condition naturally excludes the situation that $H(t)$ is real analytic because if $H(t)$ is real analytic on $[0, 1]$, then according to Taylor's theorem, $H(t)$ is just a constant matrix over the entire time interval and the corresponding quantum dynamics

degenerates to the trivial scenario. Therefore $H(t)$ should belong to a function class between the real analytic function class and the smooth function class. One example is the Gevrey class defined in the following assumption.

Assumption 4. *$H(s)$ is in the Gevrey class G^α for $\alpha > 0$ in the sense that there exist constants $C, D > 0$ such that for all $k \geq 0$,*

$$\max_{s\in[0,1]} \|H^{(k)}(s)\| \leq CD^k \frac{(k!)^{1+\alpha}}{(k+1)^2}. \tag{1.2.4}$$

Notice that G^0 is a subspace of the real analytic function space, and for any $\alpha > 0$, there exists a function in G^α which is not real analytic. Therefore Gevrey class can be viewed as a generalization of the real analytic function class.

The adiabatic theorem developed in [62] is as follows.

Theorem 5. *Under Assumption 1, Assumption 3, Assumption 4, for any normalized vector $\psi(0) \in P(0)$, the final adiabatic error can be bounded as*

$$\|\psi(1)\psi^*(1) - P(1)\| \leq \frac{c}{\Delta_*} \exp\left(-\left(cT\Delta_*^3\right)^{\frac{1}{1+\alpha}}\right) \tag{1.2.5}$$

where c is a positive constant only depending on C, D and α.

Now we discuss the gap dependence, which refers to the dependence of the evolution time T on the gap Δ if the adiabatic error is controlled below a fixed level independent of the gap. Notice that both Theorem 2 and Theorem 5 indicate a cubic gap dependence, namely T should be at least $\mathcal{O}(1/\Delta_*^3)$ to control the adiabatic error. However, it is expected that the general gap dependence is quadratic due to numerical tests on various applications and the quadratic growth in a single integration by parts step conjectured in Theorem 2. To theoretically establish the quadratic gap dependence, [53] proves the following result.

Theorem 6. *Assume the Hamiltonian $H(t)$ satisfies the gap condition (Assumption 1), belongs to the Gevrey class (Assumption 4), and that the minimal gap $\Delta_* \ll h$ where $h = \|H(0)\| = \|H(1)\|$. If there exists a Δ-independent constant $K > 0$ such that*

$$T \geq \frac{K}{\Delta_*^2} |\log(\Delta_*/h)|^{6(\alpha+1)}, \tag{1.2.6}$$

then the adiabatic error $\|\psi(s)\psi^*(s) - P(s)\|$ is $o(1)$ for all $s \in [0,1]$ as Δ_* goes to 0.

Theorem 6 shows that to control the adiabatic error, the evolution time indeed scales almost quadratically in the inverse gap up to a logarithmic factor. This is better than the

cubic gap dependence established in Theorem 2 and Theorem 5. Furthermore, such an almost quadratic dependence is nearly tight since [34] constructs an example showing that generic adiabatic error cannot scale better than the quadratic dependence in the inverse gap. However, Theorem 6 does not reveal how the adiabatic error depends on the evolution time explicitly, *i.e.* how T should scale in ϵ if we would like to bound the error by a given ϵ.

1.3 Our result

Here, under the vanishing boundary condition and the Gevrey class assumption, we prove a new adiabatic theorem combining exponential convergence order in the inverse evolution time and almost quadratic gap dependence. This result was not previously published, and this dissertation is its first appearance. Our result can be viewed as an improvement of Theorem 5 with better gap dependence, and as an improvement of Theorem 6 with explicit convergence order in $1/T$. The generic quadratic gap dependence is also beyond Theorem 5 and is generally sharp unless there is more specific knowledge of the spectral information along the adiabatic path. A similar result has only been established for two-level systems through Landau-Zener formula [163, 99, 88], and we are not aware of any existing rigorous result beyond two-level systems. The proof of our result is largely built upon [121, 62, 7], with a more careful tracking on the spectrum gap dependence.

The organization of this section is as follows. We first state our result and provide the skeleton of the proof. Then the proof is completed by presenting technical lemmas and estimating the growth of several operators' derivatives.

Main result and proof idea

Theorem 7. *Let the Hamiltonian $H(t)$ satisfy the gap condition (Assumption 1), the vanishing boundary condition (Assumption 3) and the Gevrey class assumption (Assumption 4). Assume that there is only one eigenvalue (not necessarily simple) in between $b_-(s)$ and $b_+(s)$. Then, for any normalized vector $\psi(0) \in P(0)$,*

1. *for any positive integer M, the final adiabatic error can be bounded as*

$$\|\psi(1)\psi^*(1) - P(1)\| \leq \frac{2^{1+\alpha}}{16} A_1 A_3 (T^{-1} A_2 M^{1+\alpha})^M, \tag{1.3.1}$$

where

$$A_1 = 2(1 + 2c_f^2 + 4c_f^4)^{-1} = \mathcal{O}(1),$$
$$A_2 = \frac{4c_f^3(1 + 2c_f^2 + 4c_f^4)}{\Delta_*}\left(D + \frac{c_f CD}{\Delta_*}\right) = \mathcal{O}(1/\Delta_*^2),$$
$$A_3 = D + \frac{c_f CD}{\Delta_*} = \mathcal{O}(1/\Delta_*),$$

2. for $T > eA_2$, we have

$$\|\psi(1)\psi^*(1) - P(1)\| \leq \frac{2^{1+\alpha}}{16} A_1 A_3 \exp\left(-\left(\frac{T}{eA_2}\right)^{\frac{1}{1+\alpha}}\right), \tag{1.3.2}$$

3. in order to bound the adiabatic error by ϵ, it suffices to choose

$$T = \mathcal{O}\left(\frac{1}{\Delta_*^2}\left(\log\left(\frac{1}{\Delta_*\epsilon}\right)\right)^{1+\alpha}\right). \tag{1.3.3}$$

Proof. Let $P_0(s) = P(s)$. Motivated by the asymptotic expansion of the projection onto the invariant space of $H(s)$, we define recursively the operators E_j such that

$$[H(s), E_0(s)] = 0, \quad \mathrm{i}\partial_s E_j(s) = [H(s), E_{j+1}(s)], \quad E_j(s) = \sum_{m=0}^{j} E_m(s)E_{j-m}(s). \tag{1.3.4}$$

It has been proved in [121] that the solution of (1.3.4) with initial condition $E_0 = P_0$ is given by

$$E_0(s) = P_0(s) = -(2\pi\mathrm{i})^{-1}\oint_{\Gamma(s)} (H(s) - z)^{-1} dz, \tag{1.3.5}$$
$$E_j(s) = (2\pi)^{-1}\oint_{\Gamma(s)} (H(s) - z)^{-1}[E_{j-1}^{(1)}(s), P_0(s)](H(s) - z)^{-1} dz + S_j(s) - 2P_0(s)S_j(s)P_0(s), \tag{1.3.6}$$

where

$$S_j(s) = \sum_{m=1}^{j-1} E_m(s)E_{j-m}(s), \tag{1.3.7}$$

and $\Gamma(s)$ is a circle centered at the eigenvalue of interest with radius Δ_*.

For an arbitrary positive integer M, we define an operator $P_M(s)$ via a truncated series

$$P_M(s) = \sum_{j=0}^{M} E_j(s) T^{-j}. \tag{1.3.8}$$

Such a $P_M(s)$ is almost the projection onto the invariant space of $H(s)$. In particular,

$$\mathrm{i}\frac{1}{T}P_M^{(1)} - [H, P_M] = \mathrm{i}\frac{1}{T}\sum_{j=0}^{M} E_j^{(1)} T^{-j} - \sum_{j=0}^{M}[H, E_j]T^{-j} = \mathrm{i}T^{-(M+1)}E_M^{(1)}. \tag{1.3.9}$$

In Lemma 8, we prove that $P_M(0) = P_0(0)$ and $P_M(1) = P_0(1)$. Let $U_T(s)$ denote the evolution operator of the dynamics Eq. (1.1.1), then the adiabatic error becomes

$$\begin{aligned}
\|\psi(1)\psi^*(1) - P(1)\| &= \left\|P_0(1) - U_T(1)P_0(0)U_T(1)^\dagger\right\| \\
&= \left\|U_T(1)^\dagger P_0(1)U_T(1) - P_0(0)\right\| \\
&= \left\|U_T(1)^\dagger P_M(1)U_T(1) - P_M(0)\right\| \\
&= \left\|\int_0^1 \frac{d}{ds}(U_T^\dagger P_M U_T)ds\right\|.
\end{aligned} \tag{1.3.10}$$

Notice that (by taking derivative on the equation $U_T U_T^\dagger = I$)

$$\frac{d}{ds}U_T^\dagger = \mathrm{i}TU_T^\dagger H \tag{1.3.11}$$

and

$$\begin{aligned}
\frac{d}{ds}\left(U_T^\dagger P_M U_T\right) &= \frac{d}{ds}(U_T^\dagger)P_M U_T + U_T^\dagger \frac{d}{ds}(P_M)U_T + U_T^\dagger P_M \frac{d}{ds}(U_T) \\
&= \mathrm{i}TU_T^\dagger H P_M U_T - \mathrm{i}TU_T^\dagger [H, P_M]U_T + T^{-M}U_T^\dagger E_M^{(1)}U_T - \mathrm{i}TU_T^\dagger P_M H U_T \\
&= T^{-M}U_T^\dagger E_M^{(1)}U_T,
\end{aligned} \tag{1.3.12}$$

then we have

$$\|\psi(1)\psi^*(1) - P(1)\| \le \left\|\int_0^1 T^{-M}U_T^\dagger E_M^{(1)}U_T ds\right\| \le T^{-M}\max_{s\in[0,1]}\|E_M^{(1)}\|. \tag{1.3.13}$$

In Lemma 13 we prove that

$$\|E_M^{(1)}\| \le A_1 A_2^M A_3 \frac{((M+1)!)^{1+\alpha}}{(1+1)^2(M+1)^2}. \tag{1.3.14}$$

Using the fact that $(M+1)! \leq 2M^M$,

$$\|E_M^{(1)}\| \leq \frac{2^{1+\alpha}}{16} A_1 A_2^M A_3 (M^M)^{1+\alpha} = \frac{2^{1+\alpha}}{16} A_1 A_3 (A_2 M^{1+\alpha})^M. \tag{1.3.15}$$

Together with Eq. (1.3.13), we complete the first part of the proof.

2. Notice that the estimate in the first part holds for all positive integer M. For sufficiently large T, we choose

$$M = \left\lfloor \left(\frac{eA_2}{T} \right)^{-\frac{1}{1+\alpha}} \right\rfloor \geq 1. \tag{1.3.16}$$

Then

$$\begin{aligned} \|\psi(1)\psi^*(1) - P(1)\| &\leq \frac{2^{1+\alpha}}{16} A_1 A_3 (T^{-1} A_2 M^{1+\alpha})^M \\ &\leq \frac{2^{1+\alpha}}{16} A_1 A_3 \exp\left(- \left(\frac{T}{eA_2} \right)^{\frac{1}{1+\alpha}} \right). \end{aligned} \tag{1.3.17}$$

3. By requiring the bound in the second part to be smaller than ϵ, we can obtain a sufficient condition for T that

$$T \geq eA_2 \left(\log \left(\frac{2^{1+\alpha}}{16} \frac{A_1 A_3}{\epsilon} \right) \right)^{1+\alpha}. \tag{1.3.18}$$

The desired complexity estimate can be obtained by noticing that $A_1 = \mathcal{O}(1), A_2 = \mathcal{O}(1/\Delta_*^2)$ and $A_3 = \mathcal{O}(1/\Delta_*)$. □

Completion of the proof

We complete the technical details of the proof of Theorem 7 by stating and proving several lemmas regarding the growth of the derivatives of the resolvent, which is defined as $R(z,s) = (H(s)-z)^{-1}$, as well as $E_j(s)$ and $P_j(s)$. It is worth mentioning in advance that in the proof we will encounter derivatives taken on a contour integral. In fact all such derivatives taken on a contour integral will not involve derivatives on the contour, since derivatives are local information and we can fix the contour to be the same when we study the growth of the derivatives according to Cauchy's theorem. By writing $R^{(k)}$ we only consider the explicit time derivatives brought by H.

Lemma 8. *1. For all $k \geq 1$, $E_0^{(k)}(0) = P_0^{(k)}(0) = 0, E_0^{(k)}(1) = P_0^{(k)}(1) = 0$.*

2. For all $j \geq 1, k \geq 0$, $E_j^{(k)}(0) = E_j^{(k)}(1) = 0$.

Proof. We will repeatedly use the fact that $R^{(k)}(0) = R^{(k)}(1) = 0$. This can be proved by taking the kth order derivative of the equation $(H - z)R = I$ and

$$R^{(k)} = -R\sum_{l=1}^{k}\binom{k}{l}(H-z)^{(l)}R^{(k-l)} = -R\sum_{l=1}^{k}\binom{k}{l}H^{(l)}R^{(k-l)}.$$

1. This is a straightforward result by the definition of E_0 and the fact that $R^{(k)}$ vanish on the boundary.

2. We prove by induction with respect to j. For $j = 1$, Eq. (1.3.6) tells that

$$E_1 = (2\pi)^{-1}\oint_\Gamma R[P_0^{(1)}, P_0]Rdz.$$

Therefore each term in the derivatives of E_1 must involve at least one of the derivative of R and the derivative of P_0, which means the derivatives of E_1 much vanish on the boundary.

Assume the conclusion holds for $< j$, then for j, first each term of the derivatives of S_j much involve the derivative of some E_m with $m < j$, which means the derivatives of S_j much vanish on the boundary. Furthermore, for the similar reason, Eq. (1.3.6) tells that the derivatives of E_j must vanish on the boundary. □

The following three technical lemmas are introduced in [121, 62]. where $c_f = 4\pi^2/3$ denote an absolute constant.

Lemma 9. *Let $\alpha > 0$ be a positive real number, p, q be non-negative integers and $r = p + q$. Then*

$$\sum_{l=0}^{k}\binom{k}{l}\frac{[(l+p)!(k-l+q)!]^{1+\alpha}}{(l+p+1)^2(k-l+q+1)^2} \le c_f\frac{[(k+r)!]^{1+\alpha}}{(k+r+1)^2}.$$

Lemma 10. *Let k be a non-negative integer, then*

$$\sum_{l=0}^{k}\frac{1}{(l+1)^2(k+1-l)^2} \le c_f\frac{1}{(k+1)^2}.$$

Lemma 11. *Let $A(s), B(s)$ be two smooth matrix-valued function defined on $[0,1]$ satisfying*

$$\|A^{(k)}(s)\| \le a_1 a_2^k\frac{[(k+p)!]^{1+\alpha}}{(k+1)^2}, \quad \|B^{(k)}(s)\| \le b_1 b_2^k\frac{[(k+q)!]^{1+\alpha}}{(k+1)^2}$$

for some positive constants a_1, a_2, b_1, b_2, non-negative integers p, q and for all $k \ge 0$. Then for every $k \ge 0, 0 \le s \le 1$,

$$\|(A(s)B(s))^{(k)}\| \le c_f a_1 b_1 \max\{a_2, b_2\}^k\frac{[(k+r)!]^{1+\alpha}}{(k+1)^2}$$

where $r = p + q$.

Now we are ready to bound the growth of the derivatives of R, P_j and E_j, which is given in the following two lemmas. Notice that here the technique of bounding the contour integrals is new compared to the previous works [7, 62].

Lemma 12. *For all $k \geq 0$ and each fixed $z \in \Gamma(s)$,*

$$\|R^{(k)}(z,s)\| \leq \frac{1}{\Delta_*}\left(1+\frac{c_f C}{\Delta_*}\right)^k D^k \frac{(k!)^{1+\alpha}}{(k+1)^2}.$$

Proof. We prove it by induction. The case $k = 0$ is straightforward. Assume the lemma holds for $< k$, by taking derivatives of the equation $(H - z)R = I$, we have

$$R^{(k)} = -R\sum_{l=1}^{k}\binom{k}{l}(H-z)^{(l)}R^{(k-l)} = -R\sum_{l=1}^{k}\binom{k}{l}H^{(l)}R^{(k-l)}. \tag{1.3.19}$$

By induction assumption, Assumption 4 and Lemma 9, we have

$$\|R^{(k)}\| \leq \frac{1}{\Delta_*}\sum_{l=1}^{k}\binom{k}{l}CD^l\frac{(l!)^{1+\alpha}}{(l+1)^2}\frac{1}{\Delta_*}\left(1+\frac{c_f C}{\Delta_*}\right)^{k-l}D^{k-l}\frac{((k-l)!)^{1+\alpha}}{(k-l+1)^2}. \tag{1.3.20}$$

Notice that $c_f C/\Delta_* \leq (1 + c_f C/\Delta_*)^l$. Combining this inequality with Lemma 9, we have

$$\begin{aligned}\|R^{(k)}\| &\leq \frac{1}{c_f\Delta_*}\left(1+\frac{c_f C}{\Delta_*}\right)^k D^k\sum_{l=1}^{k}\binom{k}{l}\frac{(l!)^{1+\alpha}}{(l+1)^2}\frac{((k-l)!)^{1+\alpha}}{(k-l+1)^2}\\ &\leq \frac{1}{\Delta_*}\left(1+\frac{c_f C}{\Delta_*}\right)^k D^k\frac{(k!)^{1+\alpha}}{(k+1)^2}.\end{aligned} \tag{1.3.21}$$

This completes the proof. □

Lemma 13. *1. For all $k \geq 0$,*

$$\|E_0^{(k)}\| = \|P_0^{(k)}\| \leq \left(1+\frac{c_f C}{\Delta_*}\right)^k D^k\frac{(k!)^{1+\alpha}}{(k+1)^2}. \tag{1.3.22}$$

2. For all $k \geq 0, j \geq 1$,

$$\|E_j^{(k)}\| \leq A_1 A_2^j A_3^k \frac{[(k+j)!]^4}{(k+1)^2(j+1)^2} \tag{1.3.23}$$

with

$$\begin{aligned}A_1 &= 2(1+2c_f^2+4c_f^4)^{-1},\\ A_2 &= \frac{4c_f^3(1+2c_f^2+4c_f^4)}{\Delta_*}\left(D+\frac{c_f CD}{\Delta_*}\right),\\ A_3 &= D+\frac{c_f CD}{\Delta_*}.\end{aligned}$$

Proof. 1. We take the time derivatives of the contour integral representation of

$$P_0(s) = -(2\pi \mathrm{i})^{-1} \oint_{\Gamma(s)} (H(s) - z)^{-1} dz, \tag{1.3.24}$$

with respect to time s. As we have discussed before, although there is some time dependence in the contour $\Gamma(s)$, such a time dependence in the contour will not be taken into account when we compute the derivative, because the derivatives are local property and we can replace $\Gamma(s)$ by $\Gamma(s_0)$ locally for some fixed s_0 due to the continuity of the spectrum. With this consideration, using Lemma 12 we have

$$\left\| P_0^{(k)} \right\| = \left\| (2\pi \mathrm{i})^{-1} \oint_{\Gamma} R^{(k)} dz \right\| \leq \left(1 + \frac{c_f C}{\Delta_*} \right)^k D^k \frac{(k!)^{1+\alpha}}{(k+1)^2}. \tag{1.3.25}$$

2. First we remark that the choice of A_1, A_2 and A_3 satisfies the conditions

$$\left(\frac{1}{2} + c_f^2 + 2c_f^4 \right) A_1 \leq 1 \tag{1.3.26}$$

$$A_1 A_2 = \frac{8c_f^3}{\Delta_*} A_3. \tag{1.3.27}$$

These relations, together with the definition of A_3, will be used in the proof.

We prove this by induction on j. For $j = 1$, by the definition of E_1 (Eq. (1.3.6)), Lemma 12, Lemma 11 and part 1 of Lemma 13, we have

$$\begin{aligned}
\left\| E_1^{(k)} \right\| &= \left\| (2\pi)^{-1} \frac{d^k}{ds^k} \oint_{\Gamma} R[P_0^{(1)}, P_0] R dz \right\| \\
&\leq 2\Delta_* c_f^3 \frac{1}{\Delta_*^2} \left(1 + \frac{c_f C}{\Delta_*} \right) D \left(1 + \frac{c_f C}{\Delta_*} \right)^k D^k \frac{((k+1)!)^{1+\alpha}}{(k+1)^2} \\
&= \frac{8c_f^3}{\Delta_*} A_3 A_3^k \frac{((k+1)!)^{1+\alpha}}{(k+1)^2 (1+1)^2} \\
&\leq A_1 A_2 A_3^k \frac{((k+1)!)^{1+\alpha}}{(k+1)^2 (1+1)^2}.
\end{aligned} \tag{1.3.28}$$

Now we assume the lemma holds for all $< j$. For the case j, we first bound the derivative of

S_j. By the definition of S_j, Lemma 10, Lemma 11 and the induction assumption,

$$\begin{aligned}\|S_j^{(k)}\| &= \left\| \frac{d^k}{ds^k} \left(\sum_{m=1}^{j-1} E_m E_{j-m} \right) \right\| \\ &\leq \sum_{m=1}^{j-1} c_f A_1^2 A_2^j A_3^k \frac{((k+j)!)^{1+\alpha}}{(k+1)^2(m+1)^2(j-m+1)^2} \\ &\leq c_f^2 A_1^2 A_2^j A_3^k \frac{((k+j)!)^{1+\alpha}}{(k+1)^2(j+1)^2}. \end{aligned} \tag{1.3.29}$$

Then by the definition of E_j (Eq. (1.3.6)), Lemma 11, Lemma 12, part 1 of Lemma 13 and the induction assumption,

$$\begin{aligned}\left\| E_j^{(k)} \right\| &\leq \left\| \frac{d^k}{ds^k} \left((2\pi)^{-1} \oint_\Gamma R[E_{j-1}^{(1)}, P_0] R dz \right) \right\| + \left\| \frac{d^k}{ds^k} S_j \right\| + \left\| \frac{d^k}{ds^k} (2P_0 S_j P_0) \right\| \\ &\leq \Delta_* c_f^3 \frac{1}{\Delta_*^2} A_1 A_2^{j-1} A_3 A_3^k \frac{((k+j)!)^{1+\alpha}}{(k+1)^2 j^2} + c_f^2 A_1^2 A_2^j A_3^k \frac{((k+j)!)^{1+\alpha}}{(k+1)^2(j+1)^2} \\ &\quad + 2c_f^2 c_f^2 A_1^2 A_2^j A_3^k \frac{((k+j)!)^{1+\alpha}}{(k+1)^2(j+1)^2} \\ &\leq \frac{4c_f^3}{\Delta_*} \frac{A_3}{A_2} A_1 A_2^j A_3^k \frac{((k+j)!)^{1+\alpha}}{(k+1)^2(j+1)^2} + c_f^2 A_1^2 A_2^j A_3^k \frac{((k+j)!)^{1+\alpha}}{(k+1)^2(j+1)^2} \\ &\quad + 2c_f^4 A_1^2 A_2^j A_3^k \frac{((k+j)!)^{1+\alpha}}{(k+1)^2(j+1)^2} \\ &= \left(\frac{4c_f^3}{\Delta_*} \frac{A_3}{A_2} + c_f^2 A_1 + 2c_f^4 A_1 \right) A_1 A_2^j A_3^k \frac{((k+j)!)^{1+\alpha}}{(k+1)^2(j+1)^2} \\ &= \left[\left(\frac{1}{2} + c_f^2 + 2c_f^4 \right) A_1 \right] A_1 A_2^j A_3^k \frac{((k+j)!)^{1+\alpha}}{(k+1)^2(j+1)^2} \\ &= A_1 A_2^j A_3^k \frac{((k+j)!)^{1+\alpha}}{(k+1)^2(j+1)^2}. \end{aligned} \tag{1.3.30}$$

We complete the proof. □

1.4 Conclusion

In this part, we prove a new version of quantum linear adiabatic theorem (Theorem 7) under the Gevrey class assumption on the Hamiltonian and the vanishing boundary condition. Our result is almost sharp in both the inverse evolution time and the spectrum gap since it

simultaneously achieves the exponential error convergence in the inverse evolution time and the almost quadratic gap dependence to obtain certain precision. The proof strategy mostly follows existing works [121, 62, 7], with a more careful tracking on the gap dependence.

It can be very interesting to see whether our result is still valid or partially valid if some assumptions are loosened. For example, if the vanishing boundary condition is not satisfied, can we still prove an adiabatic theorem with quadratic gap dependence and linear error convergence in $1/T$? If so, this can be a nice improvement of both Theorem 2 and Theorem 6, and becomes sharp in both parameters ϵ and Δ in the scenario without vanishing boundary condition. Another possible direction is to investigate the case when the spectrum of interest consists of more than one eigenvalue. We conjecture that Theorem 7 still holds in this case, and the proof can be done through a cleverer way of bounding the contour integrals, for example, changing the contour from a circle to a tall thin rectangle enclosing the spectrum of interest. We leave the detailed analysis to future work.

Finally, we remark that in Theorem 7 we only consider how the errors depend on the minimal spectrum gap Δ_* instead of simultaneous gap $\Delta(s)$ for technical simplicity. Our technique can be generalized to track simultaneous gap dependence, which further improves the quadratic gap dependence in specific applications where more information is provided on the spectrum of $H(s)$. We refer to Section 6.11 for an example of how the linear gap dependence is achieved in applying adiabatic quantum computing to quantum linear system problems.

Part III

Simulating quantum dynamics on classical computers

This part is devoted to simulating quantum dynamics on classical computers, focusing on dealing with the fast oscillatory solution. The key observation is that such fast oscillations in wave functions are not physically intrinsic, and the oscillations in the density matrix can be much slower. More precisely, we propose that the gauge choice (*i.e.* degrees of freedom irrelevant to physical observables) of the Schrödinger equation can be generally non-optimal for numerical simulation. This can limit, and in some cases, severely limit the time step size. We find that the optimal gauge choice is given by a parallel transport formulation. This parallel transport dynamics can be interpreted as the dynamics driven by the residual vectors, analogous to those defined in eigenvalue problems in the time-independent setup. We remark that what this part focuses on is not to develop another numerical scheme to discretize the quantum dynamics directly, but to propose an alternative formulation that is equivalent to the original quantum dynamics and can be solved with improved numerical efficiency using existing discretization schemes.

In Chapter 2, we start with the simplest scenario of the evolution of a single pure state. We derive the dynamics under parallel transport gauge via two approaches: solving the optimization problem and evolving the dynamics under the parallel transport operator. The parallel transport dynamics with tiny modification can also be derived from a Hamiltonian structure, thus suitable to be solved using a symplectic and implicit time discretization scheme, such as the implicit midpoint rule, which allows the usage of a large time step and ensures the long time numerical stability. We analyze the parallel transport dynamics in the context of the singularly perturbed linear Schrödinger equation and demonstrate its superior performance in the near adiabatic regime. We then demonstrate the effectiveness of our method using numerical results for toy examples as well as linear and nonlinear Schrödinger equations.

In Chapter 3, we derive the dynamics under parallel transport gauge for Real-time time-dependent density functional theory (RT-TDDFT), which becomes prevalent in studying ultrafast dynamics. Under the parallel transport gauge, RT-TDDFT calculations can be significantly accelerated using a combination of the parallel transport gauge and implicit integrators, and the resulting scheme can be used to accelerate any electronic structure software that uses a Schrödinger representation. Using absorption spectrum, ultrashort laser pulse, and Ehrenfest dynamics calculations as examples, we show that the new method can utilize a time step that is on the order of $10 \sim 100$ attoseconds using a planewave basis set. Thanks to the significant increase in the size of the time step, we also demonstrate that the new method is more than 10 times faster in terms of the wall clock time when compared to the standard explicit fourth-order Runge-Kutta time integrator for silicon systems ranging from 32 to 1024 atoms.

In Chapter 4, we generalize the parallel transport dynamics for simulating pure states to general quantum states which can be possibly mixed. Going beyond the linear and near adiabatic regime in previous chapters, we find that the error of the parallel transport

dynamics can be bounded by certain commutators between Hamiltonians, density matrices, and their derived quantities. Such a commutator structure is not present in the Schrödinger dynamics. The commutator structure of the error bound and numerical results for model RT-TDDFT calculations in both linear and nonlinear regimes confirm the advantage of the parallel transport dynamics.

Chapter 2

Pure-state parallel transport dynamics

2.1 Introduction

Consider the following set of coupled nonlinear Schrödinger equations

$$\mathrm{i}\epsilon\partial_t\Psi(t) = H(t,P)\Psi(t). \tag{2.1.1}$$

Here we assume $0 < \epsilon \ll 1$. $\Psi(t) = [\psi_1(t), \ldots, \psi_N(t)]$ are N time-dependent wave functions subject to suitable initial and boundary conditions. $H(t,P)$ is a self-adjoint time-dependent Hamiltonian. $P(t)$ is called the density matrix and defined as

$$P(t) = \Psi(t)\Psi^*(t) = \sum_{j=1}^{N} \psi_j(t)\psi_j^*(t). \tag{2.1.2}$$

Note that when the initial state $\Psi(0)$ consists of N orthonormal functions, the functions in $\Psi(t)$ will remain orthonormal for all t, i.e. $(\psi_i(t), \psi_j(t)) = \delta_{ij}$, where $(\cdot,\cdot)$ denotes a suitable inner product. Then

$$P^2(t) = \sum_{j,k=1}^{N} \psi_j(t)(\psi_j(t), \psi_k(t))\psi_k^*(t) = \sum_{j=1}^{N} \psi_j(t)\psi_j^*(t) = P(t),$$

i.e. $P(t)$ is a projector. The explicit dependence of the Hamiltonian on t is often due to the existence of an external field, and we assume the partial derivatives $\frac{\partial^m H}{\partial t^m}$ are of $\mathcal{O}(1)$ in some suitable norms for all $m \geq 1$. Hence when $0 < \epsilon \ll 1$, the wave functions can oscillate on a much smaller time scale than that of the external fields, and this is called the singularly perturbed regime [70].

The equations (2.1.1) are rather general and appear in several fields of scientific computation. In the simplest setup when $N = 1$ and $H(t, P) \equiv H(t)$, this is the linear Schrödinger equation. Another example is the nonlinear Schrödinger equation (NLSE) used for modeling nonlinear photonics and Bose-Einstein condensation process [58],

$$\mathrm{i}\epsilon\partial_t\psi(t) = H_0(t)\psi(t) + g|\psi(t)|^2\psi(t), \tag{2.1.3}$$

where $H_0(t)$ is a Hermitian matrix obtained by discretizing the linear operator $-\frac{1}{2}\Delta+V(x,t)$. Since $N = 1$, $P(t) = \psi(t)\psi^*(t)$, and $|\psi(t)|^2 = \mathrm{diag}[P(t)]$ is a nonlinear local potential. When $N > 1$, the coupled set of Schrödinger equations must be solved simultaneously. This is the case in the time-dependent density functional theory (TDDFT) [133, 126].

Note that if we multiply $\Psi(t)$ by a time-dependent unitary matrix $U(t) \in \mathbb{C}^{N\times N}$, the resulting set of rotated wave functions, denoted by $\Phi(t) = \Psi(t)U(t)$, yields the same density matrix as

$$P(t) = \Phi(t)\Phi^*(t) = \Psi(t)\left[U(t)U^*(t)\right]\Psi^*(t) = \Psi(t)\Psi^*(t). \tag{2.1.4}$$

Since the unitary rotation matrix $U(t)$ is irrelevant to the density matrix which is used to represent many physical observables, $U(t)$ is called the gauge, and Eq. (2.1.4) indicates the density matrix is *gauge-invariant*. Furthermore, Eq. (2.1.1) can be directly written in terms of the density matrix as

$$\mathrm{i}\epsilon\partial_t P(t) = [H(t, P), P(t)], \tag{2.1.5}$$

where $[H, P] := HP - PH$ is the commutator between H and P. Eq. (2.1.5) is called the von Neumann equation (or quantum Liouville equation), which can be viewed as a more intrinsic representation of quantum dynamics since the gauge degrees of freedom are eliminated completely.

The simulation of the von Neumann equation can also be advantageous from the perspective of time discretization. Consider the simplified scenario that $H(t, P) \equiv H(P)$ does not explicitly depend on t, and the initial state $\Psi(0)$ consists of a set of eigenfunctions of H, *i.e.*

$$H[P]\psi_j(0) = \psi_j(0)\lambda_j(0), \quad j = 1, \ldots, N, \quad P = \sum_{j=1}^{N}\psi_j(0)\psi_j^*(0). \tag{2.1.6}$$

Eq. (2.1.6) is a set of nonlinear eigenvalue equations. When solved self-consistently, the solution to the Schrödinger equation (2.1.1) has an analytic form

$$\psi_j(t) = \exp\left(-\frac{\mathrm{i}}{\epsilon}\lambda_j(0)t\right)\psi_j(0), \quad j = 1, \ldots, N, \tag{2.1.7}$$

which oscillates on the $\mathcal{O}(\epsilon)$ time scale. Hence many numerical schemes still need to resolve the dynamics with a time step of $\mathcal{O}(\epsilon)$. On the other hand, the right hand side of the

von Neumann equation vanishes for all t, and hence nominally can be discretized with an arbitrarily large time step! Of course one can use techniques such as integration factors [46] to make this simulation using the Schrödinger equation as efficient. However this example illustrates that the gap in terms of the size of the time step generally exists between the Schrödinger representation and the von Neumann representation.

In this chapter, we identify that such gap is solely due to the gauge degrees of freedom in the Schrödinger representation. By optimizing the gauge choice, one can propagate the wave functions using a time step comparable to that of the von Neumann equation. We demonstrate that the optimized gauge is given by a parallel transport (PT) formulation. We refer to this gauge as the parallel transport gauge, and the resulting dynamics as the parallel transport dynamics. Correspondingly the trivial gauge $U(t) \equiv I_N$ in Eq. (2.1.1) is referred to as the Schrödinger gauge, and the resulting dynamics as the Schrödinger dynamics. We remark that the PT dynamics can also be interpreted as an analytic and optimal way of performing the dynamical low rank approximation [97] for Eq. (2.1.1). Note that the simulation of the von Neumann equation requires the explicit operation on the density matrix $P(t)$. When a large basis set such as finite elements or planewaves is used to discretize the partial differential equation, the storage cost of $P(t)$ can be often prohibitively expensive compared to that of the wave functions $\Psi(t)$. Hence the PT dynamics combines the advantages of both approaches, namely to perform simulation using the time step size of the von Neumann equation, but with cost comparable to that of the Schrödinger equation.

We analyze the effectiveness of the parallel transport dynamics for the linear time-dependent Schrödinger equation in the near adiabatic regime. We remark that efficient numerical methods have been recently developed in this regime based on the construction of a set of instantaneous adiabatic states [80, 152]. The assumption is that the wave functions can be approximated by the subspace spanned by low energy eigenstates of the Hamiltonian at each t. The dimension of the subspace is often chosen to be cN, where c is a relatively small constant. Compared to these methods, the PT dynamics always operates only on N wave functions, and therefore has reduced computational and the storage cost. The PT dynamics is also applicable beyond the near adiabatic regime.

By extending the quantum adiabatic theorem [121, 9] to the PT dynamics, we prove that the local truncation error of the PT dynamics gains an extra order of accuracy in terms of ϵ, when the time step is $\mathcal{O}(\epsilon)$ or smaller. The PT dynamics, after a slight modification, can be derived from a Hamiltonian system similar to that in the Schrödinger dynamics. Hence the gain of accuracy for the local truncation error can be directly translated to the global error as well for long time simulation.

We demonstrate the effectiveness of the PT dynamics using numerical results of the model linear and nonlinear Schrödinger equations. When the spectral radius of the Hamiltonian is large, it is suitable to discretize the PT dynamics using a symplectic and implicit time discretization scheme, such as the implicit midpoint rule, and the resulting scheme can

significantly outperform the same scheme for the Schrödinger dynamics. We also find that other time-reversible and implicit time discretization schemes, such as the Crank-Nicolson scheme, can yield similar performance as well. Numerical results confirm our analysis in the near adiabatic regime, and indicate that the convergence of the PT dynamics can start when the time step size is much larger than $\mathcal{O}(\epsilon)$. This is in contrast to the Schrödinger dynamics where the error stays flat until the time step reaches below $\mathcal{O}(\epsilon)$.

The rest of this chapter is organized as follows. We derive the parallel transport gauge in Section 2.2, and discuss the numerical discretization of the parallel transport dynamics in Section 2.3. We analyze the parallel transport dynamics in the singularly perturbed regime in Section 2.4. We then present the numerical results in Section 2.5, followed by the conclusion in Section 2.6.

2.2 Parallel Transport Gauge

Since the concept of the parallel transport gauge is associated with the time propagation instead of spatial discretization, for simplicity of the presentation, unless otherwise specified, we assume that Eq. (2.1.1) represents a discrete, finite dimensional quantum system, i.e. for a given time t, $\psi_j(t)$ is a finite dimensional vector, and $H(t,P)$ is a finite dimensional matrix. If the quantum system is spatially continuous, we may first find a set of orthonormal bases functions $\{e_j(\mathbf{r})\}_{j=1}^d$ satisfying $\int e_j^*(\mathbf{r})e_{j'}(\mathbf{r})\,\mathrm{d}\mathbf{r} = \delta_{jj'}$, and expand the continuous wavefunction as $\widetilde{\psi}_j(\mathbf{r},t) \approx \sum_{j=1}^d \psi_j(t)e_j(\mathbf{r})$. Then after a Galerkin projection, Eq. (2.1.1) becomes a d-dimensional quantum system, and the inner product for the coefficients $\psi_j(t)$ becomes the standard ℓ^2-inner product as $(\psi_j(t),\psi_k(t)) := \psi_j^*(t)\psi_k(t) = \delta_{jk}$. Hence we can use the linear algebra notation. The star notation is interpreted as the complex conjugation when applied to a scalar, and Hermitian conjugation when applied to a vector or a matrix.

Derivation

For simplicity let us consider the case $N = 1$ first, where the gauge matrix $U(t)$ simply becomes a phase factor $c(t) \in \mathbb{C}, |c(t)| = 1$. Note that the gauge choice cannot affect physical observables such as the density matrix. Hence conceptually we may think that the time-dependent density matrix $P(t)$ has already been obtained as the solution of the von Neumann equation (2.1.5) on some time interval $[0,T]$. Similarly the wave function $\psi(t)$ satisfying the Schrödinger dynamics is also known. Then the relation

$$P(t)\varphi(t) = \varphi(t), \quad \varphi(t) = \psi(t)c(t) \tag{2.2.1}$$

is satisfied for any gauge choice. For simplicity we use the notation $\dot{\varphi}(t) = \partial_t\varphi(t)$, and drop the explicit t-dependence in all quantities, as well as the P-dependence in the Hamiltonian

unless otherwise noted. Our goal is to find the time-dependent gauge factor $c(t)$ so that the rotated wave function $\varphi(t)$ varies *as slowly as possible*. This gives rise to the following minimization problem,

$$\begin{aligned} \min_{c(t)} \quad & \|\dot{\varphi}(t)\|_2^2 \\ \text{s.t.} \quad & \varphi(t) = \psi(t)c(t), \quad |c(t)| = 1. \end{aligned} \tag{2.2.2}$$

In order to solve (2.2.2), note that $P(t)$ is a projector, we split $\dot{\varphi}$ into two orthogonal components,

$$\dot{\varphi} = P\dot{\varphi} + (I - P)\dot{\varphi}. \tag{2.2.3}$$

By taking the time derivative with respect to both sides of the first equation in Eq. (2.2.1), we have

$$(I - P)\dot{\varphi} = \dot{P}\varphi. \tag{2.2.4}$$

Then

$$\begin{aligned} \|\dot{\varphi}\|_2^2 &= \|P\dot{\varphi}\|_2^2 + \|(I - P)\dot{\varphi}\|_2^2 \\ &= \|P\dot{\varphi}\|_2^2 + \|\dot{P}\varphi\|_2^2 \\ &= \|P\dot{\varphi}\|_2^2 + \|\dot{P}\psi\|_2^2. \end{aligned} \tag{2.2.5}$$

In the last equality, we have used that $|c(t)| = 1$. Note that the term $\|\dot{P}\psi\|_2^2$ is independent of the gauge choice, so $\|\dot{\varphi}\|_2^2$ is minimized when

$$P\dot{\varphi} = 0. \tag{2.2.6}$$

Therefore instead of writing down the minimizer of Eq. (2.2.2) directly, we define the gauge implicitly through Eq. (2.2.6).

Let us write down an equation for $\varphi(t)$ directly. Combining equations (2.2.4), (2.2.6), (2.1.5) and (2.2.1), we have

$$\dot{\varphi} = \dot{P}\varphi = \frac{1}{\mathrm{i}\epsilon}[H, P]\varphi = \frac{1}{\mathrm{i}\epsilon}(H\varphi - \varphi(\varphi^* H\varphi)), \tag{2.2.7}$$

or equivalently

$$\mathrm{i}\epsilon\partial_t\varphi = H\varphi - \varphi(\varphi^* H\varphi). \tag{2.2.8}$$

For reasons that will become clear shortly, we refer to this gauge choice as the *parallel transport gauge*, and Eq. (2.2.8) as the parallel transport (PT) dynamics. Comparing with the Schrödinger dynamics, we find that the PT dynamics only introduces one extra term

$\varphi(\varphi^* H\varphi)$. The right hand side of Eq. (2.2.8) takes the form of the residual vector in the solution of eigenvalue problem of the form (2.1.6). Hence the PT dynamics can be simply interpreted as the dynamics driven by the residuals. Therefore we expect that the PT dynamics can be particularly advantageous in the *near adiabatic* regime [80, 152], *i.e.* when φ is close to be the eigenstate of H, and all the residual vectors are therefore small.

Now we provide an alternative interpretation of the gauge choice using the parallel transport formulation associated with a family of projectors. For simplicity let us assume $H(t)$ is already discretized into a finite dimensional Hermitian matrix for each t and so is $P(t)$. Given the single parameter family of projectors $\{P(t)\}$ defined on some interval $[0, T]$, we define

$$\mathcal{A}(t) = \mathrm{i}\epsilon[\partial_t P(t), P(t)]. \tag{2.2.9}$$

It can be directly verified that $\mathcal{A}(t)$ is a Hermitian matrix for each t, and induces a dynamics

$$\mathrm{i}\epsilon\partial_t \mathcal{T}(t) = \mathcal{A}(t)\mathcal{T}(t), \quad \mathcal{T}(0) = I. \tag{2.2.10}$$

$\mathcal{T}(t)$ is a unitary matrix for each t. $\mathcal{T}(t)$ is called the parallel transport evolution operator (see *e.g.* [119, 47]). The connection between the parallel transport dynamics and the parallel transport evolution operator is given in Proposition 14.

Proposition 14. *Define $\varphi(t) = \mathcal{T}(t)\psi(0)$ where $\mathcal{T}(t)$ is the evolution operator satisfying* (2.2.10)*, and $P(t)$ satisfies the von Neumann equation* (2.1.5)*. Then $P(t) = \varphi(t)\varphi^*(t)$, and $\varphi(t)$ satisfies the parallel transport dynamics* (2.2.8)*.*

Proof. First we prove the following relation

$$P(t)\mathcal{T}(t) = \mathcal{T}(t)P(0) \tag{2.2.11}$$

by showing that both sides solve the same initial value problem. Note that $\mathcal{T}(t)P(0)$ satisfies a differential equation of the form (2.2.10) with the initial value $\mathcal{T}(0)P(0)$. We would like to derive the differential equation $P(t)\mathcal{T}(t)$ satisfies. Taking the time derivative on both sides of the identity $P(t) = P^2(t)$, we yield two useful relations

$$\dot{P} = \dot{P}P + P\dot{P}, \quad P\dot{P}P = 0. \tag{2.2.12}$$

Then using Eq. (2.2.10),

$$\mathrm{i}\epsilon\partial_t(P\mathcal{T}) = \mathrm{i}\epsilon\dot{P}\mathcal{T} + \mathrm{i}\epsilon P[\dot{P}, P]\mathcal{T} = \mathrm{i}\epsilon\dot{P}P\mathcal{T}.$$

On the other hand,

$$\mathcal{A}(P\mathcal{T}) = \mathrm{i}\epsilon(\dot{P}PP\mathcal{T} - P\dot{P}P\mathcal{T}) = \mathrm{i}\epsilon\dot{P}P\mathcal{T}.$$

Therefore

$$\mathrm{i}\epsilon\partial_t(P\mathcal{T}) = \mathcal{A}(P\mathcal{T}). \tag{2.2.13}$$

Hence $P\mathcal{T}$ also satisfies an equation of the form (2.2.10). This proves Eq. (2.2.11) by noticing further the shared initial condition $P(0)\mathcal{T}(0) = \mathcal{T}(0)P(0)$.

Using Eq. (2.2.11), we have

$$P(t)\varphi(t) = P(t)\mathcal{T}(t)\psi(0) = \mathcal{T}(t)P(0)\psi(0) = \mathcal{T}(t)\psi(0) = \varphi(t). \tag{2.2.14}$$

Since $\mathcal{T}(t)$ is unitary, we have $\|\varphi(t)\|_2 = 1$ for all t. Hence

$$P(t) = \varphi(t)\varphi^*(t). \tag{2.2.15}$$

The only thing left is to show that the gauge choice in $\varphi(t)$ is indeed the parallel transport gauge. Using Eq. (2.2.11) and (2.2.13), we have

$$\mathrm{i}\epsilon\partial_t\varphi = \mathrm{i}\epsilon\partial_t(\mathcal{T}\psi(0)) = \mathrm{i}\epsilon\partial_t(P\mathcal{T})\psi(0) = \mathrm{i}\epsilon\dot{P}P\mathcal{T}\psi(0) = HP\varphi - PHP\varphi. \tag{2.2.16}$$

Here we have used the von Neumann equation

$$\mathrm{i}\epsilon\dot{P} = HP - PH.$$

Finally using Eq. (2.2.14) and (2.2.15), we have

$$\mathrm{i}\epsilon\partial_t\varphi = H\varphi - \varphi(\varphi^* H\varphi),$$

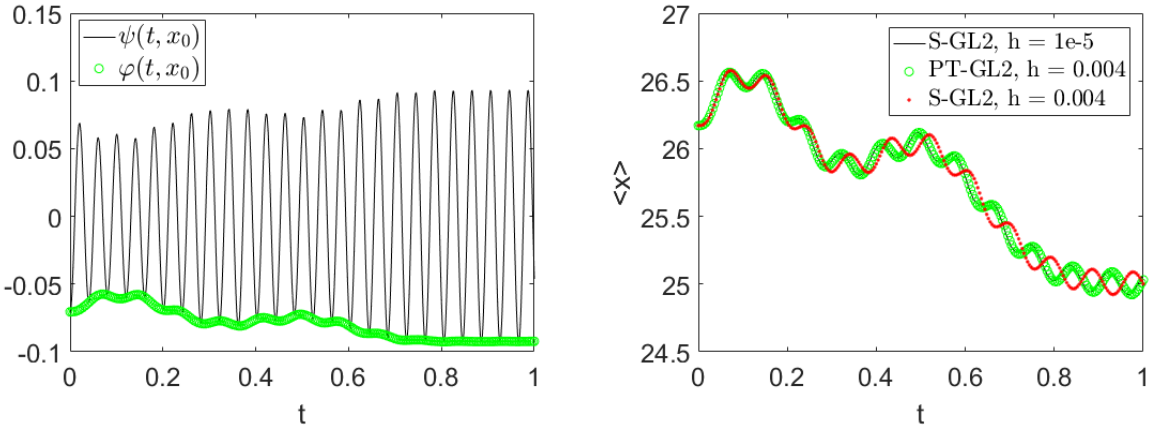
which is precisely the parallel transport dynamics. □

In order to see why the parallel transport gauge can be more advantageous, consider again the time-independent example (2.1.6) in the introduction for the case $N = 1$. We find that the right hand side of Eq. (2.2.8) vanishes, and the solution is simply

$$\varphi(t) = \varphi(0) = \psi(0)$$

for all t. This implies that the parallel transport gauge is $c(t) = \exp\left(+\frac{\mathrm{i}}{\epsilon}\lambda(0)t\right)$ that perfectly cancels with the rotating factor in (2.1.7). Hence the PT dynamics yields the slowest possible dynamics by completely eliminating the time-dependent phase factor, and the time step for propagating the PT dynamics can be chosen to be arbitrarily large as in the case of the von Neumann equation.

For a more complex example, consider a time-dependent nonlinear Schrödinger equation in one dimension to be further illustrated in Section 2.5. Fig. 2.2.1 (a) shows the evolution of the real part of the solution $\psi(t)$ from the Schrödinger dynamics, and that of $\varphi(t)$ from the PT dynamics, respectively. We find that the trajectory of $\varphi(t)$ varies considerably slower



(a) wave functions

(b) Evolution of centers

Figure 2.2.1: (a) Real parts of the wave functions at $x_0 = 25$ with the Schrödinger gauge and the PT gauge, respectively. (b) Centers of the wave functions. Parameters are chosen to be $T = 1, \epsilon = 0.005$, and the reference solution is obtained from S-GL2 with time step size $h = 10^{-5}$.

than that of $\psi(t)$, which allows us to use a much larger time step for the simulation. Fig. 2.2.1 (b) measures the accuracy of the average of the orbital center $\langle x \rangle (t)$, using simulation with the implicit midpoint rule, also known as the Gauss-Legendre method of order 2 (GL2) scheme. We compare the performance of the GL2 scheme with the Schrödinger gauge (S-GL2) and that with the PT gauge (PT-GL2) with the same step size $h = 0.004$, and the reference solution is obtained using a very small step size $h = 10^{-5}$. We observe that the solution from PT-GL2 agrees very well with the reference solution, while the phase error of the solution from S-GL2 becomes noticeable already after $t = 0.2$.

Hamiltonian structure

For simplicity let us consider the linear Schrödinger equation, *i.e.* $H(t, P) \equiv H(t)$, and assume $H(t)$ is a real symmetric matrix for all t. It is well known that the Schrödinger dynamics is a Hamiltonian system [122, 123, 68]. More specifically, we separate the solution ψ into its real and imaginary parts as

$$\psi = q + \mathrm{i}p. \tag{2.2.17}$$

The ℓ^2-inner product associated with real quantities such as p, q are denoted by $(p, q) := p^T q$. We also introduce the canonically conjugate pair of variables (τ, E) to eliminate the explicit

dependence of $H(t)$ on time [31, 68]. This gives the following energy functional

$$\mathcal{E}(\tau, q, E, p) = \frac{1}{2\epsilon}\left[q^T H(\tau) q + p^T H(\tau) p\right] + E. \tag{2.2.18}$$

The Hamiltonian system corresponding to this energy functional is

$$\begin{aligned}
\partial_t \tau &= \frac{\partial \mathcal{E}}{\partial E} = 1, \\
\partial_t q &= \frac{\partial \mathcal{E}}{\partial p} = \frac{1}{\epsilon} H(\tau) p, \\
\partial_t E &= -\frac{\partial \mathcal{E}}{\partial \tau} = -\frac{1}{2\epsilon}\left[q^T \frac{\partial H(\tau)}{\partial \tau} q + p^T \frac{\partial H(\tau)}{\partial \tau} p\right], \\
\partial_t p &= -\frac{\partial \mathcal{E}}{\partial q} = -\frac{1}{\epsilon} H(\tau) q.
\end{aligned} \tag{2.2.19}$$

Hence τ is simply the time variable, and $-E$ is the usually defined energy of the system up to a constant. By combining the equations for q, p we obtain the Schrödinger dynamics for ψ.

Although the PT dynamics only differs from the Schrödinger dynamics by the choice of the gauge, interestingly, the PT dynamics cannot be directly written as a Hamiltonian system. To illustrate this, we first separate the real and imaginary parts of φ as in (2.2.17), and the PT dynamics can be written as

$$\begin{aligned}
\partial_t q =& \frac{1}{\epsilon}(Hp - (q^T H q + p^T H p) p), \\
\partial_t p =& \frac{1}{\epsilon}(-Hq + (q^T H q + p^T H p) q).
\end{aligned} \tag{2.2.20}$$

If this dynamics can be derived from some energy functional $\mathcal{E}$, then

$$\begin{aligned}
\frac{\partial \mathcal{E}}{\partial p} &= \frac{1}{\epsilon}(Hp - (q^T H q + p^T H p) p), \\
\frac{\partial \mathcal{E}}{\partial q} &= \frac{1}{\epsilon}(Hq - (q^T H q + p^T H p) q).
\end{aligned} \tag{2.2.21}$$

Straightforward computation reveals that $\frac{\partial^2 \mathcal{E}}{\partial p \partial q} = \frac{\partial^2 \mathcal{E}}{\partial q \partial p}$ is not true in general, and hence the PT dynamics (2.2.8) cannot be a Hamiltonian system.

Fortunately, the PT dynamics can be slightly modified to become a Hamiltonian system. Consider the following modified energy functional

$$\mathcal{E}(\tau, q, E, p) = \frac{1}{2\epsilon}(q^T H(\tau) q + p^T H(\tau) p)(2 - q^T q - p^T p) + E. \tag{2.2.22}$$

The corresponding Hamiltonian equations are

$$\begin{aligned}
\partial_t \tau &= \frac{\partial \mathcal{E}}{\partial E} = 1, \\
\partial_t q &= \frac{\partial \mathcal{E}}{\partial p} = \frac{1}{\epsilon}\left[H(\tau)p(2 - q^T q - p^T p) - (q^T H(\tau) q + p^T H(\tau) p)p\right], \\
\partial_t E &= -\frac{\partial \mathcal{E}}{\partial \tau}, \\
\partial_t p &= -\frac{\partial \mathcal{E}}{\partial q} = \frac{1}{\epsilon}\left[-H(\tau)q(2 - q^T q - p^T p) + (q^T H(\tau) q + p^T H(\tau) p)q\right].
\end{aligned} \tag{2.2.23}$$

Again τ is the same as t, and the conjugate variable $E(t)$ satisfies

$$E(t) = -\frac{1}{2\epsilon}(q^T H(t) q + p^T H(t) p)(2 - q^T q - p^T p) + \text{constant}.$$

Compared to the PT dynamics (2.2.20), we have an extra factor $(2 - q^T q - p^T p)$ in the equations and the energy. Proposition 15 states that the solution to the PT dynamics (2.2.20) is the same as the solution of the Hamiltonian system (2.2.23).

Proposition 15. *If (τ, q, E, p) solves the Hamiltonian system* (2.2.23) *with normalized initial value condition $p^T(0)p(0) + q^T(0)q(0) = 1$, then $(q(t), p(t))$ solves* (2.2.20) *with the same initial value condition, and $\varphi(t) = q(t) + \mathrm{i}p(t)$ solves the PT dynamics* (2.2.8).

Proof. Comparing Eq. (2.2.23) with Eq. (2.2.20), we only need to show the identity

$$p^T p + q^T q = 1$$

holds for all t. By computing

$$\begin{aligned}
\frac{d}{dt}(p^T p + q^T q) =& 2(p^T \partial_t p + q^T \partial_t q) \\
=& \frac{1}{\epsilon}(-2(2 - q^T q - p^T p)p^T H q + 2(q^T H q + p^T H p)p^T q \\
&+ 2(2 - q^T q - p^T p)q^T H p - 2(q^T H q + p^T H p)q^T p) = 0,
\end{aligned}$$

we find that $p^T p + q^T q$ is invariant during the propagation. Together with the normalized initial condition, we complete the proof. □

Proposition 15 suggests that the Hamiltonian form of the PT dynamics is

$$\mathrm{i}\epsilon \partial_t \varphi = H\varphi(2 - \varphi^* \varphi) - \varphi(\varphi^* H \varphi), \tag{2.2.24}$$

which shares exactly the same solution with (2.2.8) using the condition $\varphi^*\varphi = 1$.

At the end of this part, we briefly discuss the Hamiltonian structure of the nonlinear Schrödinger equation and the associated PT dynamics. Let us consider the discretized nonlinear Schrödinger equation (2.1.3), which can be reformulated as a Hamiltonian system driven by the energy functional

$$\mathcal{E}(\tau, q, E, p) = \frac{1}{2\epsilon}\left[q^T H_0(\tau) q + p^T H_0(\tau) p + \frac{g}{2}\mathrm{Tr}((|q|^2 + |p|^2)^2)\right] + E. \tag{2.2.25}$$

The PT dynamics corresponding to Eq. (2.1.3) can be written as

$$\mathrm{i}\epsilon\partial_t\varphi = H_0\varphi + g|\varphi|^2\varphi - \varphi(\varphi^* H_0\varphi) - g\varphi(\varphi^*|\varphi|^2\varphi). \tag{2.2.26}$$

Similar to the linear case, the PT dynamics itself cannot be reformulated as a Hamiltonian system in general, but can be slightly modified to become a Hamiltonian system. More precisely, define the energy functional

$$\begin{aligned}\mathcal{E}(\tau, q, E, p) = &\frac{1}{2\epsilon}\left[q^T H_0(\tau) q + p^T H_0(\tau) p + g\mathrm{Tr}((|q|^2 + |p|^2)^2)\right](2 - q^T q - p^T p)\\ &- \frac{g}{4\epsilon}\mathrm{Tr}((|q|^2 + |p|^2)^2) + E,\end{aligned} \tag{2.2.27}$$

then the Hamiltonian system driven by this energy functional can be written as

$$\mathrm{i}\epsilon\partial_t\varphi = (H_0\varphi + 2g|\varphi|^2\varphi)(2 - \varphi^*\varphi) - \varphi(\varphi^* H_0\varphi) - g\varphi(\varphi^*|\varphi|^2\varphi) - g|\varphi|^2\varphi. \tag{2.2.28}$$

Again this equation shares the same solution with Eq. (2.2.26) using the condition $\varphi^*\varphi = 1$.

General case

The PT dynamics derived in the previous sections can be directly generalized to Eq. (2.1.1) with $N > 1$. Define the transformed set of wave functions

$$\Phi(t) = \Psi(t)U(t) = [\varphi_1(t), \ldots, \varphi_N(t)],$$

where $U(t) \in \mathbb{C}^{N\times N}$ is a gauge matrix. Following the same derivation in Section 2.2, we find that the parallel transport gauge is given by the condition

$$P\dot{\Phi} = 0. \tag{2.2.29}$$

This gives rise to the following PT dynamics

$$\mathrm{i}\epsilon\partial_t\Phi(t) = H(t, P(t))\Phi(t) - \Phi(t)[\Phi^*(t)H(t, P(t))\Phi(t)], \quad P(t) = \Phi(t)\Phi^*(t). \tag{2.2.30}$$

Again the PT dynamics is driven by the residual vectors as in eigenvalue problems.

In addition, the Hamiltonian structure is also preserved for the PT dynamics. For simplicity let us consider the linear Hamiltonian $H(t)$. We separate the set of PT wave functions Φ into real and imaginary parts as

$$\Phi(t) = \mathfrak{q}(t) + \mathrm{i}\mathfrak{p}(t).$$

Define the energy functional

$$\mathcal{E}(\tau, \mathfrak{q}, E, \mathfrak{p}) = \frac{1}{2\epsilon}\mathrm{Tr}\Big((\mathfrak{q}^T H(\tau)\mathfrak{q} + \mathfrak{p}^T H(\tau)\mathfrak{p})(2I_N - \mathfrak{q}^T\mathfrak{q} - \mathfrak{p}^T\mathfrak{p})\Big) + E. \tag{2.2.31}$$

The associated Hamiltonian system is

$$\begin{aligned}
\partial_t \tau &= \frac{\partial \mathcal{E}}{\partial E} = 1, \\
\partial_t \mathfrak{q} &= \frac{\partial \mathcal{E}}{\partial \mathfrak{p}} = \frac{1}{\epsilon}(H(\tau)\mathfrak{p}(2I_N - \mathfrak{q}^T\mathfrak{q} - \mathfrak{p}^T\mathfrak{p}) - \mathfrak{p}(\mathfrak{q}^T H(\tau)\mathfrak{q} + \mathfrak{p}^T H(\tau)\mathfrak{p})), \\
\partial_t E &= -\frac{\partial \mathcal{E}}{\partial \tau}, \\
\partial_t \mathfrak{p} &= -\frac{\partial \mathcal{E}}{\partial \mathfrak{q}} = \frac{1}{\epsilon}(-H(\tau)\mathfrak{q}(2I_N - \mathfrak{q}^T\mathfrak{q} - \mathfrak{p}^T\mathfrak{p}) + \mathfrak{q}(\mathfrak{q}^T H(\tau)\mathfrak{q} + \mathfrak{p}^T H(\tau)\mathfrak{p})).
\end{aligned} \tag{2.2.32}$$

Similar with the case when $N = 1$ (Proposition 15), we can show that

$$\mathfrak{p}^T\mathfrak{p} + \mathfrak{q}^T\mathfrak{q} = I_N$$

provided the orthonormal initial value condition. Therefore the solution to the Hamiltonian system (2.2.32) can exactly form a set of solutions to the PT dynamics.

Due to the straightforward generalization as described above, unless otherwise noted, we will focus on the case $N = 1$ for the rest of the chapter.

2.3 Time discretization

When the spectral radius of the Hamiltonian is relatively small and $\epsilon \sim \mathcal{O}(1)$, explicit time integrators such as the 4th order Runge-Kutta method (RK4) and the Strang splitting method can be very efficient, and can be applied to both the Schrödinger dynamics and the PT dynamics. However, the advantage of propagating the PT dynamics can become clearer when ϵ becomes small or when the spectral radius of H becomes very large, which is typical in *e.g.* TDDFT calculations. In this scenario, all explicit time integrators must take a very small time step, which may become very costly. It should be noted that in the

Schrödinger dynamics, the solution often oscillates rapidly on the time scale of ϵ as indicated in Eq. (2.1.7). Standard implicit discretization schemes, such as the implicit midpoint rule and the Crank-Nicolson scheme, aim at interpolating such rapidly moving curves by low order polynomials. Therefore the time step must still be kept on the order of ϵ to meet the accuracy requirement, even though the numerical scheme itself may have a large stability region or even A-stable [69].

On the other hand, as discussed in Section 2.2, the PT dynamics transforms the fast oscillating wave function $\psi(t)$ into a potentially slowly oscillating wave function $\varphi(t)$ (as in Fig. 2.2.1 (a)). This makes it feasible to approximate $\varphi(t)$ using a low order polynomial approximation. This statement will be further quantified by numerical results in Section 2.5. Combined with an implicit time discretization scheme with a large stability region, we may expect that the PT dynamics can be discretized with a much larger time step than that in the Schrödinger dynamics.

The Hamiltonian structure of the PT dynamics further invites the usage of a symplectic scheme for achieving long time accuracy and stability. The simplest symplectic and implicit scheme is the implicit mid-point rule, also known as the Gauss-Legendre method of order 2 (GL2). We use a uniform time discretization $t_n = nh$, and h is the time step size. With some abuse of notations, we denote by $\varphi(t_n)$ the exact solution at t_n, and φ_n the numerical approximation to $\varphi(t_n)$. Correspondingly we define

$$P_n = \varphi_n \varphi_n^*, \quad H_n = H(t_n, P_n).$$

It would also be helpful to define the effective nonlinear Hamiltonian $H^e(t, \varphi)$ as

$$\begin{aligned} H^e &= H(2 - \varphi^*\varphi) - (\varphi^* H \varphi) I, \text{ for Eq. (2.2.24)}, \\ H^e &= (H_0 + 2g|\varphi|^2)(2 - \varphi^*\varphi) - (\varphi^* H_0 \varphi) I - g(\varphi^*|\varphi|^2\varphi) I - g|\varphi|^2, \text{ for Eq. (2.2.28)}. \end{aligned}$$

Then the Hamiltonian equations (2.2.24) and (2.2.28) can be written in a uniform form

$$\mathrm{i}\epsilon \partial_t \varphi = H^e \varphi. \tag{2.3.1}$$

The PT-Ham-GL2 discretization for discretizing the Hamiltonian equation Eq. (2.2.24) and Eq. (2.2.28) therefore becomes

$$\begin{aligned} \varphi_{n+1} &= \varphi_n + \frac{h}{\mathrm{i}\epsilon} H^e_{n+\frac{1}{2}} \widetilde{\varphi}, \\ \widetilde{\varphi} &= \frac{1}{2}(\varphi_n + \varphi_{n+1}), \end{aligned} \tag{2.3.2}$$

Here $\widetilde{\varphi}$ can be interpreted as the approximation to $\varphi(t_{n+\frac{1}{2}})$ at the half time step, and

$$H^e_{n+\frac{1}{2}} := H^e(t_{n+\frac{1}{2}}, \widetilde{\varphi}).$$

Note that the normalization condition $\widetilde{\varphi}^*\widetilde{\varphi} \to 1$ holds only in the limit $h \to 0$, but $\widetilde{\varphi}^*\widetilde{\varphi} \neq 1$ in general. Eq. (2.3.2) is a set of nonlinear equations for φ_{n+1}, and need to be solved iteratively. This can be viewed as a fixed point problem of the form

$$\varphi = \mathfrak{F}(\varphi),$$

where the mapping $\mathfrak{F}$ is explicitly defined as

$$\mathfrak{F}(\varphi) = \varphi_n + \frac{h}{\mathrm{i}\epsilon} H^e_{n+\frac{1}{2}} \widetilde{\varphi}, \quad \widetilde{\varphi} = \frac{1}{2}(\varphi_n + \varphi). \tag{2.3.3}$$

Assuming the fixed point exists and is unique, we may associate φ_{n+1} with the fixed point, and then move to the next time step. We may use any nonlinear equation solving technique to solve such fixed point problem [94]. In this work, we use the Anderson mixing [8] method, which is a simplified Broyden-type method widely used in electronic structure calculations [105].

The PT-Ham-GL2 scheme can be simplified by directly applying the GL2 discretization to the PT dynamics (2.2.8) and (2.2.26), with the efficient Hamiltonians to be defined as

$$\begin{aligned} H^e &= H - (\varphi^* H \varphi) I, \text{ for Eq. (2.2.8)}, \\ H^e &= H_0 + g|\varphi|^2 - (\varphi^* H_0 \varphi) I - g(\varphi^* |\varphi|^2 \varphi) I, \text{ for Eq. (2.2.26)}. \end{aligned}$$

Again note that, unlike the continuous case, PT-GL2 is not equivalent to PT-Ham-GL2 since $\widetilde{\varphi}^*\widetilde{\varphi} \neq 1$ in general. Nevertheless, the norm of the numerical solutions obtained by GL2 at the discretized time points t_n are indeed conserved, which is summarized in the following proposition.

Proposition 16. *Suppose φ_n is the numerical solution obtained by applying GL2 to one of the following PT dynamics,* (2.2.24)*,* (2.2.28)*,* (2.2.8) *and* (2.2.26)*. Assume that $I - \frac{h}{2\mathrm{i}\epsilon} H^e_{n+\frac{1}{2}}$ is always invertible in each step, then $\|\varphi_n\|_2 = \|\varphi_0\|_2$.*

Proof. We consider the GL2 scheme (2.3.2) for the uniform form (2.3.1). It suffices to prove that $\|\varphi_{n+1}\|_2 = \|\varphi_n\|_2$ for any n. We first substitute $\widetilde{\varphi}$ by $\frac{1}{2}(\varphi_n + \varphi_{n+1})$ and rewrite GL2 as

$$\left(I - \frac{h}{2\mathrm{i}\epsilon} H^e_{n+\frac{1}{2}}\right) \varphi_{n+1} = \left(I + \frac{h}{2\mathrm{i}\epsilon} H^e_{n+\frac{1}{2}}\right) \varphi_n.$$

Note that for all defined H^e, $H^{e*} = H^e$, then

$$\begin{aligned}
&\varphi_{n+1}^* \varphi_{n+1} \\
=&\varphi_n^* \left(I + \frac{h}{2\mathrm{i}\epsilon} H^e_{n+\frac{1}{2}}\right)^* \left(I - \frac{h}{2\mathrm{i}\epsilon} H^e_{n+\frac{1}{2}}\right)^{*-1} \left(I - \frac{h}{2\mathrm{i}\epsilon} H^e_{n+\frac{1}{2}}\right)^{-1} \left(I + \frac{h}{2\mathrm{i}\epsilon} H^e_{n+\frac{1}{2}}\right) \varphi_n \\
=&\varphi_n^* \left(I - \frac{h}{2\mathrm{i}\epsilon} H^e_{n+\frac{1}{2}}\right) \left(I + \frac{h}{2\mathrm{i}\epsilon} H^e_{n+\frac{1}{2}}\right)^{-1} \left(I - \frac{h}{2\mathrm{i}\epsilon} H^e_{n+\frac{1}{2}}\right)^{-1} \left(I + \frac{h}{2\mathrm{i}\epsilon} H^e_{n+\frac{1}{2}}\right) \varphi_n \\
=&\varphi_n^* \left(I - \frac{h}{2\mathrm{i}\epsilon} H^e_{n+\frac{1}{2}}\right) \left(I - \frac{h}{2\mathrm{i}\epsilon} H^e_{n+\frac{1}{2}}\right)^{-1} \left(I + \frac{h}{2\mathrm{i}\epsilon} H^e_{n+\frac{1}{2}}\right)^{-1} \left(I + \frac{h}{2\mathrm{i}\epsilon} H^e_{n+\frac{1}{2}}\right) \varphi_n \\
=&\varphi_n^* \varphi_n,
\end{aligned}$$

where the second to the last line uses the fact that $I - \frac{h}{2\mathrm{i}\epsilon} H^e_{n+\frac{1}{2}}$ and $I + \frac{h}{2\mathrm{i}\epsilon} H^e_{n+\frac{1}{2}}$ commute. □

Similarly we may use other time-reversible (but not symplectic) schemes [68], such as the trapezoidal rule discretization (known in this context as the Crank-Nicolson method). So the PT-CN scheme becomes

$$\varphi_{n+1} = \varphi_n + \frac{h}{2\mathrm{i}\epsilon} H^e_n \varphi_n + \frac{h}{2\mathrm{i}\epsilon} H^e_{n+1} \varphi_{n+1}, \tag{2.3.4}$$

Here $H^e_n = H^e(t_n, \varphi_n), H^e_{n+1} = H^e(t_{n+1}, \varphi_{n+1})$. In both PT-GL2 and PT-CN schemes, we need to solve φ_{n+1} with nonlinear equation solvers as before. Although these schemes are not symplectic schemes and the 2-norm of the numerical solution by PT-CN is not strictly conserved as in PT-Ham-GL2, numerical results in Section 2.5 indicate that the performance of all the three schemes can be very comparable in practice.

Following the discussion above, we may readily obtain the corresponding scheme for $N > 1$ case, as well as higher order and symplectic time discretization schemes, such as the Gauss-Legendre collocation methods [79] for the PT dynamics.

2.4 Analysis in the near adiabatic regime

In this section, we demonstrate the advantage of the PT dynamics by analyzing the accuracy of the discretized PT dynamics in the near adiabatic regime. Our main result is that for $h \leq \mathcal{O}(\epsilon)$, a proper discretization of the PT dynamics gains one extra order of accuracy in ϵ compared to that of the Schrödinger dynamics.

We extend the quantum adiabatic theorem [93, 9, 147] to the PT dynamics, which shows that the PT wave function $\varphi(t)$ can be decomposed into a component of which the oscillation is independent of ϵ and the magnitude is $\mathcal{O}(1)$, and a component that is highly oscillatory

with $\mathcal{O}(\epsilon)$ magnitude. This leads to the desired result in terms of the local truncation error. We then obtain the global error estimate from the standard results of symplectic integrators due to the Hamiltonian structure of the dynamics.

Again, we restrict the scope of the theoretical analysis to the time-dependent linear system with $N = 1$. While the generalization to the case $N > 1$ is straightforward, the analysis beyond the linear system can be considerably more difficult. One important difficulty is the lack of the spectral theory and the corresponding adiabatic theorem for general nonlinear operators [141], which play important roles as being shown in our proof, though progress has been made in recent years for certain types of the nonlinear problems such as the Schrödinger equation with weak nonlinearity [141], and certain quantum-classical molecular dynamics (QCMD) models [24]. We remark that there has been recent progress [57] proving the adiabatic theorem under a more general nonlinear setting. Extension of the work of [57] to the nonlinear PT dynamics will be our future work.

We make the following assumptions through this section, which defines the near adiabatic regime:

1. $H : [0, T] \to \mathbb{C}^{d\times d}$ is a Hermitian-valued and smooth map. The norms $\|H(t)\|_2$ and $\|H^{(k)}(t)\|_2$ for all the time derivatives are bounded independently of ϵ and $t \in [0, T]$.

2. There exists a continuous function $\lambda(t) \in \mathrm{spec}(H(t))$ which is a simple eigenvalue of $H(t)$ and stays separated from the rest of the spectrum, *i.e.* there exists a positive constant Δ such that

$$\mathrm{dist}(\lambda(t), \mathrm{spec}(H(t))\backslash\{\lambda(t)\}) \geq \Delta, \quad \forall t \in [0, T]. \tag{2.4.1}$$

3. The initial state $\varphi(0)$ is the normalized eigenvector of $H(0)$ associated with the eigenvalue $\lambda(0)$.

The assumption 1 ensures that the solutions of both the Schrödinger dynamics and the PT dynamics are smooth with respect to t. The assumption 2 is called the gap condition [147].

Before we continue, we would like to investigate a useful conclusion which can be directly derived from the above assumptions. Let $Q(t)$ denote the projector on the eigenspace corresponding to $\lambda(t)$. $Q(t)$ can be expressed by the Riesz representation of the projector as

$$Q(t) = -\frac{1}{2\pi\mathrm{i}} \int_{\Gamma(t)} R(z, t)dz \tag{2.4.2}$$

in which $R(z, t) = (H(t) - z)^{-1}$ is the resolvent at time t and the complex contour can be chosen as $\Gamma(t) = \{z \in \mathbb{C} : |z - \lambda(t)| = \Delta/2\}$. Note that the assumption 2 assures that such representation is well-defined and, together with assumption 1, $Q(t)$ is actually also a smooth bounded map, which is summarized in the following lemma.

Lemma 17. *The norms of all time derivatives $\|Q^{(k)}(t)\|$ are bounded independently of ϵ.*

Proof. We follow the technique in [147]. The boundedness of $Q(t)$ directly follows from the Riesz representation (2.4.2) and the boundedness of $R(z,t)$ over the contour $\Gamma(t)$. The contour $\Gamma(t)$ depends on t. To avoid taking time derivatives over the contour, note that the continuity of $\lambda(t)$ implies that for any $s \in [0,T]$, there exists a neighborhood $B(s,\delta_s)$ such that

$$|z - \lambda(t)| \geq \Delta/4, \quad \forall t \in B(s,\delta_s) \cap [0,T],\ z \in \Gamma(s).$$

By finding a finite cover $\bigcup_{j=1}^{n} B(s_j,\delta_{s_j}) \supset [0,T]$, for each $t \in [0,T]$, there exists a s_j such that $t \in B(s_j,\delta_{s_j})$ and we can rewrite $Q(t)$ as

$$Q(t) = -\frac{1}{2\pi \mathrm{i}} \int_{\Gamma(s_j)} R(z,t)dz. \tag{2.4.3}$$

Such s_j remains unchanged locally, hence

$$Q^{(k)}(t) = -\frac{1}{2\pi \mathrm{i}} \int_{\Gamma(s_j)} R^{(k)}(z,t)dz.$$

The boundedness of $Q^{(k)}(t)$ can be directly assured by the boundedness of $H^{(k)}(t)$. □

Adiabatic theorem

First let us define the adiabatic evolution $\varphi_A(t)$ as the solution to the following initial value problem

$$\mathrm{i}\epsilon\partial_t\varphi_A = \mathrm{i}\epsilon[\dot{Q},Q]\varphi_A, \quad \varphi_A(0) = \varphi(0). \tag{2.4.4}$$

Since the matrix $\mathrm{i}\epsilon[\dot{Q},Q]$ is Hermitian, $\|\varphi_A\|_2 = 1$ holds for all t. Following the same proof of Eq. (2.2.14) in Proposition 14, we find that φ_A is an eigenvector of $H(t)$ corresponding to $\lambda(t)$, *i.e.* $Q(t)\varphi_A(t) = \varphi_A(t)$ holds for all $t \in [0,T]$.

In the near adiabatic regime, we may separate $\varphi(t)$ into the smooth component φ_A and a remainder term. This is called the adiabatic theorem and is given in Theorem 18.

Theorem 18. *Let $\varphi(t)$ follow the PT dynamics* (2.2.8)*, and let $\varphi_A(t)$ follow the adiabatic evolution as defined in Eq.* (2.4.4)*. Then the following decomposition*

$$\varphi(t) = \varphi_A(t) + \epsilon\varphi_R(t) \tag{2.4.5}$$

holds up to time $T = \mathcal{O}(1)$. Furthermore, $\varphi_R(t)$ is infinitely differentiable, and $\|\varphi_R(t)\|_2$ is bounded independently of ϵ.

Proof. The proof is organized according to the following three steps.

1. Define another adiabatic evolution φ_B, which satisfies an equation that resembles the PT dynamics.

2. Prove the adiabatic decomposition with respect to φ_B, *i.e.* there exists an infinitely differentiable function $\eta(t)$ such that

$$\varphi(t) = \varphi_B(t) + \epsilon\eta(t), \quad \forall t \in [0, T],$$

   where $\|\eta(t)\|_2$ is bounded independently of ϵ.

3. Prove that the difference between φ_B and φ_A is of $\mathcal{O}(\epsilon)$.

1. Define $\mathcal{T}_B$ as the solution to the initial value problem

$$\mathrm{i}\epsilon\partial_t\mathcal{T}_B = (H - \varphi^* H\varphi + \mathrm{i}\epsilon[\dot{Q}, Q])\mathcal{T}_B, \quad \mathcal{T}_B(0) = I, \tag{2.4.6}$$

We define φ_B according to

$$\varphi_B(t) := \mathcal{T}_B(t)\varphi(0),$$

which solves the initial value problem

$$\mathrm{i}\epsilon\partial_t\varphi_B = (H - \varphi^* H\varphi + \mathrm{i}\epsilon[\dot{Q}, Q])\varphi_B, \quad \varphi_B(0) = \varphi(0). \tag{2.4.7}$$

Since the matrix $(H - \varphi^* H\varphi + \mathrm{i}\epsilon[\dot{Q}, Q])$ is Hermitian, $\mathcal{T}_B$ is a unitary evolution, and φ_B is a normalized vector.

Next we show that $\varphi_B(t)$ is an eigenvector of $H(t)$ corresponding to $\lambda(t)$, *i.e.*

$$Q(t)\varphi_B(t) = \varphi_B(t). \tag{2.4.8}$$

This can be done by showing that $Q\varphi_B$ and φ_B solve the same initial value problem. By the Leibniz rule and Eq. (2.4.7), we have

$$\begin{aligned}\partial_t(Q\varphi_B) &= \dot{Q}\varphi_B + Q\dot{\varphi_B} \\ &= \dot{Q}\varphi_B + Q[\dot{Q}, Q]\varphi_B - \frac{\mathrm{i}}{\epsilon}Q(H - \varphi^* H\varphi)\varphi_B.\end{aligned}$$

Use the identities similar to (2.2.12),

$$\dot{Q} = \dot{Q}Q + Q\dot{Q}, \quad Q\dot{Q}Q = 0, \quad Q^2 = Q,$$

we have

$$\begin{aligned}\dot{Q} + Q[\dot{Q}, Q] &= \dot{Q}Q + Q\dot{Q} + Q\dot{Q}Q - Q^2\dot{Q} \\ &= \dot{Q}Q = (\dot{Q}Q - Q\dot{Q})Q = [\dot{Q}, Q]Q.\end{aligned}$$

Hence

$$\partial_t(Q\varphi_B) = [\dot{Q}, Q]Q\varphi_B - \frac{\mathrm{i}}{\epsilon}Q(H - \varphi^* H\varphi)\varphi_B.$$

Together with the identity $QH = HQ$, we have

$$\begin{aligned}\partial_t(Q\varphi_B) &= [\dot{Q}, Q]Q\varphi_B - \frac{\mathrm{i}}{\epsilon}(H - \varphi^* H\varphi)Q\varphi_B \\ &= -\frac{\mathrm{i}}{\epsilon}(H - \varphi^* H\varphi + i\epsilon[\dot{Q}, Q])(Q\varphi_B).\end{aligned}$$

Furthermore, the initial condition satisfies $Q(0)\varphi_B(0) = \varphi_B(0) = \varphi(0)$. Hence $Q\varphi_B$ solves the same initial value problem (2.4.7) as φ_B.

In summary, in step 1 we define another adiabatic evolution $\varphi_B(t)$ which is also an eigenstate of $H(t)$ corresponding to $\lambda(t)$ (Eq. (2.4.8)). Therefore, $\varphi_A(t)$ and $\varphi_B(t)$ are both eigenstates of $H(t)$ differing at most by a choice of gauge.

2. Now we estimate the distance between $\varphi(t)$ and $\varphi_B(t)$. This can be done by mimicking the standard proof of the adiabatic theorem [9] with some modifications. By the definition of φ_B,

$$\|\varphi(t) - \varphi_B(t)\|_2 = \|\varphi(t) - \mathcal{T}_B(t)\varphi(0)\|_2 = \|\mathcal{T}_B^{-1}(t)\varphi(t) - \varphi(0)\|_2.$$

Define $w(t) = \mathcal{T}_B^{-1}(t)\varphi(t)$, then

$$\|\varphi(t) - \varphi_B(t)\|_2 = \|w(t) - w(0)\|_2 = \left\|\int_0^t \dot{w}(s)ds\right\|_2. \tag{2.4.9}$$

In order to estimate $\dot{w}$, differentiate the equation $\mathcal{T}_B w = \varphi$ and we get

$$\dot{w} = -\mathcal{T}_B^{-1}[\dot{Q}, Q]\mathcal{T}_B w. \tag{2.4.10}$$

Note that if we define

$$X(t) = -\frac{1}{2\pi\mathrm{i}}\int_{\Gamma(s_j)} R(z,t)\dot{Q}(t)R(z,t)dz$$

where $\Gamma(s_j)$ and $R(z,t)$ are defined in the proof of Lemma 17, then $\|X\|_2$ and $\|\dot{X}\|_2$ are bounded independently of ϵ, and [9, 147]

$$[\dot{Q}, Q] = [H, X].$$

Then

$$\dot{w} = -\mathcal{T}_B^{-1}[H, X]\mathcal{T}_B w = -(\mathcal{T}_B^{-1}H)X\mathcal{T}_B w + \mathcal{T}_B^{-1}X(H\mathcal{T}_B)w. \tag{2.4.11}$$

To compute the first part of Eq. (2.4.11), we first take the time derivative of the identity $I = \mathcal{T}_B^{-1}\mathcal{T}_B$ and get

$$\mathcal{T}_B^{-1}H = -\mathrm{i}\epsilon\partial_t(\mathcal{T}_B^{-1}) + (\varphi^* H\varphi)\mathcal{T}_B^{-1} - \mathrm{i}\epsilon\mathcal{T}_B^{-1}[\dot{Q}, Q]. \tag{2.4.12}$$

Then the first part of Eq. (2.4.11) can be rewritten as

$$-(\mathcal{T}_B^{-1}H)X\mathcal{T}_B w = \mathrm{i}\epsilon\partial_t(\mathcal{T}_B^{-1})X\mathcal{T}_B w + \mathrm{i}\epsilon\mathcal{T}_B^{-1}[\dot{Q}, Q]X\mathcal{T}_B w - (\varphi^* H\varphi)\mathcal{T}_B^{-1}X\mathcal{T}_B w. \tag{2.4.13}$$

To compute the second part of Eq. (2.4.11), rewrite Eq. (2.4.6) as

$$H\mathcal{T}_B = \mathrm{i}\epsilon\dot{\mathcal{T}}_B + (\varphi^* H\varphi)\mathcal{T}_B - \mathrm{i}\epsilon[\dot{Q}, Q]\mathcal{T}_B, \tag{2.4.14}$$

and then

$$\mathcal{T}_B^{-1}X(H\mathcal{T}_B)w = \mathrm{i}\epsilon\mathcal{T}_B^{-1}X\dot{\mathcal{T}}_B w - \mathrm{i}\epsilon\mathcal{T}_B^{-1}X[\dot{Q}, Q]\mathcal{T}_B w + (\varphi^* H\varphi)\mathcal{T}_B^{-1}X\mathcal{T}_B w. \tag{2.4.15}$$

Sum up Eq. (2.4.13) and (2.4.15), then Eq. (2.4.11) becomes

$$\dot{w} = \mathrm{i}\epsilon(\partial_t(\mathcal{T}_B^{-1})X\mathcal{T}_B + \mathcal{T}_B^{-1}X\dot{\mathcal{T}}_B)w + \mathrm{i}\epsilon\mathcal{T}_B^{-1}[[\dot{Q}, Q], X]\mathcal{T}_B w. \tag{2.4.16}$$

In Eq. (2.4.16), the second term of the right hand side is already of $\mathcal{O}(\epsilon)$. Now we turn to the first term to treat the derivatives $\partial_t(\mathcal{T}_B^{-1})$ and $\dot{\mathcal{T}}_B$. By repeated usage of the Leibniz rule, Eq. (2.4.16) becomes

$$\begin{aligned}
\dot{w} &= \mathrm{i}\epsilon\partial_t(\mathcal{T}_B^{-1}X\mathcal{T}_B)w - \mathrm{i}\epsilon\mathcal{T}_B^{-1}\dot{X}\mathcal{T}_B w + \mathrm{i}\epsilon\mathcal{T}_B^{-1}[[\dot{Q}, Q], X]\mathcal{T}_B w \\
&= \mathrm{i}\epsilon\partial_t(\mathcal{T}_B^{-1}X\mathcal{T}_B w) - \mathrm{i}\epsilon\mathcal{T}_B^{-1}X\mathcal{T}_B\dot{w} - \mathrm{i}\epsilon\mathcal{T}_B^{-1}\dot{X}\mathcal{T}_B w + \mathrm{i}\epsilon\mathcal{T}_B^{-1}[[\dot{Q}, Q], X]\mathcal{T}_B w \\
&= \mathrm{i}\epsilon\partial_t(\mathcal{T}_B^{-1}X\varphi) + \mathrm{i}\epsilon\mathcal{T}_B^{-1}X[H, X]\varphi - \mathrm{i}\epsilon\mathcal{T}_B^{-1}\dot{X}\varphi + \mathrm{i}\epsilon\mathcal{T}_B^{-1}[[\dot{Q}, Q], X]\varphi.
\end{aligned} \tag{2.4.17}$$

In the last equation we use again Eq. (2.4.10). Substitute Eq. (2.4.17) back to Eq. (2.4.9), we get

$$\begin{aligned}
\|\varphi(t) - \varphi_B(t)\|_2 &= \| \int_0^t \dot{w}(s)ds\|_2 \\
&\leq \epsilon\|(\mathcal{T}_B^{-1}X\varphi)(t) - (\mathcal{T}_B^{-1}X\varphi)(0)\|_2 \\
&\quad + \epsilon\left\| \int_0^t \left(\mathcal{T}_B^{-1}X[H, X]\varphi - \mathcal{T}_B^{-1}\dot{X}\varphi + \mathcal{T}_B^{-1}\big[[\dot{Q}, Q], X\big]\varphi\right)ds \right\|_2 \\
&= \mathcal{O}(\epsilon).
\end{aligned} \tag{2.4.18}$$

Therefore there exists $\eta(t)$ such that

$$\varphi(t) = \varphi_B(t) + \epsilon\eta(t), \tag{2.4.19}$$

where $\|\eta(t)\|_2$ is bounded independently of ϵ. The differentiability of $\eta(t)$ follows directly from that of $\varphi(t)$ and $\varphi_B(t)$.

3. Comparing Eq. (2.4.19) with our goal, the only thing that we need to prove is that the distance between φ_B and φ_A is also $\mathcal{O}(\epsilon)$. Note that φ_A can be written as [58]

$$\varphi_A(t) = \mathfrak{T}\left[\exp\left(\int_0^t [\dot{Q}(s), Q(s)]ds\right)\right]\varphi_A(0), \tag{2.4.20}$$

where $\mathfrak{T}$ is the time ordering operator due to the explicit time dependence of Q. Using the power series representation, the time-ordered exponential is defined as

$$\mathfrak{T}\left[e^{\int_0^t A(s)\,\mathrm{d}s}\right] = I + \int_0^t A(s)\,\mathrm{d}s + \frac{1}{2!}\int_0^t\int_0^t \mathfrak{T}[A(s_1)A(s_2)]\,\mathrm{d}s_1\,\mathrm{d}s_2 + \cdots, \tag{2.4.21}$$

where the time-ordered product of two matrices $\mathfrak{T}[A(s_1)A(s_2)]$ is given by

$$\mathfrak{T}[A(s_1)A(s_2)] = \begin{cases} A(s_1)A(s_2), & s_1 \geqslant s_2; \\ A(s_2)A(s_1), & s_1 < s_2. \end{cases} \tag{2.4.22}$$

Using Duhamel's principle, we have from Eq. (2.4.4) and (2.4.7)

$$\varphi_B(t) = \varphi_A(t) + \int_0^t \mathfrak{T}\left[\exp\left(\int_s^t [\dot{Q}(s'), Q(s')]ds'\right)\right]\cdot\frac{1}{\mathrm{i}\epsilon}(H(s) - \varphi^*(s)H(s)\varphi(s))\varphi_B(s)ds \tag{2.4.23}$$

By Eq. (2.4.8), (2.4.19), and the normalization condition of φ and φ_B,

$$\begin{aligned}(H - \varphi^* H\varphi)\varphi_B &= -\lambda(\epsilon\eta^*\varphi_B + \epsilon\varphi_B^*\eta)\varphi_B - \epsilon^2(\eta^* H\eta)\varphi_B \\ &= -\lambda[(\varphi_B + \epsilon\eta)^*(\varphi_B + \epsilon\eta) - \varphi_B^*\varphi_B - \epsilon^2\eta^*\eta]\varphi_B - \epsilon^2(\eta^* H\eta)\varphi_B \\ &= \epsilon^2\lambda(\eta^*\eta)\varphi_B - \epsilon^2(\eta^* H\eta)\varphi_B \\ &= \mathcal{O}(\epsilon^2).\end{aligned} \tag{2.4.24}$$

Hence Eq. (2.4.23) implies

$$\varphi_B - \varphi_A = \mathcal{O}(\epsilon). \tag{2.4.25}$$

Therefore, $\varphi_R := \eta + (\varphi_B - \varphi_A)/\epsilon$ is infinitely differentiable, and $\|\varphi_R(t)\|_2$ is bounded independently of ϵ. This proves the decomposition of the solution to the PT dynamics

$$\varphi = \varphi_B + \epsilon\eta = \varphi_B + \epsilon\varphi_R - (\varphi_B - \varphi_A) = \varphi_A + \epsilon\varphi_R. \tag{2.4.26}$$

□

Theorem 18 gives a decomposition near the adiabatic regime with respect to the PT wave function. As a corollary, we also have the adiabatic theorem with respect to the projector.

Corollary 19. *For the projector $P(t)$, there exists an infinitely differentiable matrix-valued function $R(t)$ such that*

$$P(t) = Q(t) + \epsilon R(t) \tag{2.4.27}$$

holds for all t up to $T = \mathcal{O}(1)$, where $\|R(t)\|_2$ is bounded independently of ϵ.

Proof. This follows directly from theorem 18

$$\begin{aligned} P = \varphi\varphi^* &= (\varphi_A + \epsilon\varphi_R)(\varphi_A + \epsilon\varphi_R)^* \\ &= Q + \epsilon(\varphi_R\varphi_A^* + \varphi_A\varphi_R^* + \epsilon\varphi_R\varphi_R^*). \end{aligned} \tag{2.4.28}$$

□

Remark 20. *The adiabatic theorem for the Schrödinger wave function $\psi(t)$ has been well established in the literature e.g. [93, 9, 147], where the decomposition takes the form $\psi = \psi_A + \epsilon\tilde{\psi}_R$, and the adiabatic evolution ψ_A satisfies*

$$\mathrm{i}\epsilon\partial_t\psi_A = (H + \mathrm{i}\epsilon[\dot{Q}, Q])\psi_A. \tag{2.4.29}$$

We compare our result with previous well-established ones from two aspects. First, there is an important difference between the PT eigenfunction φ_A, governed by Eq. (2.4.4)*, and the standard one ψ_A, governed by Eq.* (2.4.29)*. Although both φ_A and ψ_A are eigenfunctions of $H(t)$, their phase factors are different, resulting in different oscillatory behavior. More specifically, the standard wavefunction ψ_A oscillates on the scale of $\mathcal{O}(\epsilon^{-1})$ since (at least intuitively) Eq.* (2.4.29) *is just a small perturbation of the original Schrödinger dynamics. The PT eigenfunction φ_A does not depend on ϵ, and thus oscillates on the scale of $\mathcal{O}(1)$. When projected to the eigenspace, the PT dynamics leads to the optimal phase factor, and this verifies the effectiveness of the definition of PT (to minimize unnecessary oscillations) and provides another theoretical explanation of the performance shown in Fig. 2.2.1a. Second, our proof largely follows the existing works of the adiabatic theorem [93, 9, 147]. Our main modification is to address the special non-linear term in the PT dynamics, even though the original Schrödinger dynamics is linear.*

Remark 21. *As mentioned at the end of step 1, φ_B is also an eigenstate, and Eq.* (2.4.19) *indeed leads to another version of the adiabatic theorem, but with notable differences from the decomposition in Theorem 18. First, the definition of φ_B still relies on the information of φ, and thus is not a self-contained equation. Second, the norms of the derivatives of φ_B still depend on ϵ (more precisely one can prove that $\|\varphi_B^{(k)}\|_2 \sim \mathcal{O}(1/\epsilon^{k-2})$ for $k \geq 3$), which indicates that the gauge choice of φ_B is not optimal either.*

Local truncation error

In this section, we show that after time discretization, the local truncation error of the discretized PT dynamics improves by one order in terms of ϵ compared to that of the discretized Schrödinger dynamics in the near adiabatic regime. This is given in Lemma 22.

For simplicity we will focus on the numerical integrators in the classes of Runge-Kutta methods and linear multistep methods, both of which are widely used for simulating the Schrödinger equation. We will refer numerical integrator to either a Runge-Kutta method or a linear multistep method in our context. Recall that a numerical integrator with a given time step h, denoted by I_h, can be generally written as

$$u_{n+1} = I_h(u_n, \cdots, u_{n-l}), \tag{2.4.30}$$

for some integer $l \geqslant 0$, and u_n is the numerical approximation to the true solution $u(t_n)$. If I_h is of order k, then the local truncation error at step $n+1$, defined as

$$L_{n+1} = I_h(u(t_n), \cdots, u(t_{n-l})) - u_{n+1},$$

should satisfy

$$\|L_{n+1}\|_2 \leqslant Ch^{k+1}\|u^{(k+1)}(\xi_{n+1})\|_2,$$

for some $\xi_{n+1} \in [t_n, t_{n+1}]$. When applied to the Schrödinger dynamics, the PT dynamics, or the associated Hamiltonian form, we may identify u with ψ, φ, or the equivalent (q,p) representation.

Lemma 22. *Apply a numerical integrator of order k to the Schrödinger dynamics or its Hamiltonian form* (2.2.19)*. Then the local truncation error is bounded by Ch^{k+1}/ϵ^r up to the time $T \sim \mathcal{O}(1)$, with $r = k+1$ and C is a constant independent of h and ϵ. The same result holds for the PT dynamics* (2.2.8) *or its corresponding Hamiltonian form* (2.2.23) *with $r = k$.*

Proof. It is sufficient to show that the derivatives satisfy $\|\psi^{(k+1)}\|_2 \leq \mathcal{O}(1/\epsilon^{k+1})$, and $\|\varphi^{(k+1)}\|_2 \leq \mathcal{O}(1/\epsilon^k)$ for any $k \geqslant 0$. This can be proved by induction.

1. For ψ, the case $k = 0$ directly follows from Eq. (2.1.1). Assume the estimate holds for all the integers smaller than k, differentiate the Schrödinger equation k times and we get

$$\psi^{(k+1)} = \frac{1}{\mathrm{i}\epsilon}\sum_{j=0}^{k}\binom{k}{j}H^{(k-j)}\psi^{(j)}. \tag{2.4.31}$$

By the induction and the assumption 1,

$$\|\psi^{(k+1)}\|_2 \leq \frac{C}{\epsilon}\sum_{j=0}^{k}\binom{k}{j}\frac{1}{\epsilon^j} \sim \mathcal{O}(\epsilon^{-(k+1)}). \tag{2.4.32}$$

2. For φ, we first study the derivatives of P, and then use the PT condition (2.2.6) to obtain the derivatives of φ.

By Corollary 19, the von Neumann equation (2.1.5) and the identity $HQ = QH$, the first order derivative of P satisfies

$$\|\dot{P}\|_2 = \frac{1}{\epsilon}\|HP - PH\|_2 = \|HR - RH\|_2 \leq \mathcal{O}(1).$$

Furthermore, differentiate the von Neumann equation (2.1.5) k times, we get

$$P^{(k+1)} = \frac{1}{\mathrm{i}\epsilon}\sum_{j=0}^{k}\binom{k}{j}[H^{(j)}, P^{(k-j)}], \tag{2.4.33}$$

from which we can show by induction that

$$\|P^{(k+1)}\|_2 \leq \mathcal{O}(\epsilon^{-k}). \tag{2.4.34}$$

Now use the PT condition $P\dot{\varphi} = 0$, we find for $k = 0$,

$$\dot{\varphi} = \partial_t(P\varphi) = \dot{P}\varphi \leq \mathcal{O}(1). \tag{2.4.35}$$

Furthermore,

$$\varphi^{(k+1)} = \sum_{j=0}^{k}\binom{k}{j}P^{(j+1)}\varphi^{(k-j)}, \tag{2.4.36}$$

from which we can prove by induction and Eq. (2.4.34) that

$$\varphi^{(k+1)} \leq \mathcal{O}(\epsilon^{-k}). \tag{2.4.37}$$

□

Global error

The analysis of the local truncation error directly extends to the global error up to $T \sim \mathcal{O}(\epsilon)$, following the classical stability analysis. However, the Lipschitz constants corresponding to the right hand side of the Schrödinger dynamics and the PT dynamics are generally $\mathcal{O}(1/\epsilon)$, which leads to an exponentially growing factor $\exp(T/\epsilon)$ in the global error bounds. Hence we cannot directly obtain the global error estimate up to $\mathcal{O}(1)$ time.

However, if we adopt the Hamiltonian formulation of the dynamics and employ a symplectic integrator, we can indeed obtain long time error estimates. This is stated in Theorem 23, of which the proof directly follows from Lemma 22 and Theorem X.3.1 in [68].

Theorem 23. *Apply a symplectic integrator of order k to the Hamiltonian system* (2.2.19) *and* (2.2.23)*, then there exist constants c, C, independent of h and ϵ, such that for the time step $h \leq c\epsilon$, the numerical solutions up to the time $T \sim \mathcal{O}(1)$ satisfy*

$$\|(q_n, p_n) - (q(t), p(t))\|_2 \leq C\frac{h^k}{\epsilon^r}. \tag{2.4.38}$$

Here $r = k + 1$ for the Schrödinger dynamics (2.2.19) *and $r = k$ for the PT dynamics* (2.2.23)*.*

Remark 24. *In Theorem X.3.1 in [68], all terms are bounded by $\mathcal{O}(1)$ terms and there is no ϵ dependence. In order to adapt its proof to the current situation, we observe the key fact in Theorem X.3.1 in [68] that the global error of a symplectic integrator accumulates linearly in time with no exponential growing factor. Therefore the local truncation error which is $\mathcal{O}(h^{k+1}/\epsilon^r)$ directly sums up linearly to the global error of $\mathcal{O}(h^k/\epsilon^r)$.*

Remark 25. *The nontrivial restriction on the time step size $h \leq c\epsilon$ is because Theorem X.3.1 in [68] holds only for sufficiently small time steps. In general, h must be no larger than c/L where L is the Lipschitz constant of the right hand side of the Hamiltonian system, and is $\mathcal{O}(1/\epsilon)$ in the singularly perturbed regime. Nonetheless, numerical results in Section 2.5 indicate that the PT dynamics may admit a considerably larger time step in practice.*

Remark 26. *When a symplectic integrator is used, Theorem 23 is directly applicable to the Schrödinger dynamics. However, the PT dynamics* (2.2.8) *and the Hamiltonian system* (2.2.23) *share the same exact solution, but lead to different numerical schemes even when the same integrator is used. Despite such difference, numerical results in Section 2.5 indicate that the symplectic integrators, and even certain non-symplectic schemes, can still perform very well in the PT dynamics* (2.2.8)*.*

Remark 27. *Theorem 23 also indicates that the PT dynamics is relatively more effective when combined with low order methods. For instance, if we would like to achieve some desired accuracy δ (assuming δ is sufficiently small), then for the Schrödinger dynamics, we should choose the time step size to be*

$$h \sim \mathcal{O}(\delta^{\frac{1}{k}}\epsilon^{1+\frac{1}{k}}).$$

For the PT dynamics, we should choose

$$h \sim \mathcal{O}(\delta^{\frac{1}{k}}\epsilon).$$

From this perspective, the gain of the PT dynamics is less significant when k is large.

2.5 Numerical results

In this section we study the effectiveness of the PT dynamics using two examples. The first one is a toy example, which is a linear Schrödinger equation in $\mathbb{C}^2$. This example gives a clear illustration of the performance of different numerical methods near and beyond the adiabatic regime. The second example is a nonlinear Schrödinger equation in a one-dimensional space, where we also compare the computational cost between the propagation of the Schrödinger dynamics and the PT dynamics.

The test programs are written in MATLAB. All calculations are carried out using the BRC High Performance Computing service. Each node consists of two Intel Xeon 10-core Ivy Bridge processors (20 cores per node) and 64 gigabyte (GB) of memory. We use the Anderson mixing for solving all the nonlinear fixed point problems, including those in the PT dynamics and the nonlinear Schrödinger equation. Here no preconditioner is used for the tests.

A toy example

First we present a linear example, in which $H(t)$ is chosen to be

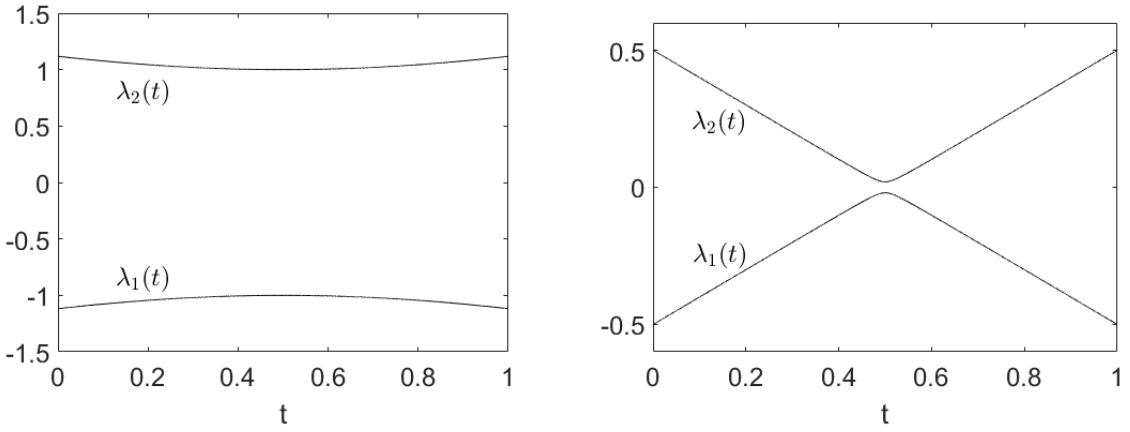
$$H(t) = \begin{pmatrix} t - t_0 & \delta \\ \delta & -(t - t_0) \end{pmatrix}. \tag{2.5.1}$$

Here $H(t)$ has the eigenvalues $\lambda_{1,2}(t) = \mp\sqrt{(t-t_0)^2 + \delta^2}$, where $\delta > 0$ ensures the gap condition and controls the size of the gap. When δ is large, the dynamics stays closer to the adiabatic regime, while the dynamics can go beyond the adiabatic regime with a smaller δ (see Fig. 2.5.1). The initial value is always chosen to be the normalized eigenvector of $H(0)$ corresponding to $\lambda_1(0) = -\sqrt{t_0^2 + \delta^2}$. We propagate the wave functions up to $T = 1$. For the choices of the parameters in the Anderson Mixing in propagating PT dynamics, the step length $\alpha = 1$, the mixing dimension is 20, and the tolerance is 10^{-8}.

Near adiabatic regime

First we consider the near adiabatic case with $\delta = 1$. We compare the following numerical methods:

- S-RK4: fourth order Runge-Kutta method (RK4) applied to the Schrödinger equation (2.1.1)
- PT-RK4: fourth order Runge-Kutta method (RK4) applied to the PT dynamics (2.2.8)
- S-GL2: implicit midpoint rule (GL2) applied to the Schrödinger equation (2.1.1)



(a) $\delta = 1$ (b) $\delta = 0.02$

Figure 2.5.1: Eigenvalues of $H(t)$ in the toy example with $t_0 = 0.5$ and two choices of δ.

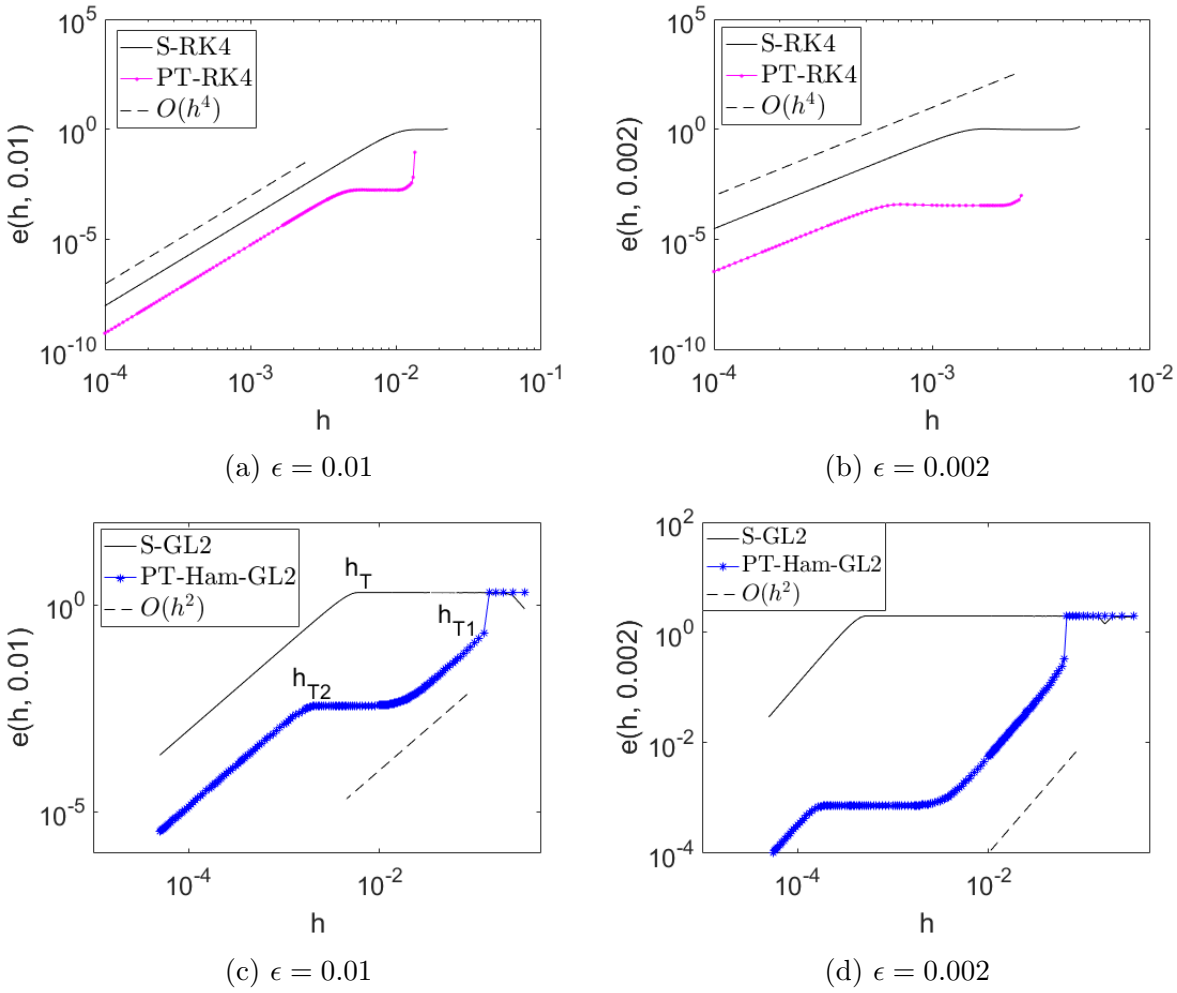
- PT-Ham-GL2: implicit midpoint rule (GL2) applied to the PT Hamiltonian system (2.2.23)
- PT-GL2: implicit midpoint rule (GL2) applied to the PT dynamics (2.2.8)
- PT-CN: trapezoidal rule (or the Crank-Nicolson method, CN) applied to the PT dynamics (2.2.8)

Fig. 2.5.2 compares the performance of different methods for this toy example. The numerical error is computed by

$$\mathsf{e}(h,\epsilon) = \max_{n \text{ s.t. } nh\in[0,T]} \|u_n - u(t_n)\|_2$$

where u denotes ψ for the Schrödinger dynamics, φ for the PT dynamics and (q,p) for the Hamiltonian systems, respectively.

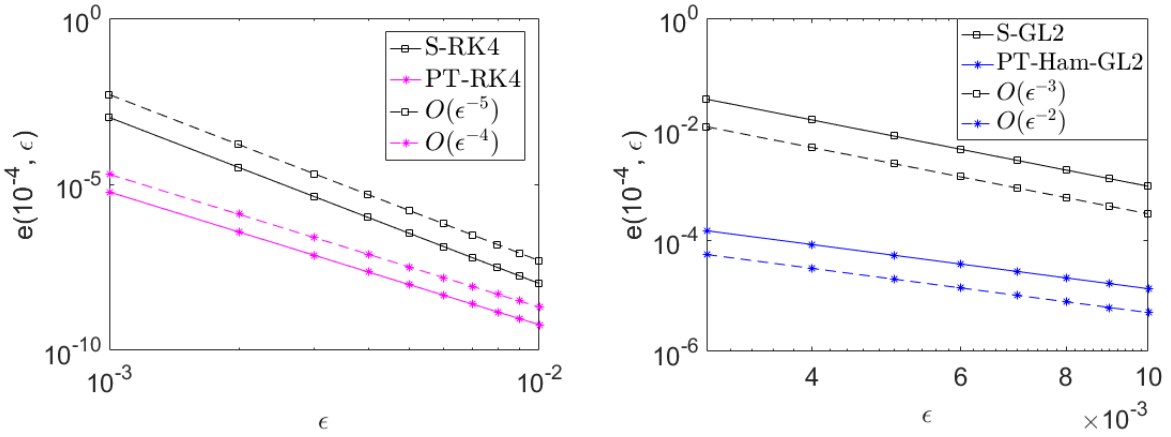
We first consider the explicit numerical methods. Fig. 2.5.2a and 2.5.2b give a comparison between S-RK4 and PT-RK4. Not surprisingly, as an explicit method, RK4 is numerically unstable for large time steps under both cases, and achieves fourth order convergence for small time steps. Furthermore, when h is small enough, $\mathsf{e}(h,\epsilon)$ of the PT dynamics is smaller than that of the Schrödinger dynamics. Fig. 2.5.3a presents a study on how $\mathsf{e}(h,\epsilon)$ depends on ϵ, which reveals that by propagating the PT dynamics we gain one extra order of accuracy in terms of ϵ. This agrees with the theoretical results in Section 2.4.



(a) $\epsilon = 0.01$ (b) $\epsilon = 0.002$

(c) $\epsilon = 0.01$ (d) $\epsilon = 0.002$

Figure 2.5.2: Numerical errors of different numerical methods in the near adiabatic regime of the toy example. (a)(b) compare S-RK4 and PT-RK4 for $\epsilon = 0.01, 0.002$, respectively. (c)(d) compare S-GL2 and PT-Ham-GL2 for $\epsilon = 0.01, 0.002$, respectively.



(a) (b)

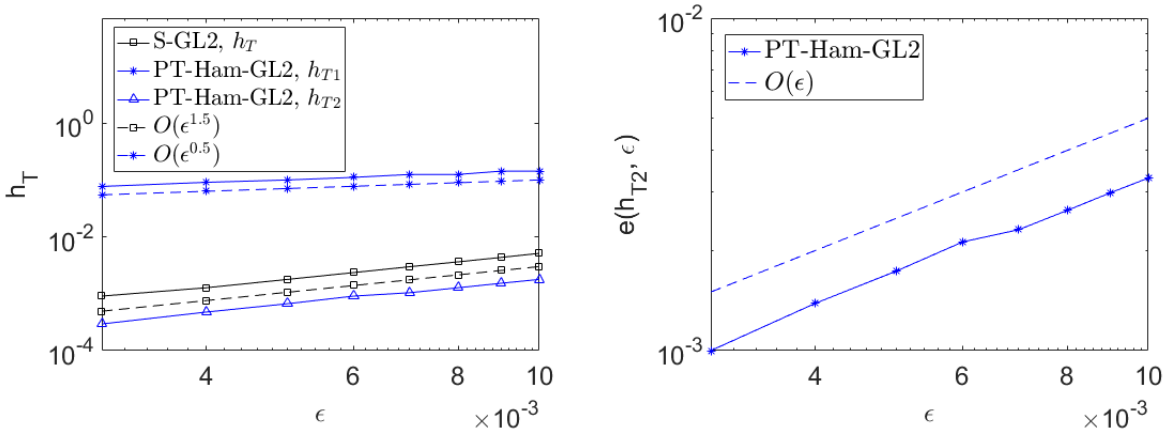
Figure 2.5.3: Relationship between the asymptotic errors and ϵ in the near adiabatic regime of the toy example. Here we fix the time step size to be $h = 10^{-4}$ for both (a)(b).

Next we test GL2 as an example of implicit symplectic methods applied to the Hamiltonian systems. Fig. 2.5.2c compares the numerical performances of S-GL2 and PT-Ham-GL2. For small h, we observe a smaller error using the PT formulation, *i.e.* $\mathsf{e}(h,\epsilon)$ of S-GL2 is $\mathcal{O}(h^2/\epsilon^3)$ and $\mathsf{e}(h,\epsilon)$ of PT-Ham-GL2 is $\mathcal{O}(h^2/\epsilon^2)$ (see Fig. 2.5.3b for a study on the ϵ dependence). This verifies the estimate in Theorem 23. Despite that GL2 is a numerically stable scheme with a large time step, the step size of S-GL2 is constrained by the requirement of the accuracy, while the step size of PT-Ham-GL2 can be chosen to be considerably larger.

More specifically, let us define the "turning point" h_T to be the largest time step size when a scheme starts to converge. Numerically for second order schemes the turning point can be computed as

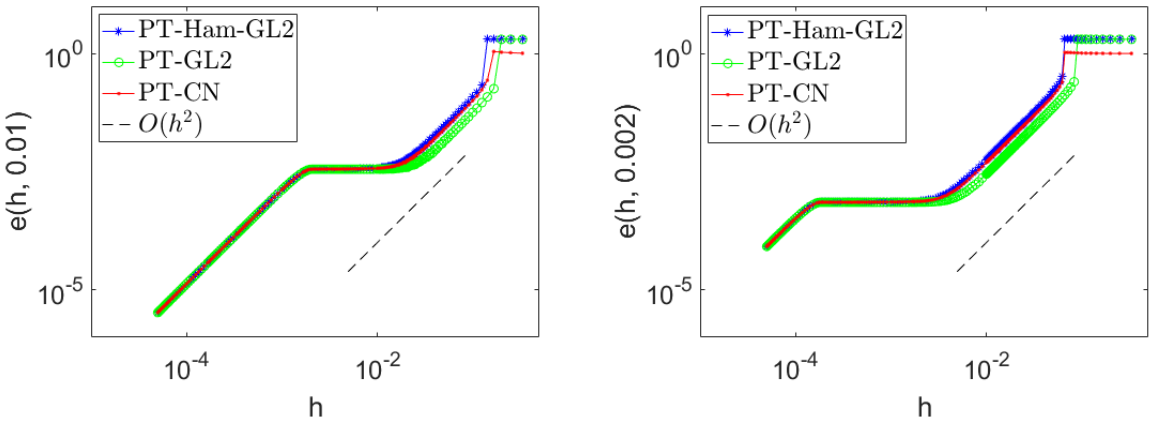
$$h_T = \arg\max \left\{ h \in [h_1, h_2] : \frac{\partial(\log \mathsf{e})}{\partial(\log h)} > 1 \right\}$$

where $[h_1, h_2]$ is a suitable interval containing the convergence interval of interests. In Fig. 2.5.2c we mark the turning points in S-GL2 and PT-Ham-GL2, and study their dependence on ϵ in Fig. 2.5.4a. For S-GL2, the convergence starts at $h_T = \mathcal{O}(\epsilon^{3/2})$. For PT-Ham-GL2, a two-stage convergence behavior is observed. As h decreases, the scheme first starts to converge with second order at a relatively large time step $h_{T1} = \mathcal{O}(\epsilon^{1/2})$. This first stage ends at $h = \mathcal{O}(\epsilon)$ when $\mathsf{e}(h,\epsilon)$ reaches a plateau with its magnitude being $\mathcal{O}(\epsilon)$ (see Fig. 2.5.4b). Then the second-stage convergence starts at $h_{T2} = \mathcal{O}(\epsilon^{3/2})$.



(a) (b)

Figure 2.5.4: (a) Relationship between the turning points and ϵ in S-GL2 and PT-Ham-GL2 in the near adiabatic regime of the toy example. (b) Relationship between the magnitude of the plateau of the numerical error and ϵ in PT-Ham-GL2.



(a) $\epsilon = 0.01$ (b) $\epsilon = 0.002$

Figure 2.5.5: Performance of PT-Ham-GL2, PT-GL2 and PT-CN in the near adiabatic regime of the toy example.

In the end we compare the schemes PT-Ham-GL2, PT-GL2 and PT-CN. Although we only justified the behavior of the global error for PT-Ham-GL2, numerical results in Fig. 2.5.5a and 2.5.5b indicate that there is no essential difference among these methods in practice.

Beyond adiabatic regime

As the value of δ is reduced, the second eigenstate corresponding to λ_2 may contribute significantly to the wave function, which leads to the violation of the adiabatic regime.

Fig. 2.5.6 investigates the Schrödinger wave function and the PT wave function with $\epsilon = 0.002, \delta = 0.05$. Fig. 2.5.6a and 2.5.6b compare the real parts of the Schrödinger wave function and the PT wave function. When $t < t_0 = 0.5$, the system stays close to the adiabatic regime and the PT wave function is nearly flat. However, when $t > t_0$, the PT wave function starts to oscillate as well. Fig. 2.5.6c shows an orthogonal decomposition of the PT wave function into two orthogonal eigenspaces. Fig. 2.5.6d shows the evolution of the probability that the eigenstate corresponding to $\lambda_2(t)$ is occupied, which can be computed as $|c_2|^2 = |(\varphi(t), e_2(t))|^2$ and $e_2(t)$ is the normalized eigenstate of $H(t)$ corresponding to $\lambda_2(t)$. These results confirm that the oscillatory behavior originates from the excited state corresponding to λ_2.

As discussed before, such oscillatory nature in the wave functions may increase the computational difficulty and require a smaller time step even for the PT dynamics. Fig. (2.5.7) compares $\mathsf{e}(h, \epsilon)$ for S-GL2, PT-Ham-GL2, PT-GL2 and PT-CN respectively. The results confirm that the PT dynamics is always more accurate than the Schrödinger dynamics using the same step size, but the gain becomes smaller as δ decreases.

Nonlinear Schrödinger equation in one dimension

Next we study the performance of the PT dynamics in a singularly perturbed nonlinear Schrödinger equation in one dimension.

$$\begin{aligned} \mathrm{i}\epsilon\partial_t\psi(x,t) &= -\frac{1}{2}\partial_x^2\psi(x,t) + V(x,t)\psi(x,t) + g|\psi(x,t)|^2\psi(x,t), \quad x \in [0, L] \\ \psi(x,0) &= \psi_0(x) \\ \psi(0,t) &= \psi(L,t). \end{aligned} \tag{2.5.2}$$

We set $L = 50$, and the external potential is chosen to be a time-dependent Gaussian function modeling a moving potential well (Fig. 2.5.8)

$$V(x,t) = -\exp(-0.1(x - R(t))^2) \tag{2.5.3}$$

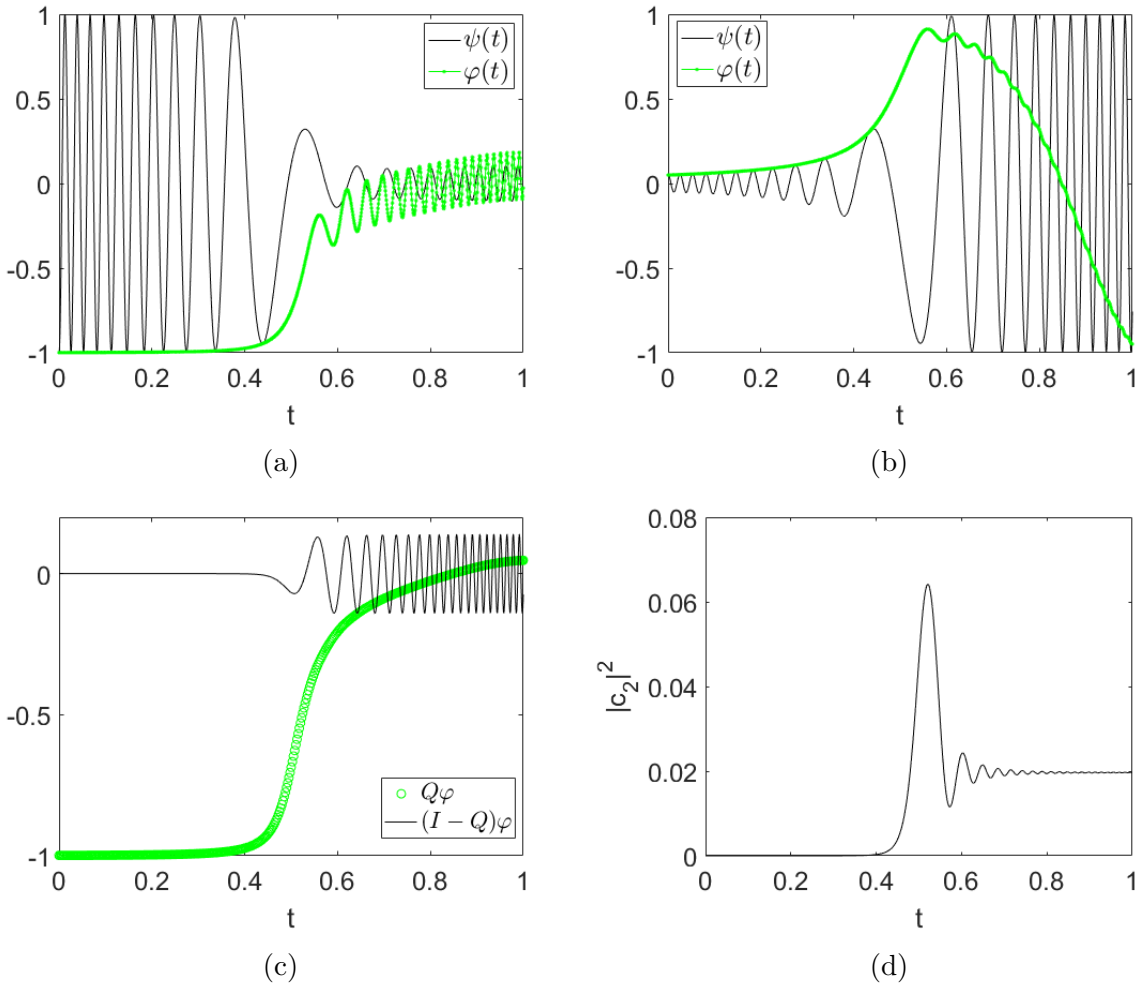
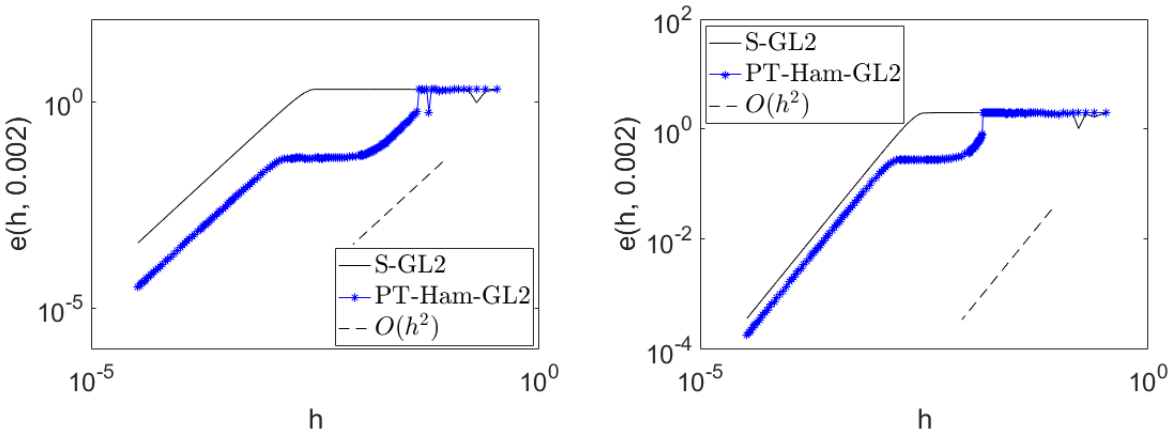
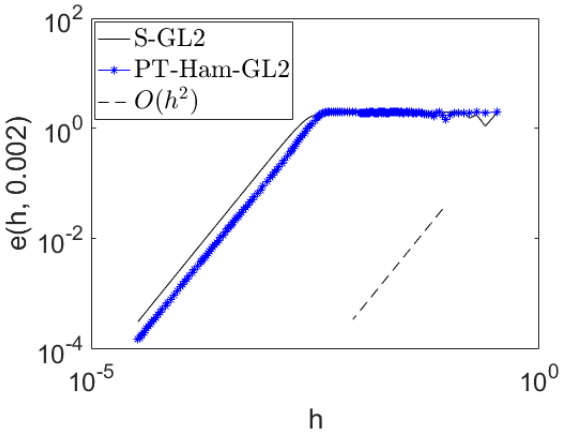


Figure 2.5.6: The Schrödinger and the PT wave functions beyond the adiabatic regime in the toy example. In all sub-figures, parameters are chosen to be $\epsilon = 0.002, \delta = 0.05$, and the solutions are obtained by GL2 with the time step $h = 10^{-6}$. (a)(b) show the first and second entry of the real part of the Schrödinger wave function and the PT wave function, respectively. (c) shows a decomposition of the PT wave function into the two orthogonal eigenspaces (in the sub-figure we only present the real part of the first entry). (d) shows the time evolution of the probability that the eigenstate corresponding to λ_2 is occupied.

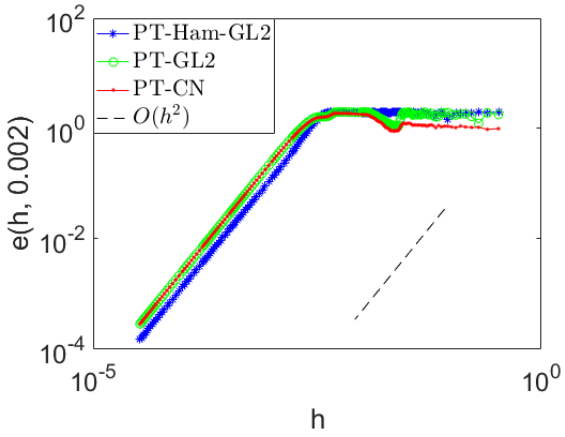


(a) $\delta = 0.07$

(b) $\delta = 0.05$

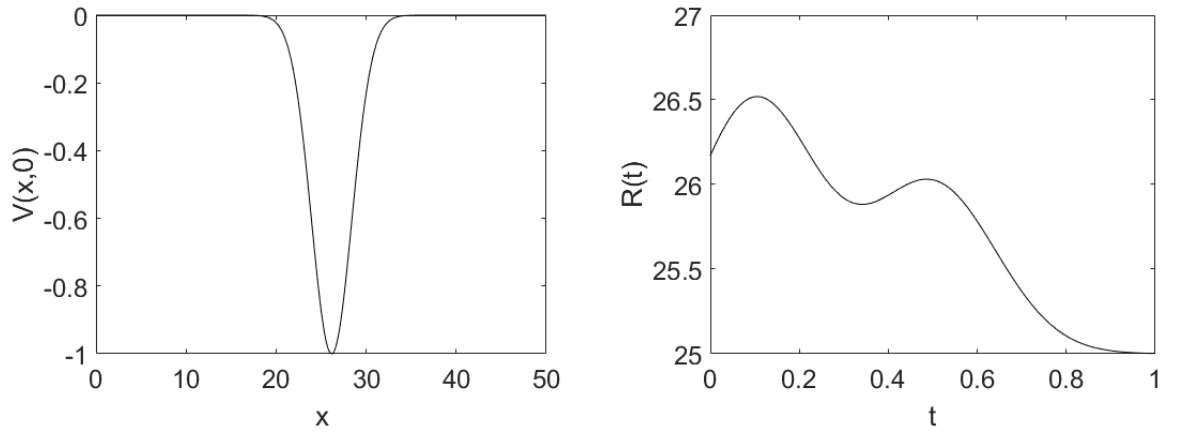


(c) $\delta = 0.03$



(d) $\delta = 0.03$

Figure 2.5.7: Numerical errors of different numerical methods beyond the adiabatic regime in the toy example. In all sub-figures $\epsilon = 0.002$. (a)(b)(c) compare the numerical performances between S-GL2 and PT-Ham-GL2 for $\delta = 0.07, 0.05, 0.03$, respectively. (d) gives a comparison of PT-Ham-GL2, PT-GL2 and PT-CN with $\delta = 0.03$.



(a) (b)

Figure 2.5.8: External potential and the time-dependent center for the nonlinear Schrödinger equation.

with a time-dependent center

$$R(t) = 25 + 1.5\exp(-25(t-0.1)^2) + \exp(-25(t-0.5)^2). \tag{2.5.4}$$

Note that $R(t)$ varies on the $\mathcal{O}(1)$ time scale.

We use equidistant nodes $x_k = kh_x$ and the second-order finite difference scheme for spacial discretization, and we fix $h_x = 0.025$. Other parameters in this example are chosen to be $g = 2.5, T = 1, \epsilon = 0.0025$. For the choices of the parameters in the Anderson Mixing, the step length $\alpha = 1$, the mixing dimension is 20, and the tolerance is 10^{-8}. Fig. 2.5.9 compares $\mathsf{e}(h, \epsilon)$ of S-GL2, PT-Ham-GL2, PT-GL2 and PT-CN, and confirms the same numerical behavior as in the toy example.

Next we study the computational cost by comparing the total number of the Anderson mixing steps versus the numerical error $\mathsf{e}(h, \epsilon)$ up to $T = 1$. Fig. 2.5.10 clearly demonstrates that in order to achieve the same level of accuracy, all the methods propagating the PT dynamics, including PT-Ham-GL2, PT-GL2 and PT-CN, are much more efficient than S-GL2. This is valid across the entire range of the step sizes under study.

2.6 Conclusion

Quantum dynamics can be equivalently written in terms of the Schrödinger equation for the wave function, and the von Neumann equation for the density matrix. However, the

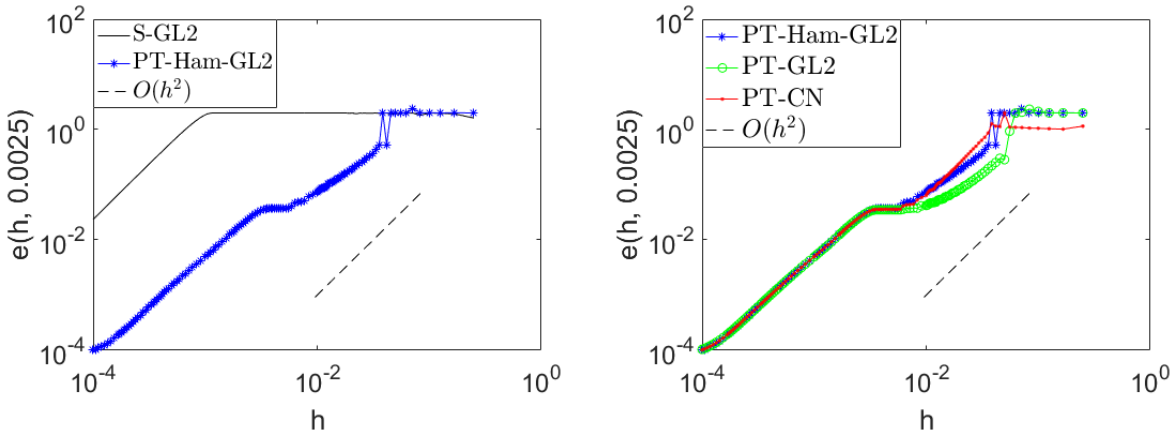


Figure 2.5.9: Numerical errors of different numerical methods in the example of the nonlinear Schrödinger equation. Parameters are chosen to be $T = 1, \epsilon = 0.0025$. (a) compares S-GL2 and PT-Ham-GL2. (b) compares PT-Ham-GL2, PT-GL2 and PT-CN.

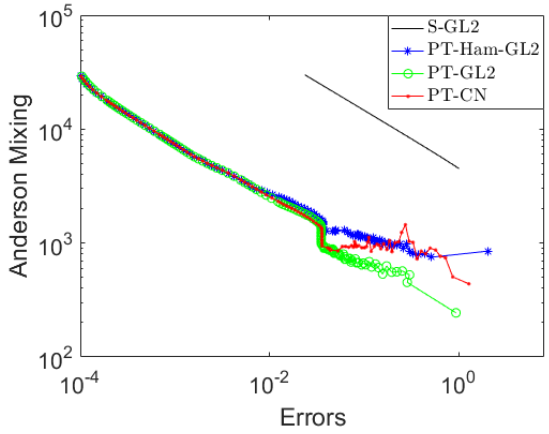


Figure 2.5.10: Total numbers of the Anderson mixing versus the numerical error.

Schrödinger dynamics may require a very small time step in numerical simulation due to the non-optimal gauge choice. In this chapter, we propose to close this gap by identifying the optimal gauge choice, which is obtained from the parallel transport formulation. The solution of the resulting parallel transport (PT) dynamics can be significantly less oscillatory to that of the Schrödinger dynamics, especially in the near adiabatic regime. The PT dynamics is suitable to be combined with implicit time integrators, which allows the usage of large time steps even when the spectral radius of the Hamiltonian is large, and/or when ϵ is small. Although our global error analysis only applies to the Hamiltonian form of the PT dynamics with symplectic integrators and a relatively small time step, our numerical results indicate that the PT dynamics can be effectively discretized with more general numerical schemes and with much larger time steps.

Chapter 3

Parallel transport dynamics for TDDFT

3.1 Introduction

[1]One of the most widely used techniques for studying ultrafast properties is the real-time time-dependent density functional theory (RT-TDDFT) [133, 126], which has achieved successes in a number of fields including *e.g.* nonlinear optical response [146] and the collision of an ion with a substrate [98]. In TDDFT, the system is described by a set of wave functions $\Psi(t) = \{\psi_j(t)\}$ satisfying the time-dependent Schrödinger equation

$$\mathrm{i}\partial_t \Psi(t) = H(t, P)\Psi(t), \quad P(t) = \Psi(t)\Psi^*(t), \tag{3.1.1}$$

and the TDDFT Hamiltonian takes the form

$$H(t, P) = -\frac{1}{2}\Delta + V_{\mathrm{ext}}(\mathbf{r}, t) + V_{\mathrm{PP}}(\mathbf{r}) + V_{\mathrm{Hxc}}[P(t)]. \tag{3.1.2}$$

Here V_{PP} is the pseudopotential operator due to the electron-ion interaction. After spatial discretization, V_{PP} becomes a matrix independent of the time t and the density matrix P. V_{Hxc} is the sum of the Hartree and exchange-correlation potentials. $V_{\mathrm{ext}}(\mathbf{r}, t)$ represents the possible external potential such as a time-dependent electric field.

As we discuss before, the range of applicability of RT-TDDFT is often hindered by the very small time step needed to propagate the Schrödinger equation. The parallel transport gauge can potentially "flatten" the wave functions thus allow much larger time step size. When combined with implicit time integrators to propagate the parallel transport dynamics, it is possible to significantly increase the time step size without sacrificing accuracy.

[1]Adapted with permission from [85]. Copyright 2018 American Chemical Society.

In this chapter, we generalize the parallel transport dynamics to the TDDFT setup with multiple wave functions, and numerical test its performance via three TDDFT calculations. In particalur, using absorption spectrum, ultrashort laser pulse, and Ehrenfest dynamics calculations for example, we show that the new method can utilize a time step that is on the order of $10 \sim 100$ attoseconds in a planewave basis set, and is no less than $5 \sim 10$ times faster when compared to the standard explicit 4th order Runge-Kutta time integrator. Please note that, since TDDFT allows electron excitation, our numerical results demonstrate that parallel transport gauge can also benefit the simulation beyond near adiabatic regime.

The rest of this chapter is organized as follows. In Section 3.2, we derive the parllel transport formalism for the TDDFT equations. Section 3.3 discusses numerical integrators for TDDFT equations under the parallel transport gauge, followed by our numerical results in Section 3.4.

3.2 Derivation of the parallel transport gauge

In order to derive the parallel transport gauge, let us first consider the RT-TDDFT equations

$$\mathrm{i}\partial_t \psi_i(t) = H(t, P(t))\psi_i, \quad i = 1, \ldots, N_e. \tag{3.2.1}$$

Here $\Psi(t) = [\psi_1, \ldots, \psi_{N_e}]$ are the electron orbitals, and the Hamiltonian can depend explicitly on t and nonlinearly on the density matrix $P(t) = \Psi(t)\Psi^*(t)$ or the electron density $\rho(t) = \sum_{i=1}^{N_e} |\psi_i(t)|^2$. Eq. (3.2.1) can be equivalently written using a set of transformed orbitals $\Phi(t) = \Psi(t)U(t)$, where the gauge matrix $U(t)$ is a unitary matrix of size N_e. An important property of the density matrix is that it is gauge-invariant: $P(t) = \Psi(t)\Psi^*(t) = \Phi(t)\Phi^*(t)$, and always satisfies the von Neumann equation (or quantum Liouville equation)

$$\mathrm{i}\partial_t P = [H, P] = HP - PH. \tag{3.2.2}$$

Our goal is to optimize the gauge matrix, so that the transformed orbitals $\Phi(t)$ vary *as slowly as possible*, without altering the density matrix. This results in the following variational problem

$$\min_{U(t)} \quad \|\dot{\Phi}\|_F^2, \text{ s.t. } \Phi(t) = \Psi(t)U(t), U^*(t)U(t) = I_{N_e}. \tag{3.2.3}$$

Here $\|\dot{\Phi}\|_F^2 := \mathrm{Tr}[\dot{\Phi}^*\dot{\Phi}]$ measures the Frobenius norm of the time derivative of the transformed orbitals.

In order to solve (3.2.3), we first split $\dot{\Phi}$ into two orthogonal components

$$\dot{\Phi} = P\dot{\Phi} + (I - P)\dot{\Phi}. \tag{3.2.4}$$

Then we have

$$\|\dot{\Phi}\|_F^2 = \|P\dot{\Phi}\|_F^2 + \|(I-P)\dot{\Phi}\|_F^2. \tag{3.2.5}$$

To reformulate the second term, we take the time derivative on the equation $P\Phi = \Phi$ and get

$$\dot{P}\Phi = \dot{\Phi} - P\dot{\Phi} = (I-P)\dot{\Phi}. \tag{3.2.6}$$

Thus Eq. (3.2.5) becomes

$$\|\dot{\Phi}\|_F^2 = \|P\dot{\Phi}\|_F^2 + \|\dot{P}\Phi\|_F^2 = \|P\dot{\Phi}\|_F^2 + \|\dot{P}\Psi\|_F^2, \tag{3.2.7}$$

where the last equality comes from that $\Phi = \Psi U$ and U is a unitary gauge matrix.

Eq. (3.2.7) has a clear physical interpretation. The second term

$$\|\dot{P}\Psi\|_F^2 = \mathrm{Tr}[\Psi^*\dot{P}^2\Psi] = \mathrm{Tr}[\dot{P}^2\Psi\Psi^*] = \mathrm{Tr}[\dot{P}^2 P] \tag{3.2.8}$$

is defined solely from the density matrix and is thus gauge-invariant. Therefore the variation of Φ is minimized when

$$P\dot{\Phi} = 0, \tag{3.2.9}$$

which is exactly the parallel transport condition.

Now we would like to directly write down the governing equation of Φ. First, the equation $\Phi = P\Phi$ and the parallel transport condition (3.2.21) imply that

$$\dot{\Phi} = \partial_t(P\Phi) = \dot{P}\Phi + P\dot{\Phi} = \dot{P}\Phi. \tag{3.2.10}$$

Together with the von Neumann equation, we have

$$\mathrm{i}\dot{\Phi} = \mathrm{i}\dot{P}\Phi = [H,P]\Phi = HP\Phi - PH\Phi = H\Phi - \Phi(\Phi^* H\Phi). \tag{3.2.11}$$

This is exactly the parallel transport dynamics.

The name "parallel transport gauge" originates from the parallel transport formulation associated with a family of density matrices $P(t)$, which generates a parallel transport evolution operator $\mathcal{T}(t)$ as (see *e.g.* [119, 47])

$$\mathrm{i}\partial_t\mathcal{T} = [\mathrm{i}\partial_t P, P]\mathcal{T}, \quad \mathcal{T}(0) = I. \tag{3.2.12}$$

We demonstrate that starting from an initial set of orbitals Ψ_0, the solution to the parallel transport dynamics (3.2.23) is simply evolved by the parallel transport evolution operator according to $\Phi(t) = \mathcal{T}(t)\Psi_0$. To show this, we first prove the following relation

$$P(t)\mathcal{T}(t) = \mathcal{T}(t)P(0) \tag{3.2.13}$$

by showing that both sides solve the same initial value problem. Note that $\mathcal{T}(t)P(0)$ satisfies

$$\mathrm{i}\partial_t(\mathcal{T}(t)P(0)) = [\mathrm{i}\partial_t P, P](\mathcal{T}(t)P(0)). \tag{3.2.14}$$

We then would like to derive the differential equation $P(t)\mathcal{T}(t)$ satisfies. Taking the time derivative on both sides of the identity $P = P^2$, we have

$$\dot{P} = \dot{P}P + P\dot{P} \tag{3.2.15}$$

and thus

$$P\dot{P}P = (\dot{P} - \dot{P}P)P = \dot{P}(P - P^2) = 0. \tag{3.2.16}$$

Then

$$\mathrm{i}\partial_t(P\mathcal{T}) = \mathrm{i}\dot{P}\mathcal{T} + \mathrm{i}P\dot{\mathcal{T}} = \mathrm{i}\dot{P}\mathcal{T} + \mathrm{i}P[\dot{P}, P]\mathcal{T} = \mathrm{i}\dot{P}P\mathcal{T}.$$

On the other hand,

$$[i\dot{P}, P](P\mathcal{T}) = \mathrm{i}(\dot{P}PP\mathcal{T} - P\dot{P}P\mathcal{T}) = \mathrm{i}\dot{P}P\mathcal{T}.$$

Therefore

$$\mathrm{i}\partial_t(P\mathcal{T}) = [i\dot{P}, P](P\mathcal{T}). \tag{3.2.17}$$

Together with the same initial value $P(0)\mathcal{T}(0) = \mathcal{T}(0)P(0) = P(0)$, we have proved that $P(t)\mathcal{T}(t) = \mathcal{T}(t)P(0)$. Using this relation, we have

$$P(t)(\mathcal{T}(t)\Psi_0) = \mathcal{T}(t)P(0)\Psi_0 = \mathcal{T}(t)\Psi_0. \tag{3.2.18}$$

Since $\mathcal{T}(t)$ is unitary, we have $(\mathcal{T}(t)\Psi_0)^*(\mathcal{T}(t)\Psi_0) = I$ for all t. Hence $\mathcal{T}(t)\Psi_0$ forms an orthogonal basis in the image of $P(t)$. Therefore

$$P(t) = (\mathcal{T}(t)\Psi_0)(\mathcal{T}(t)\Psi_0)^*. \tag{3.2.19}$$

By Eq. (3.2.13), (3.2.17) and the von Neumann equation, we have

$$\begin{aligned}\mathrm{i}\partial_t(\mathcal{T}\Psi_0) =& \mathrm{i}\partial_t(P\mathcal{T})\Psi_0 = [\mathrm{i}\dot{P}, P]P\mathcal{T}\Psi_0 \\ =& \mathrm{i}\dot{P}P\mathcal{T}\Psi_0 = HP\mathcal{T}\Psi_0 - PHP\mathcal{T}\Psi_0.\end{aligned} \tag{3.2.20}$$

Finally using Eq. (3.2.18) and (3.2.19), we have

$$\mathrm{i}\partial_t(\mathcal{T}\Psi_0) = H(\mathcal{T}\Psi_0) - (\mathcal{T}\Psi_0)((\mathcal{T}\Psi_0)^* H(\mathcal{T}\Psi_0)),$$

thus $\mathcal{T}\Psi_0$ precisely solves the parallel transport dynamics, indicating $\Phi(t) = \mathcal{T}(t)\Psi_0$.

In summary, the minimizer of (3.2.3), in terms of Φ, satisfies

$$P\dot{\Phi} = 0. \tag{3.2.21}$$

Eq. (3.2.21) implicitly defines a gauge choice for each $U(t)$, and this gauge is called the *parallel transport gauge*. The governing equation of each transformed orbital φ_i can be concisely written down as

$$\mathrm{i}\partial_t \varphi_i = H\varphi_i - \sum_{j=1}^{N_e} \varphi_j \left\langle \varphi_j | H | \varphi_i \right\rangle, \quad i = 1, \cdots, N_e, \tag{3.2.22}$$

or more concisely in the matrix form

$$\mathrm{i}\partial_t \Phi = H\Phi - \Phi(\Phi^* H \Phi), \quad P(t) = \Phi(t)\Phi^*(t). \tag{3.2.23}$$

The right hand side of Eq. (3.2.23) is analogous to the residual vectors of an eigenvalue problem in the time-independent setup. Hence $\Phi(t)$ follows the dynamics driven by residual vectors and is expected to vary slower than $\Psi(t)$.

3.3 Numerical discretization

In order to propagate the parallel transport dynamics numerically, all the RT-TDDFT propagation methods can be used since Eq. (3.2.23) only differs from Eq. (3.2.1) in one extra term $\Phi(\Phi^* H \Phi)$.

We list several propagation schemes used in this chapter, but the parallel transport dynamics can be discretized with any propagator. Here all the $H_n = H(t_n, P_n)$ is the Hamiltonian at step t_n, and $t_{n+\frac{1}{2}} = t_n + \frac{1}{2}\Delta t$, $t_{n+1} = t_n + \Delta t$. For implicit time integrators, Ψ_{n+1} or Φ_{n+1} needs to be solved self-consistently.

The standard explicit 4th order Runge-Kutta scheme for the Schrödinger dynamics (S-

RK4):

$$
\begin{aligned}
k_1 &= -\mathrm{i}\Delta t H_n \Psi_n, \\
\Psi_n^{(1)} &= \Psi_n + \frac{1}{2}k_1, \quad H_n^{(1)} = H(t_{n+\frac{1}{2}}, \Psi_n^{(1)}\Psi_n^{(1)*}) \\
k_2 &= -\mathrm{i}\Delta t H_n^{(1)} \Psi_n^{(1)}, \\
\Psi_n^{(2)} &= \Psi_n + \frac{1}{2}k_2, \quad H_n^{(2)} = H(t_{n+\frac{1}{2}}, \Psi_n^{(2)}\Psi_n^{(2)*}) \\
k_3 &= -\mathrm{i}\Delta t H_n^{(2)} \Psi_n^{(2)}, \\
\Psi_n^{(3)} &= \Psi_n + k_3, \quad H_n^{(3)} = H(t_{n+1}, \Psi_n^{(3)}\Psi_n^{(3)*}) \\
k_4 &= -\mathrm{i}\Delta t H_n^{(3)} \Psi_n^{(3)}, \\
\Psi_{n+1} &= \Psi_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4).
\end{aligned}
\tag{3.3.1}
$$

The standard explicit 4th order Runge-Kutta scheme for the parallel transport dynamics (PT-RK4):

$$
\begin{aligned}
k_1 &= -\mathrm{i}\Delta t\{H_n\Phi_n - \Phi_n(\Phi_n^* H_n \Phi_n)\}, \\
\Phi_n^{(1)} &= \Phi_n + \frac{1}{2}k_1, \quad H_n^{(1)} = H(t_{n+\frac{1}{2}}, \Phi_n^{(1)}\Phi_n^{(1)*}) \\
k_2 &= -\mathrm{i}\Delta t\{H_n^{(1)}\Phi_n^{(1)} - \Phi_n^{(1)}(\Phi_n^{(1)*} H_n^{(1)} \Phi_n^{(1)})\}, \\
\Phi_n^{(2)} &= \Phi_n + \frac{1}{2}k_2, \quad H_n^{(2)} = H(t_{n+\frac{1}{2}}, \Phi_n^{(2)}\Phi_n^{(2)*}) \\
k_3 &= -\mathrm{i}\Delta t\{H_n^{(2)}\Phi_n^{(2)} - \Phi_n^{(2)}(\Phi_n^{(2)*} H_n^{(2)} \Phi_n^{(2)})\}, \\
\Phi_n^{(3)} &= \Phi_n + k_3, \quad H_n^{(3)} = H(t_{n+1}, \Phi_n^{(3)}\Phi_n^{(3)*}) \\
k_4 &= -\mathrm{i}\Delta t\{H_n^{(3)}\Phi_n^{(3)} - \Phi_n^{(3)}(\Phi_n^{(3)*} H_n^{(3)} \Phi_n^{(3)})\}, \\
\Phi_{n+1} &= \Phi_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4).
\end{aligned}
\tag{3.3.2}
$$

The implicit Crank-Nicolson scheme for the Schrödinger dynamics (S-CN):

$$
\left(I + \mathrm{i}\frac{\Delta t}{2}H_{n+1}\right)\Psi_{n+1} = \left(I - \mathrm{i}\frac{\Delta t}{2}H_n\right)\Psi_n. \tag{3.3.3}
$$

The implicit Crank-Nicolson scheme for the parallel transport dynamics (PT-CN):

$$
\begin{aligned}
&\Phi_{n+1} + \mathrm{i}\frac{\Delta t}{2}\left\{H_{n+1}\Phi_{n+1} - \Phi_{n+1}\left(\Phi_{n+1}^* H_{n+1}\Phi_{n+1}\right)\right\} \\
=&\Phi_n - \mathrm{i}\frac{\Delta t}{2}\left\{H_n\Phi_n - \Phi_n\left(\Phi_n^* H_n \Phi_n\right)\right\}.
\end{aligned}
\tag{3.3.4}
$$

In Eq. (3.3.4), the solution Φ_{n+1} needs to be solved self-consistently. This is a set of nonlinear equations with respect to the unknowns Φ_{n+1}, and can be efficiently solved by *e.g.* the preconditioned Anderson mixing scheme [8]. The propagation of $\Phi(t)$ can also be naturally combined with the motion of nuclei discretized *e.g.* by the Verlet scheme for the simulation of Ehrenfest dynamics [111].

3.4 Numerical results

Next, we demonstrate the performance of the PT-CN scheme for RT-TDDFT calculations for three real systems representing three prototypical usages of RT-TDDFT. Our method is implemented in PWDFT code, which uses the planewave basis set and is a self-contained module in the massively parallel DGDFT (Discontinuous Galerkin Density Functional Theory) software package [104, 76]. We use the Perdew-Burke-Ernzerhof (PBE) exchange correlation functional [129], and the Optimized Norm-Conserving Vanderbilt (ONCV) pseudopotentials [71, 138].

Absorption spectrum

The first example is the computation of the absorption spectrum of an anthracene molecule ($C_{14}H_{10}$, Fig. 3.4.1). The simulation is performed using a cubic supercell of size $(20\text{Å})^3$, and the kinetic energy cutoff is 20 au. In order to compute the absorption spectrum, a δ-pulse of strength 0.005 au is applied to the x, y, z directions to the ground state wavefunctions respectively, and the system is then propagated for 4.8 fs along each direction. This gives the polarization tensor $\chi(\omega)$, and the optical absorption cross-section is evaluated as

$$\sigma(\omega) = (4\pi\omega/c)\,\mathrm{Im}\,\mathrm{Tr}[\chi(\omega)].$$

We set the time step size of PT-CN to be 12 attoseconds (as), and that of S-RK4 to be 1 as (it becomes unstable when the step size is larger). Fig. 3.4.2 compares the absorption spectrum obtained from PT-CN and S-RK4 with PWDFT. This result is benchmarked against the linear response time-dependent density functional theory (LR-TDDFT) calculation using the turboTDDFT module [115] from the Quantum ESPRESSO software package [63], which performs 3000 Lanczos steps along each perturbation direction to evaluate the polarization tensor. A Lorentzian smearing of 0.27 eV is applied to all calculations. We find that the absorption spectrum calculations from the three methods agree very well. The spectrum obtained from PT-CN and that from S-RK4 are nearly indistinguishable below 10 eV, and becomes slightly different above 15 eV. Note that the δ-pulse simultaneously excites all eigenstates from the entire spectrum, and $\omega = 15$ eV already amounts to the time scale of 40 as, which is approaching the step size of the PT-CN method. Since the

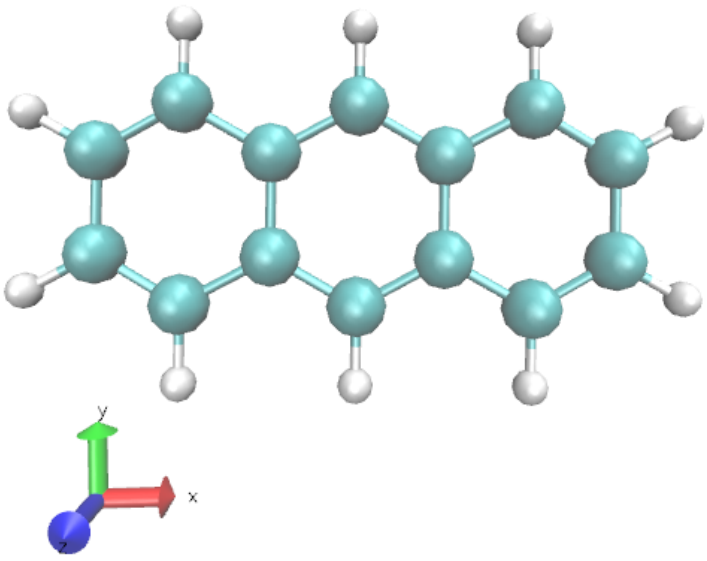


Figure 3.4.1: Atomic configuration of anthracene.

computational cost of RT-TDDFT calculations is mainly dominated by the cost of applying the Hamiltonian operator to orbitals, we measure the numerical efficiency using the number of such matrix-vector multiplications per orbital. The PT-CN method requires on average 4.9 matrix-vector multiplications for each orbital. This is comparable to the S-RK4 method which requires 4 matrix-vector multiplications per time step. Hence for this example, the PT-CN method is around 10 times faster than the S-RK4 method.

Ultrafast laser

The second system is a benzene molecule driven by an ultrashort laser pulse, where the external potential $V_{\text{ext}}(\mathbf{r}, t) = \mathbf{r} \cdot \mathbf{E}(t)$ is given by a time-dependent electric field

$$\mathbf{E}(t) = \hat{\mathbf{k}} E_{\max} \exp\Big[-\frac{(t-t_0)^2}{2a^2} \Big] \sin[\omega(t-t_0)], \tag{3.4.1}$$

where $\hat{\mathbf{k}}$ is a unit vector defining the polarization of the electric field. The parameters $a, t_0, E_{\max}, \omega$ define the width, the initial position of the center, the maximum amplitude of the Gaussian envelope, and the frequency of the laser, respectively. In practice ω and a are often determined by the wavelength λ and the full width at half maximum (FWHM)

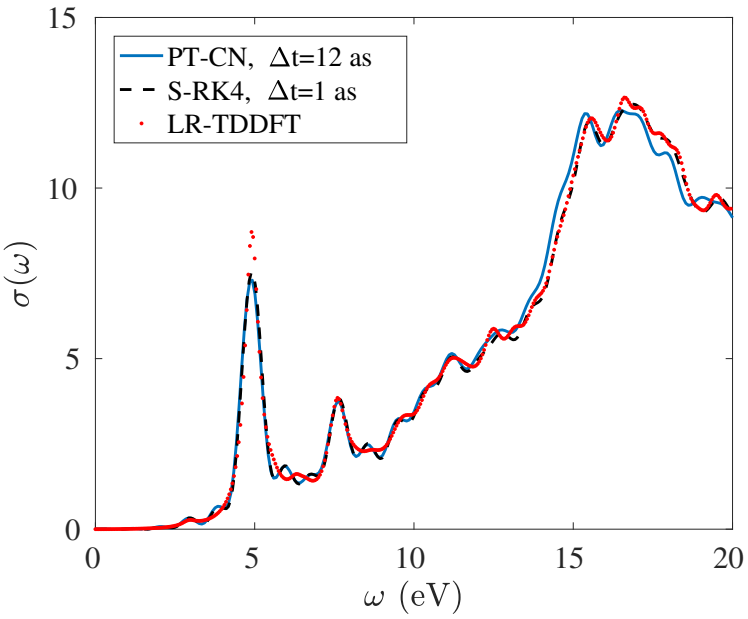


Figure 3.4.2: Absorption spectrum for anthracene.

pulse width [134], *i.e.* $\lambda\omega = 2\pi c$ and FWHM $= 2a\sqrt{2\log 2}$, where c is the speed of the light. In this example, the peak electric field $E_{\max}$ is 1.0 eV/Å, occurring at $t_0 = 15.0$ fs. The FWHM pulse width is 6.0 fs, and the polarization of the laser field is aligned along the x axis (the benzene molecule is in x-y plane, see Fig. 3.4.3a). We consider one relatively slow laser with wavelength 800 nm, and another faster laser with wavelength 250 nm, respectively (Fig. 3.4.3). The electron dynamics for the first laser is in the near adiabatic regime, where the system stays near the ground state after the active time interval of the laser, while the second laser drives electrons to excited states. We implement S-RK4 and PT-CN in the PWDFT package, and propagate TDDFT to $T = 30.0$ fs. For the parameters in the Anderson mixing, the step length α is 0.2, the mixing dimension is 10, and the tolerance is 10^{-6}. We measure the accuracy using the dipole moment $\mathbf{D}(t) := \mathrm{Tr}[\mathbf{r}P(t)]$, as well as the energy difference $E(t) - E(0)$ along the trajectory.

Figure 3.4.4 shows the numerical results for the 800 nm laser using S-RK4 with a step size 0.0005 fs and PT-CN with a step size 0.05 fs. In this case, the system stays near the ground state after the active time interval of the laser. After 25.0 fs, the total energy for S-RK4 only increases by 2.00×10^{-4} eV, and hence we may use the results from S-RK4 as our benchmark. We remark that S-RK4 becomes unstable at large time step sizes. Even when increasing the time step to be 0.001 fs, S-RK4 blows up within 100 time steps. We

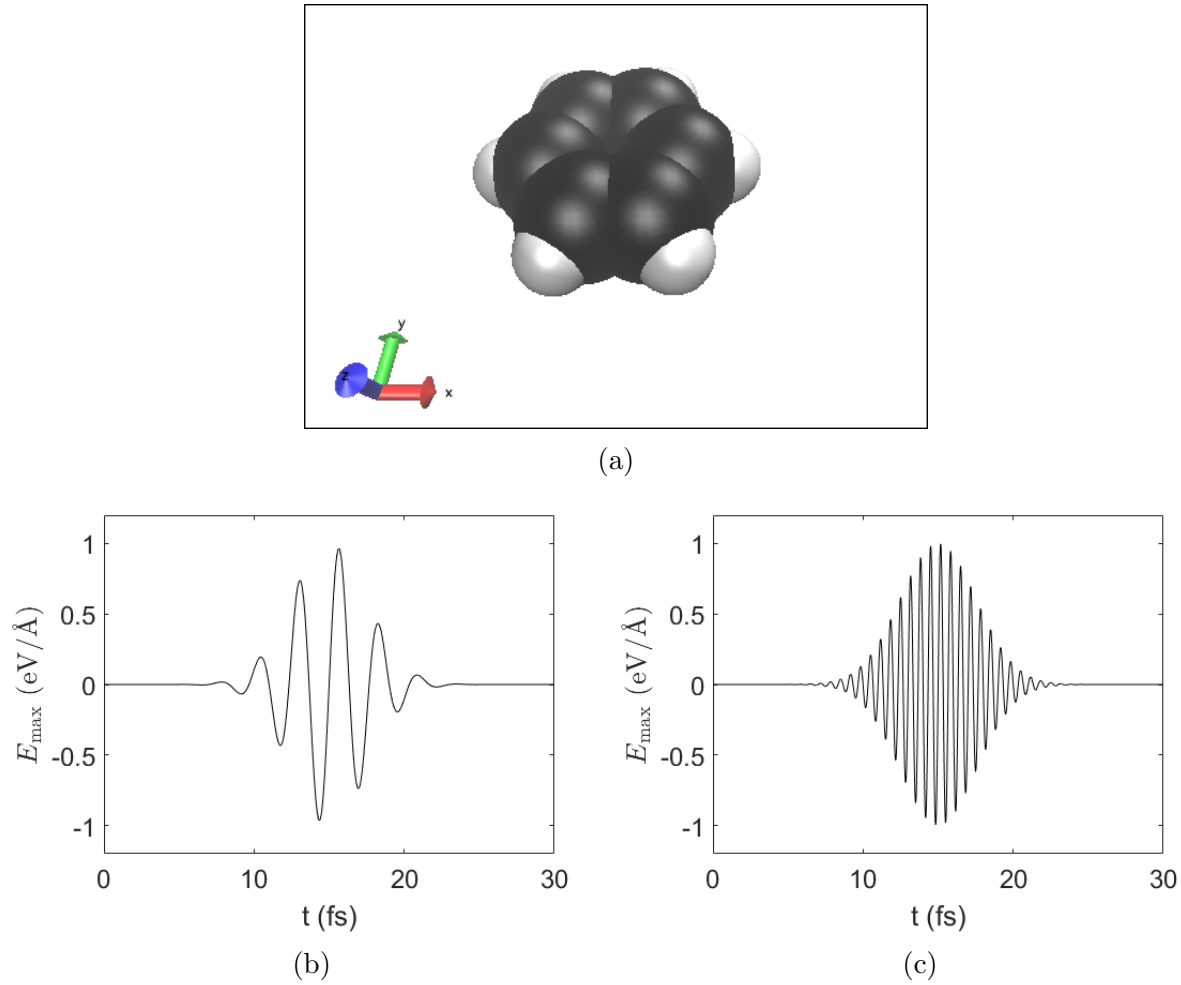
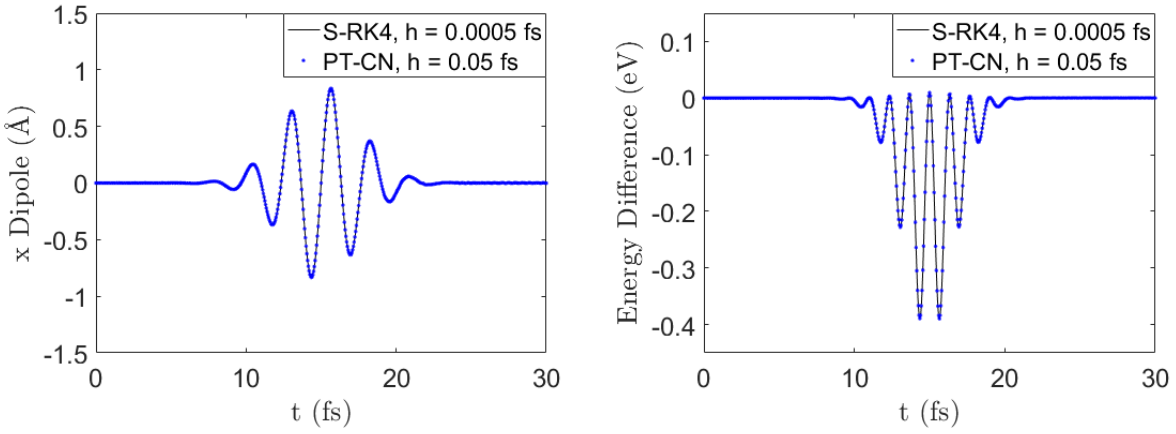


Figure 3.4.3: (a) The benzene molecule. The direction of the external electric field is along the x-axis. This figure is generated by VMD package [77]. (b)(c) The intensity of the electric field. The peak electric field $E_{\max}$ is 1.0 eV/Å, occurring at $t_0 = 15.0$ fs, and the FWHM pulse width is 6.0 fs. The wavelength is 800 nm in (b), and 250 nm in (c).



(a) (b)

Figure 3.4.4: (a) Dipole moment along the x-direction and (b) total energy difference with the 800 nm laser.

observe that PT-CN agrees perfectly with S-RK4 in terms of the dipole moment along the x direction, and the total energy difference. After 25.0 fs, the total energy is nearly constant and only slightly increases by 2.44×10^{-4} eV compared to that of the initial state.

Since the computational cost of TDDFT calculations is mainly dominated by the cost of applying the Hamiltonian matrix to wave functions, we measure the numerical efficiency using the number of such matrix-vector multiplications. Although PT-CN requires more matrix-vector multiplications in each time step, the total number of matrix-vector multiplications is still significantly reduced due to the larger time step size, and PT-CN usually achieves a significant speedup. More specifically, in this case, during the time interval for which the laser is active (from 5.5 fs to 24.5 fs), the average number of matrix-vector multiplications in each PT-CN time step is 12.6, and the total number of matrix-vector multiplications in the simulation is 4798. On the other hand, the number of matrix-vector multiplications in each S-RK4 time step is 4, and the total number of matrix-vector multiplications during this period using time step 0.0005 fs is 152000. Hence the overall speedup of PT-CN over RK4 is 31.7.

Figure 3.4.5 shows the numerical results for the 250 nm laser. In this case, the laser carries more energy and hence a significant amount of electrons can reach the excited states. According to the S-RK4 benchmark, the total energy of the system increases by 0.5260 eV after 25.0 fs. Furthermore, the dipole moment along the x direction oscillates more strongly due to the excitation. PT-CN needs to adopt a smaller time step size 0.005 fs, and still gives

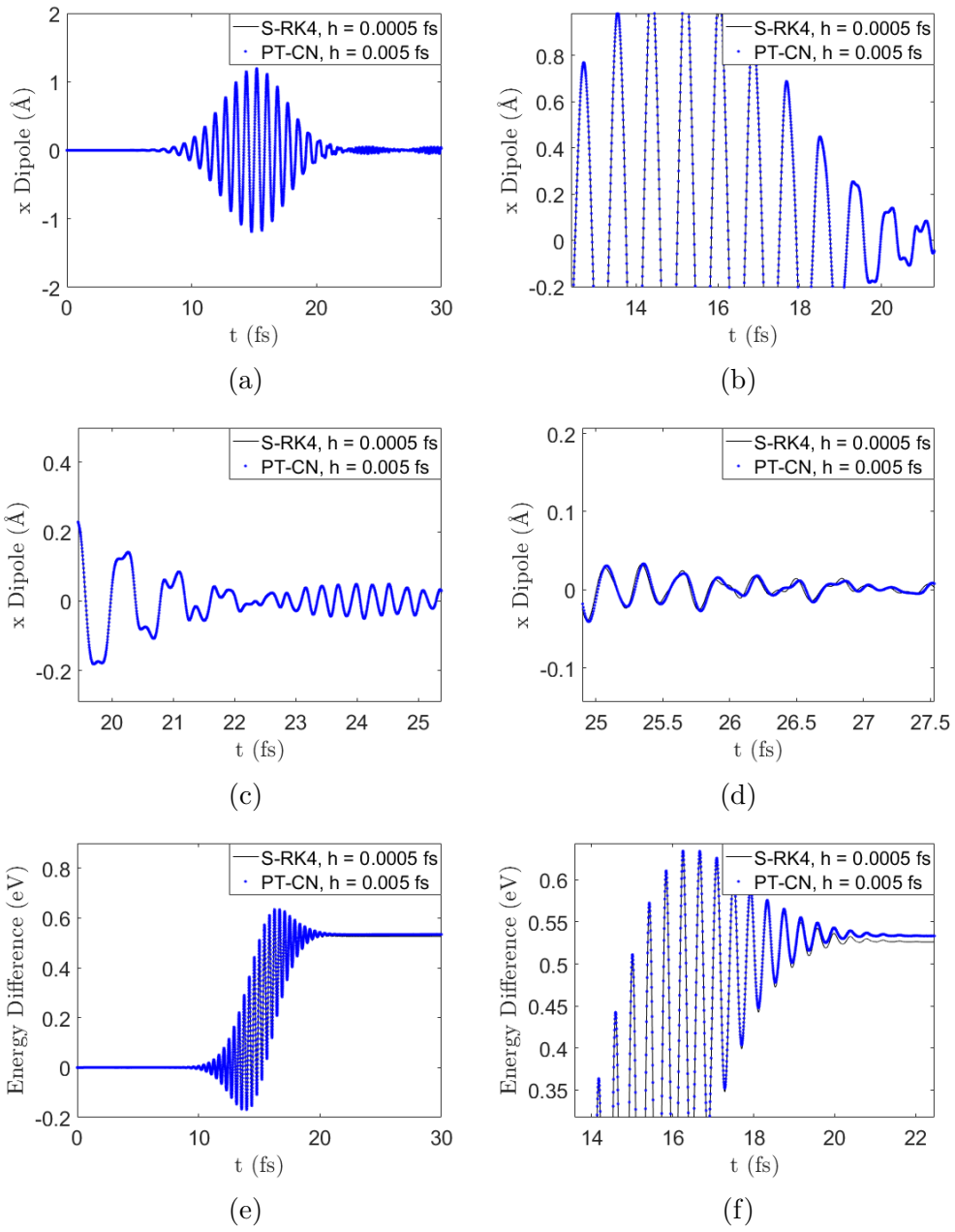


Figure 3.4.5: (a) Dipole moment along the x-direction and (e) total energy difference with the 250 nm laser, with zoom-in views provided in (b)(c)(d)(f).

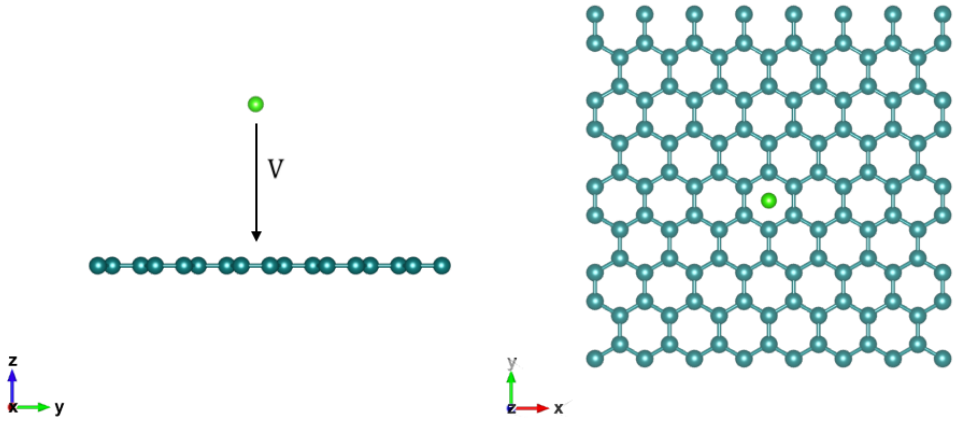
Method	h (fs)	AEI (eV)	AOE (eV)	MVM	Speedup
S-RK4	0.0005	0.5260	/	152000	/
PT-CN	0.005	0.5340	0.0080	28610	5.3
PT-CN	0.0065	0.5347	0.0087	22649	6.7
PT-CN	0.0075	0.5362	0.0102	21943	6.9
PT-CN	0.01	0.5435	0.0175	15817	9.6
PT-CN	0.02	0.5932	0.0672	12110	12.6

Table 3.1: Accuracy and efficiency of PT-CN for the electron dynamics with the 250 nm laser compared to S-RK4. The accuracy is measured using the average energy increase (AEI) after 25.0 fs and the average overestimated energy (AOE) after 25.0 fs. The efficiency is measured using the total number of matrix-vector multiplications (MVM) during the time interval from 5.5 fs to 24.5 fs, and the computational speedup.

a very good approximation to the electron dynamics compared to S-RK4, For the dipole moment, PT-CN results match very well with S-RK4 benchmark during (Fig. 3.4.5b) and after (Fig. 3.4.5c and 3.4.5d) the active time interval of the laser. The total energy obtained by PT-CN matches very well with that in S-RK4 benchmark during the active interval and stays at a constant level with an average increase of 0.5340 eV by the end of the simulation (Fig. 3.4.5e and 3.4.5f). In this case, PT-CN slightly overestimates the total energy after the laser's action by 7.96×10^{-3} eV.

For the computational costs within the period from 5.5 fs to 24.5 fs, the total number of matrix-vector multiplications is still 152000 for S-RK4. The average number of matrix-vector multiplications in each PT-CN time step is 7.5 due to the reduced step size, and the total number of matrix-vector multiplications is 28610. Therefore in this case PT-CN achieves 5.3 times speedup over S-RK4.

We remark that even the electron dynamics is beyond the adiabatic regime, PT-CN can still be stable with a larger time step. Table 3.1 measures the accuracy of PT-CN with $h =$ 0.005 fs, 0.0065 fs, 0.0075 fs, 0.01 fs and 0.02 fs, respectively. We find that the number of matrix-vector multiplications systematically reduces as the step size increases. When the step size is 0.02 fs, the speed up over S-RK4 is 12.6, and this is at the expense of overestimating the energy by 0.0672 eV after the active interval of the laser. Hence one can use PT-CN to quickly study the electron dynamics with a large time step, while this is not possible using an explicit scheme like S-RK4.



(a) x direction (b) z direction

Figure 3.4.6: Model for the collision of Cl/Cl^- and a graphene nanoflake.

Ehrenfest dynamics

As the last example, we use the RT-TDDFT based Ehrenfest dynamics to study the process of a chlorine ion (Cl^-) colliding to a graphene nanoflake consisting of 112 atoms (shown in Fig. 3.4.6). This models the ion implantation procedure for doping a substrate. At the beginning of the simulation, the Cl^- is placed at 6 Å away from the graphene and is given an initial velocity perpendicular to the plane of the graphene pointing towards the center of one hexagonal ring formed by the carbon atoms. The simulation is terminated before the ion reaches the boundary of the supercell. For instance, we set $T = 10$ fs when the velocity is 2.0 Bohr/fs. In such case, the time step size for PT-CN and S-RK4 is set to be 50 as and 0.5 as, respectively. Each PT-CN step requires on average 28 matrix-vector multiplication operations per orbital, and the overall speedup of PT-CN over S-RK4 is 14.2.

We compare the result obtained from the Ehrenfest dynamics with that from the Born-Openheimer Molecular Dynamics (BOMD). In the BOMD simulation, since the extra electron of Cl^- will localize on the conduction band of the graphene conduction rather than on Cl during the self-consistent field iteration, we replace the Cl^- ion by the Cl atom. Fig. 3.4.7 (a) illustrates the energy transfer with different initial kinetic energies. As the Cl/Cl^- initial kinetic energy increases, the gain of the kinetic energy by the graphene atoms decreases due to that Cl/Cl^- can pass through the system faster. When the initial kinetic energy of Cl/Cl^- is smaller than 500 eV, the losses of the kinetic energy for Cl/Cl^- are similar between RT-TDDFT and BOMD. However, when the initial kinetic energy of Cl/Cl^- further

increases, the RT-TDDFT predicts an increase of the loss of the Cl/Cl^- kinetic energy, while the gain of the graphene kinetic energy remains decreasing. This is a consequence of the electron excitation, which is absent in the BOMD simulation. Such excitation is illustrated in Fig. 3.4.7 (b) for the occupied electron density of states in the higher energy regimes. The occupied density of states is calculated as $\rho(\varepsilon) := \sum_{j=1}^{N_e} \sum_{i=1}^{\infty} |\langle \phi_i(T) | \psi_j(T) \rangle|^2 \tilde{\delta}(\varepsilon - \varepsilon_i(T))$. Here $\psi_j(T)$ is the j-th orbital obtained at the end of the RT-TDDFT simulation at time T, and $\varepsilon_i(T), \phi_i(T)$ are the eigenvalues and wavefunctions corresponding to the Hamiltonian at time T. $\tilde{\delta}$ is a Dirac-δ function with a Gaussian broadening of 0.05 eV.

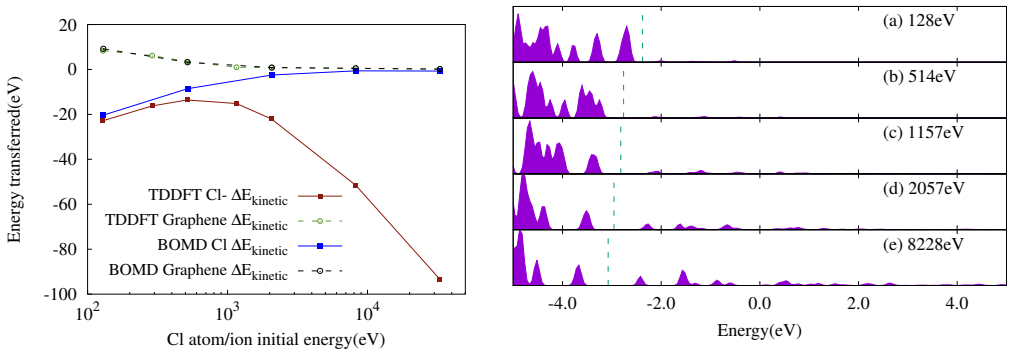


Figure 3.4.7: Energy transfer and density of states. (a) BOMD and RT-TDDFT energy transfer with different initial kinetic energies. (b) Density of state after the ion collision. Green dashed line: Fermi energy.

Fig. 3.4.8 presents further details of the energy transfer along the trajectory of the RT-TDDFT and BOMD simulation when the initial velocity is 2.0 Bohr/fs (2057 eV). When the collision occurs at around $T = 6$ fs, the loss of the Cl/Cl^- kinetic energy is 44 eV and 58 eV under RT-TDDFT and BOMD, respectively. However, after collision Cl regains almost all the kinetic energy in BOMD, and the final kinetic energy is only 2.5 eV less than the initial one. Correspondingly, the kinetic energy of the graphene increases by 0.86 eV and the potential energy increases by 1.63 eV. On the other hand, RT-TDDFT predicts that the Cl^- ion should lose 22.5 eV kinetic energy, which is mostly transferred to the potential energy of the excited electrons. The increase of the kinetic energy of the graphene is 0.84 eV and is similar to the BOMD result. Therefore, in RT-TDDFT, the Cl^- loses its kinetic energy to electron excitation in graphene.

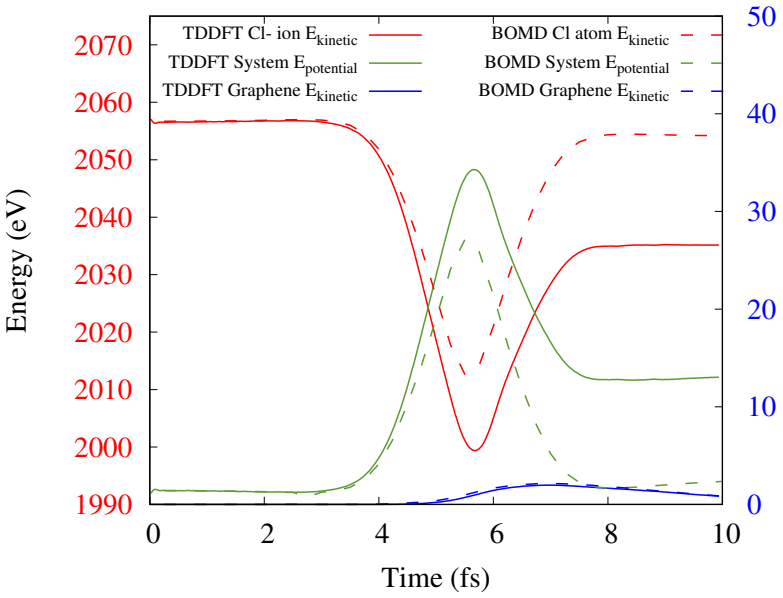


Figure 3.4.8: BOMD and RT-TDDFT energy transfer with time, projectile speed is 2.0 Bohr/fs.

3.5 Conclusion

In this chapter, we demonstrate that one significant factor leading to the very small time step size in RT-TDDFT calculations is the non-optimal gauge choice in the Schrödinger dynamics. Since all physical observables should be gauge-independent, we may optimize the gauge choice to improve the numerical efficiency without sacrificing accuracy. The resulting scheme can be beneficial to any RT-TDDFT integrator, and can even be nearly symplectic. With the increased time step size, we hope that RT-TDDFT can be used to study many ultrafast problems unamenable today.

Chapter 4

Mixed-state parallel transport dynamics

4.1 Introduction

Consider the problem of solving a finite dimensional, (possibly) nonlinear von Neumann equation

$$\mathrm{i}\partial_t \rho(t) = [H(t, \rho(t)), \rho(t)], \quad \rho(0) = \rho_0, \tag{4.1.1}$$

where $\rho_0 \in \mathbb{C}^{N_g \times N_g}$ is a Hermitian matrix satisfying $\rho_0^2 \preceq \rho_0$. Here $[A, B] = AB - BA$ is the commutator of A and B, and $A \preceq B$ means that $A - B$ is a negative semidefinite matrix. The initial quantum state ρ_0 is called a pure state if $\rho_0^2 = \rho_0$, and a mixed state if $\rho_0^2 \prec \rho_0$. Eq. (4.1.1) can be used to describe the dynamics of a closed quantum system in a very general setting, and we allow the time-dependent Hamiltonian $H(t, \rho(t)) \in \mathbb{C}^{N_g \times N_g}$ to have a nonlinear dependence on entries of $\rho(t)$. One prominent application is the real-time time-dependent density functional theory (rt-TDDFT) [133, 159, 126, 151], which is one of the most widely used techniques for studying ultrafast properties of electrons, and has resulted in a variety of applications in quantum physics, chemistry, and materials science.

In practice, ρ_0 is often of low-rank, or can be very well approximated by a low-rank matrix. For simplicity, let the rank of ρ_0 be denoted by N and assume $N \ll N_g$. The von Neumann dynamics neglects the low-rank structure and propagates $\rho(t)$ directly as a dense matrix. For rt-TDDFT calculations with a fine discretization scheme (e.g. the planewave basis set, or the finite difference discretization), N_g can be 10^6 or larger, and the direct propagation of Eq. (4.1.1) becomes extremely expensive. In this case, the von Neumann dynamics is often replaced by a set of nonlinear Schrödinger equations (see Eq. (4.2.1)), and the simulation variables become the electron wavefunctions described by a much smaller matrix $\Psi(t) \in \mathbb{C}^{N_g \times N}$. However, such a rank reduction can come at the cost of the time

step size, denoted by h. In many applications, h for the von Neumann dynamics (4.1.1) can be chosen to be at the sub-femtosecond scale (1 fs= 10^{-15} s), while h for the Schrödinger dynamics needs to be sub-attosecond scale (1 as= 10^{-18} s) [36, 137, 65].

Given a pure initial state, among all possible gauge choices, the parallel transport (PT) gauge yields the slowest gauge-transformed dynamics at any given time, as discussed in previous chapters. Compared to the Schrödinger dynamics, the time step h in the PT dynamics can be chosen to be much larger and is comparable to that of the von Neumann dynamics Eq. (4.1.1). When combined with implicit integrators (such as the Crank-Nicolson method or the implicit midpoint rule), the PT dynamics has been applied to rt-TDDFT simulations for real materials with thousands of atoms at the level of generalized gradient approximation exchange-correlation functionals (GGA, such as the Perdew–Burke–Ernzerhof [129] functional) and hybrid exchange-correlation functionals (such as the Heyd–Scuseria–Ernzerhof [74] functional) [86, 87].

In previous chapters, the PT dynamics is derived for a pure initial state, and its efficiency has been justified in the linear, near adiabatic regime in terms of a singularly perturbed linear system. The pure initial state is suitable for describing molecules and insulating materials at zero temperature. On the other hand, in practice, the initial state is often a low-energy excited state [54, 60, 28], or a thermal state [161] especially for metallic systems. This inspires us to consider the most general setting when ρ_0 is given by a mixed state (for instance, the occupation number of ρ_0 is given by the Fermi-Dirac distribution).

By assuming a dynamical low-rank factorization $\rho(t) \approx \Phi(t)\sigma(t)\Phi^\dagger(t)$, where $\Phi(t) \in \mathbb{C}^{N_g \times N}$ and $\sigma(t) \in \mathbb{C}^{N\times N}$, we derive the PT dynamics in terms of its low-rank factors $\Phi(t), \sigma(t)$. The PT dynamics with a pure initial state is recovered by setting $\sigma(t) = I_N$. When the spectral radius of the Hamiltonian is large, the time step h is simultaneously constrained by accuracy and stability requirements, and implicit integrators are more suited for efficient propagation of the PT dynamics. Using the implicit midpoint (IM) rule (also known as the second order Gauss-Legendre method, GL2) as an example, we derive the discretized numerical scheme, and prove that the resulting PT-IM scheme has certain orthogonality and trace-preserving properties.

We then derive a new error bound for the discretized PT dynamics. Instead of relying on the linear quantum adiabatic theorem to obtain an *a priori* error bound of the solution, our new error bound expresses the local truncation error directly in terms of the Hamiltonian, density matrix, and their derived quantities. Our analysis shows that an upper bound of the local truncation error of PT dynamics only involves certain commutators between the Hamiltonian (or its time derivatives) and the density matrix (or the associated spectral projector), while that of the Schrödinger gauge involves additional terms lacking such commutator structures. Using the commutator type error bound, in the near adiabatic regime when the *a priori* estimate is available from the quantum adiabatic theorem, our new result shows the PT dynamics gains one extra order of accuracy in terms of the singularly per-

turbed parameter ϵ than the Schrödinger dynamics, which reproduces the previous result in Chapter 2. Recently, the quantum adiabatic theorem has been extended to certain weakly nonlinear systems [57, 61]. Our commutator type error analysis can be directly combined with such analysis leading to results comparable to that in Chapter 2 in the weakly nonlinear regime. Away from the near adiabatic regime, the commutator scaling of the PT dynamics can still lead to a significantly smaller error than that of the Schrödinger dynamics. We illustrate the numerical performance of the PT dynamics for a number of one-dimensional model metallic systems, which also verifies the effectiveness of the new error bound.

Related works:

Numerical integrators for rt-TDDFT simulation following the Schrödinger dynamics is a well-studied subject (see an early paper [36], and also [65, 130] for recent comparative studies of a variety of standard numerical integrators), but the importance and the benefit of gauge-transformed dynamics have only been realized recently (see [162] for another type of gauge-transformed dynamics using Wannier functions).

At the continuous level, the PT dynamics is a special case of the dynamical low-rank approximation (DLRA) developed by Lubich *et al.* (see [97, 111] for examples; DLRA is intimately related to the Dirac–Frenkel/McLachlan variational principle in the physics literature). The basic strategy of DLRA is to update a low-rank decomposition (such as eigenvalue or singular value decomposition) of a large matrix (in this case $\rho(t)$) on the fly. For a mixed initial state, a direct application of DLRA involves $\sigma^{-1}(t)$ in the equation of the low-rank factors, which in general can be a source of numerical instability [97, 113]. Our derivation of the PT dynamics with a mixed initial state uses the structure of the von Neumann equation and can be viewed as a simplified derivation of DLRA. It also naturally shows that the pathological term $\sigma^{-1}(t)$ does not appear, so the PT dynamics is numerically stable even if one overestimates the numerical rank of $\rho(t)$.

Regarding the time discretization, existing works of DLRA mostly use explicit integrators, although the possibility of using implicit integrators has also been mentioned in certain settings [113]. Our previous studies in Chapter 2 and Chapter 3 suggest that for rt-TDDFT calculations, the combined use of the PT dynamics and implicit integrators is the key for efficient propagation in real chemical and materials systems. The PT dynamics with a mixed initial state can also be viewed as a special case of the low-rank approximation for solving Lindblad equations by Le Bris *et al.* [100, 101] (since the von Neumann equation can be viewed as the Lindblad equation without the decoherence operator), which is also derived independently of DLRA. It is worth pointing out that [100] introduces an arbitrary Hermitian matrix that can be freely determined. We demonstrate that in the context of the von Neumann dynamics, setting this arbitrary matrix to $H(t)$ (the instantaneous Hamiltonian matrix), and 0 (the zero matrix) leads to the Schrödinger dynamics and the PT dynamics, respectively.

Organization:

The rest of this chapter is organized as follows. In Section 4.2, we introduce some preliminaries of rt-TDDFT and the PT dynamics with a pure initial state. We derive the PT dynamics with a mixed initial state in Section 4.3. For completeness, an alternative derivation of the PT dynamics that explicitly uses the structure of the tangent manifold (which is also a simplified derivation of [100]) is given in Section 4.4. We then derive an implicit numerical propagator for the PT dynamics in Section 4.5. Section 4.6 analyzes the numerical errors of the PT and the Schrödinger dynamics. Finally, we validate the error analysis with numerical results in Section 4.7.

4.2 Preliminaries

In this section, we briefly review the key idea of deriving PT dynamics with a pure initial state discussed in previous chapters. In the setting with a pure initial state, real-time time-dependent density functional theory (rt-TDDFT) solves the following set of Schrödinger equations

$$\mathrm{i}\partial_t \Psi(t) = H(t, \rho(t))\Psi(t), \quad \Psi(0) = \Psi_0. \tag{4.2.1}$$

Here $\Psi(t) = [\psi_1(t), \dots, \psi_N(t)]$ is the collection of electron wavefunctions (also called electron orbitals), and the number of columns N is equal to the number of electrons denoted by N_e (spin degeneracy omitted). The initial set of wavefunctions satisfy the orthonormality condition $\Psi(0)^\dagger \Psi(0) = I_N$. Here $A^\dagger$ denotes the Hermitian conjugate of a matrix or vector A. The density matrix is $\rho(t) = \Psi(t)\Psi^\dagger(t) \equiv \sum_{i=1}^N \psi_i(t)\psi_i^\dagger(t)$, and in particular $\rho_0 := \rho(0) = \Psi(0)\Psi(0)^\dagger$ is a pure state satisfying $\rho_0^2 = \rho_0$.

Throughout the chapter we are concerned with time propagation instead of spatial discretization. Unless otherwise specified, Eq. (4.2.1) represents a discrete, finite dimensional quantum system, i.e. $H(t, \rho)$ is a Hermitian matrix with finite dimension N_g. If the quantum system is spatially continuous, we may first find a set of orthonormal basis functions and expand the continuous wavefunction under this basis. Then after a Galerkin projection, Eq. (4.2.1) becomes an N_g-dimensional quantum system, and $\psi_j(t)$ represents the coefficient vector under the basis for the j-th wavefunction.

The time-dependent Hamiltonian operator $H(t, \rho(t))$ is Hermitian for all t and ρ, and its precise form is not important for the purpose of this chapter. Starting from a pure initial state ρ_0, the orthogonality condition $\Psi(t)^\dagger \Psi(t) = I_N$ is satisfied for all $t \geqslant 0$, and hence $\rho(t)$ is a pure state for all t satisfying $\rho^2(t) = \rho(t)$. Throughout the chapter, we may use the notations $\partial_t \rho = \rho_t = \dot{\rho}$ interchangeably for the time-derivatives. For composite functions such as $H(t, \rho(t))$, we use the notation $\dot{H} := \frac{d}{dt} H(t, \rho(t)) = H_t + H_\rho \rho_t$, where the tensor contractions are defined such that the chain rule holds. For example, the tensor contraction between the 4-tensor H_ρ and the matrix ρ_t are defined such that the chain rule

$\frac{d}{dt}H(t,\rho(t)) = H_t + H_\rho \rho_t$ follows the element-wise operation

$$\frac{d}{dt}H_{ij}(t,\rho(t)) = \partial_t H_{ij}(t,\rho(t)) + \sum_{k,l} \frac{\partial H_{ij}}{\partial \rho_{kl}}(t,\rho(t))\frac{\partial \rho_{kl}(t)}{\partial t}.$$

The set of Schrödinger equations (4.2.1) is equivalent to the von Neumann dynamics (4.1.1). Note that if we right multiply $\Psi(t)$ by a time-dependent unitary matrix $U(t) \in \mathbb{C}^{N\times N}$ and let $\Phi(t) = \Psi(t)U(t)$, then

$$\rho(t) = \Psi(t)\Psi^\dagger(t) = \Phi(t)\left[U^\dagger(t)U(t)\right]\Phi^\dagger(t) = \Phi(t)\Phi^\dagger(t). \tag{4.2.2}$$

The unitary rotation matrix $U(t)$ is called the gauge matrix, and Eq. (4.2.2) indicates that the density matrix is *gauge-invariant*. In particular, the choice $U(t) = I_N$ is referred to as the Schrödinger gauge.

Since all physical observables can be derived from the von Neumann equation (4.1.1) and the density matrix $\rho(t)$, the choice of the gauge matrix $U(t)$ has no measurable effects. On the other hand, the gauge matrix introduces additional degrees of freedom, and can oscillate at a different time scale from that of the corresponding wavefunctions. It is then desirable to optimize the gauge matrix, so that the transformed wavefunctions $\Phi(t)$ vary *as slowly as possible*, without changing the density matrix. This results in the following variational problem

$$\min_{U(t)} \quad \|\dot{\Phi}\|_F^2, \text{ s.t. } \Phi(t) = \Psi(t)U(t), U^\dagger(t)U(t) = I_N. \tag{4.2.3}$$

Here $\|\dot{\Phi}\|_F^2 := \mathrm{Tr}[\dot{\Phi}^\dagger\dot{\Phi}]$ measures the Frobenius norm of the time derivative of the transformed orbitals.

The minimizer of (4.2.3), in terms of Φ, satisfies the following equation

$$\rho\dot{\Phi} = 0. \tag{4.2.4}$$

We refer readers to Chapter 2 for the derivation. Eq. (4.2.4) has an intuitive explanation that the optimal dynamics should minimize the "internal" rotations within the range of ρ. Eq. (4.2.4) implicitly defines a gauge choice for each $U(t)$, and this gauge is called the *parallel transport gauge*. The name "parallel transport" comes from that $\Phi(t)$ can be identified as the unique horizontal lift [119] of $\rho(t)$ from the Grassmann manifold to the Stiefel manifold, starting from the initial condition Ψ_0. This will be further explained in Section 4.6.

From Eq. (4.2.4), the governing equation of $\Phi(t)$ can be concisely written down as

$$\mathrm{i}\partial_t\Phi(t) = H(t,\rho(t))\Phi(t) - \Phi(t)(\Phi^\dagger(t)H(t,\rho(t))\Phi(t)), \quad \Phi(0) = \Psi_0, \tag{4.2.5}$$

where $\rho(t) = \Phi(t)\Phi^\dagger(t)$. Notice that Eq. (4.2.5) introduces one extra term compared to the original dynamics Eq. (4.2.1) under Schrödinger gauge, and directly provides a self-contained

definition of the transformed wavefunctions under the optimal gauge. In practice, we can directly solve Eq. (4.2.5) by numerical schemes to approximate the dynamics, instead of computing the PT gauge explicitly.

To observe the advantage of the parallel transport dynamics, consider the extreme case that each column of Ψ_0 is already an eigenstate of $H(0)$ and $H(t, \rho(t)) \equiv H(0)$ is a time-independent matrix. Then Eq. (4.2.5) is reduced to

$$\mathrm{i}\partial_t \Phi(t) = 0.$$

Hence $\Phi(t) = \Phi(0)$ holds for all $t \geqslant 0$, while each column of the solution Schrödinger dynamics (4.2.1) rotates with a time-dependent phase factor. For less trivial dynamics, the temporal oscillation of gauge-transformed wavefunctions $\Phi(t)$ can still be significantly slower than that of $\Psi(t)$.

4.3 Parallel transport dynamics with a mixed initial state

In rt-TDDFT calculations, the pure initial state can be used for simulating insulating systems starting from the ground state, or a well-defined excited state. In many other cases the initial state should be a mixed state. For instance, for metallic systems at finite temperature, the initial state often takes the form of the Fermi-Dirac distribution

$$\rho(0) = \left(1 + \exp\left(\beta(H(0) - \mu)\right)\right)^{-1}, \tag{4.3.1}$$

where $\beta = 1/(k_{\mathrm{B}}T)$, k_{B} is the Boltzmann constant, T is the temperature. The chemical potential μ is a Lagrange multiplier, which should be adjusted to satisfy the normalization condition

$$\mathrm{Tr}[\rho(0)] = N_e, \tag{4.3.2}$$

where N_e is the number of electrons. If we diagonalize $H(0)$ according to $H(0)\psi_i(0) = \varepsilon_i(0)\psi_i(0)$, then the occupation number

$$s_i(0) := \langle \psi_i(0)|\rho(0)|\psi_i(0)\rangle = \left(1 + \exp\left(\beta(\varepsilon_i(0) - \mu)\right)\right)^{-1}.$$

Hence when β is large (e.g. at room temperature 300K, $\beta \approx 10^3$ in the atomic unit), $s_i(0)$ is very close to 0 when $\varepsilon_i(0) - \mu \gg \beta^{-1}$. Therefore, $\rho(0)$ can be very well approximated by a low rank matrix, with its approximate rank denoted by N. In other words, we can set

$$\rho(0) = \sum_{i=1}^{N} \psi_i(0) s_i(0) \psi_i^\dagger(0) = \Psi(0)\sigma(0)\Psi^\dagger(0),$$

with the chemical potential μ slightly adjusted so that the normalization condition (4.3.2) is still satisfied. Here $\sigma_0 := \sigma(0) = \mathrm{diag}[s_1(0), \ldots, s_N(0)]$ is a diagonal matrix. Since the occupation number satisfies $0 < s_i(0) < 1$, we have $N > N_e$, and $\rho^2(0) \prec \rho(0)$. We also assume $N_g \gg N$.

If we solve the nonlinear Schrödinger equation (4.2.1) to obtain $\Psi(t)$, then

$$\rho(t) = \Psi(t)\sigma_0\Psi^\dagger(t), \tag{4.3.3}$$

is the unique solution of Eq. (4.1.1) (viewed as a large ODE system) with the initial state $\rho(0)$. Hence in practice, we only need to solve Eq. (4.2.1) in the same way as for the pure initial state, but weigh the contribution of each time-dependent vector $\psi_i(t)$ always by the *initial* occupation number $\sigma_i(0)$. This fact that the occupation number $\sigma(t)$ remains as a constant matrix σ_0 can also be derived directly (see Eq. (4.4.9) in Section 4.4). However, similar to the case with a pure initial state, Eq. (4.2.1) can require a relatively small time step size.

Note that we may still apply a gauge matrix $U(t) \in \mathbb{C}^{N\times N}$ and define $\Phi(t) = \Psi(t)U(t)$ with initial condition $U(0) = I_N$. In such a case, we must also redefine the occupation number matrix as

$$\sigma(t) = U^\dagger(t)\sigma_0 U(t), \tag{4.3.4}$$

so that

$$\rho(t) = \Phi(t)\sigma(t)\Phi^\dagger(t) \tag{4.3.5}$$

is satisfied. Here $\sigma(t)$ is now a Hermitian matrix of size N and may no longer be diagonal for $t > 0$. We would like to solve again the optimization problem in Eq. (4.2.3) so that the gauge-transformed wavefunctions $\Phi(t)$ vary as slowly as possible. This leads to Eq. (4.2.4) and hence Eq. (4.2.5), with $\rho(t)$ defined in Eq. (4.3.5). For simplicity, we may also define a gauge-invariant projector

$$P(t) = \Psi(t)\Psi^\dagger(t) = \Phi(t)\Phi^\dagger(t),$$

so that Eq. (4.2.5) can be rewritten as

$$\mathrm{i}\partial_t\Phi(t) = (I - P(t))H(t, \rho(t))\Phi(t).$$

Here the identity matrix is given as $I = I_{N_g}$, and we have used that $P(t)\Phi(t) = \Phi(t)$.

In order to close the equation, it remains to identify the equation of motion of $\sigma(t)$. First, by differentiating the equation $\Psi(t)U(t) = \Phi(t)$ and using (4.2.5), we may derive the dynamics of the gauge $U(t)$, i.e.

$$(\mathrm{i}\partial_t\Psi(t))U(t) + \Psi(t)(\mathrm{i}\partial_t U(t)) = H(t, \rho(t))\Psi(t)U(t) - \Psi(t)\Psi^\dagger(t)H(t, \rho(t))\Psi(t)U(t).$$

This gives

$$\mathrm{i}\partial_t U(t) = -(\Psi^\dagger(t)H(t, \rho(t))\Psi(t))U(t). \tag{4.3.6}$$

By differentiating both sides of Eq. (4.3.4) and using Eq. (4.3.6), we have

$$\begin{aligned}
\mathrm{i}\partial_t\sigma(t) &= (\mathrm{i}\partial_t U^\dagger(t))\sigma_0 U(t) + U^\dagger(t)\sigma_0(\mathrm{i}\partial_t U(t)) \\
&= U^\dagger(t)(\Psi^\dagger(t)H(t,\rho(t))\Psi(t))\sigma_0 U(t) - U^\dagger(t)\sigma_0(\Psi^\dagger(t)H(t,\rho(t))\Psi(t))U(t) \\
&= \Phi^\dagger(t)H(t,\rho(t))\Phi(t)\sigma(t) - \sigma(t)\Phi^\dagger(t)H(t,\rho(t))\Phi(t) \\
&= [\Phi^\dagger(t)H(t,\rho(t))\Phi(t), \sigma(t)].
\end{aligned}$$

The equation of motion for $\sigma(t)$ only depends on the slowly varying gauge-transformed wavefunctions $\Phi(t)$.

In summary, the parallel transport dynamics with a general initial state consists of the following set of equations

$$\begin{aligned}
\mathrm{i}\partial_t\Phi(t) &= (I - P(t))H(t,\rho(t))\Phi(t), \\
\mathrm{i}\partial_t\sigma(t) &= [\Phi^\dagger(t)H(t,\rho(t))\Phi(t), \sigma(t)], \\
\rho(t) &= \Phi(t)\sigma(t)\Phi^\dagger(t), \quad P(t) = \Phi(t)\Phi^\dagger(t), \\
\Phi(0) &= \Psi_0, \quad \sigma(0) = \sigma_0.
\end{aligned} \tag{4.3.7}$$

Eq. (4.3.7) gives a self-contained definition of the transformed wavefunctions $\Phi(t)$ and the matrix $\sigma(t)$ under the optimal gauge. Therefore, we can directly solve Eq. (4.3.7) to numerically approximate the state $\rho(t)$ without computing the PT gauge matrix explicitly. Notice that, compared to the Schrödinger dynamics in which one can set $\sigma(t) = \sigma(0)$, the PT dynamics in Eq. (4.3.7) introduces one extra nonlinear term in the propagation of the wavefunctions, and enlarges the size of the ODE system via a non-trivial dynamics of the transformed $\sigma(t)$. This is different from the pure state setting where only an extra term in the equation of Φ is added. However, due to the assumption that $N_g \gg N$, the increase of the number of variables by N^2 due to $\sigma(t)$ does not add too much overhead in the numerical simulation.

4.4 Alternative derivation of the parallel transport dynamics using the tangent space formulation

In this section, we provide an alternative derivation of the PT dynamics using the tangent space formulation, which follows the derivation in [100, 101] for Lindblad equations, and is more analogous to the derivation of the dynamical low-rank approximation. The derivation also provides an alternative perspective of the gauge choice in terms of an auxiliary Hamiltonian. The presentation of this derivation consists of three parts: we first write down

Eq. (4.3.5) as well as the domain of the low-rank factors. We then derive the equation of motion of the low-rank factors of the rank-N density matrix. Finally, we introduce the optimal gauge (i.e. PT gauge).

Let ρ be of rank N ($N_e \leqslant N < N_g$). Recall Eq. (4.3.5):

$$\rho(t) = \Phi(t)\sigma(t)\Phi^\dagger(t), \tag{4.4.1}$$

where $\sigma(t)$ is an $N \times N$ positive semidefinite Hermitian matrix, and Φ is an $N_g \times N$ complex-valued matrix that satisfies $\Phi^\dagger\Phi = I_N$, in other words, Φ belongs to the Stiefel manifold $\mathrm{St}(N, N_g)$ defined as

$$\mathrm{St}(N, N_g) = \{\Phi \in \mathbb{R}^{N_g \times N} : \Phi^\dagger\Phi = I_N\}.$$

As is explained in Section 4.3, this decomposition (4.4.1) in fact admits an equivalence relationship $(\Phi, \sigma) \equiv (\Phi U, U^\dagger \sigma U)$, namely, for any $N \times N$ unitary matrix U,

$$\rho = \Phi\sigma\Phi^\dagger = (\Phi U)(U^\dagger \sigma U)(U^\dagger \Phi^\dagger).$$

Consider the infinitesimal variation of the tangent map of $(\Phi, \sigma) \mapsto \Phi\sigma\Phi^\dagger$. As is shown in [32, Lemma 4], the tangent space of the Stiefel manifold admits the parametrization $\mathrm{i}\omega\Phi$, where ω is a Hermitian matrix of size N_g. Denote

$$\dot\Phi = \mathrm{i}\omega\Phi, \quad \dot\sigma = \xi,$$

where ξ is a traceless Hermitian matrix of size N. The infinitesimal variation of ρ can thus be represented as

$$\mathrm{i}[\omega, \rho] + \Phi\xi\Phi^\dagger = \Phi(\mathrm{i}[\Phi^\dagger\omega\Phi, \sigma] + \xi)\Phi^\dagger.$$

We then project $\dot\rho$ onto this tangent space by minimizing the distance between them, namely,

$$\min \left\| -\mathrm{i}[H(t, \rho(t)), \rho(t)] - \mathrm{i}[\omega, \rho] - \Phi\xi\Phi^\dagger \right\|_F.$$

Hereafter, we drop the (t, ρ) in H for simplicity. The two stationary conditions in ω and ξ read

$$\left[-\mathrm{i}[H, \rho] - \mathrm{i}[\omega, \rho] - \Phi\xi\Phi^\dagger, \rho\right] = 0, \tag{4.4.2}$$

$$\Phi^\dagger(-\mathrm{i}[H, \rho] - \mathrm{i}[\omega, \rho] - \Phi\xi\Phi^\dagger)\Phi = \lambda I_N, \tag{4.4.3}$$

where λ is the Lagrange multiplier introduced to satisfy the traceless condition of ξ. By taking trace on both sides of (4.4.3), we find that

$$\lambda = \frac{1}{N}\mathrm{Tr}\left[\Phi^\dagger\left(-\mathrm{i}[H, \rho] - \mathrm{i}[\omega, \rho]\right)\Phi\right] = 0,$$

because for any X,

$$\mathrm{Tr}(\Phi^\dagger[X,\rho]\Phi) = \mathrm{Tr}(\Phi^\dagger X\Phi\sigma\Phi^\dagger\Phi - \Phi^\dagger\Phi\sigma\Phi^\dagger X\Phi) = \mathrm{Tr}(\Phi^\dagger X\Phi\sigma - \sigma\Phi^\dagger X\Phi) = 0.$$

Define projection operators

$$P := \Phi\Phi^\dagger, \quad Q = I - P,$$

and one can express $\Phi\xi\Phi^\dagger$ using (4.4.3) as

$$\xi = \Phi^\dagger\left(-\mathrm{i}[H,\rho] - \mathrm{i}[\omega,\rho]\right)\Phi, \quad \Phi\xi\Phi^\dagger = P\left(-\mathrm{i}[H,\rho] - \mathrm{i}[\omega,\rho]\right)P. \tag{4.4.4}$$

Together with Eq. (4.4.2), we obtain

$$Q\left(-\mathrm{i}H\rho - \mathrm{i}\omega\rho\right)\rho = \rho\left(\mathrm{i}\rho H + \mathrm{i}\rho\omega\right)Q.$$

Note that $P\rho = \rho$ and hence the left-hand side stays in the range of Q while the right-hand side remains in the range of P. By orthogonality, both sides of the equation vanish, which imposes some constraints on $Q\omega P$ and $P\omega Q$. To be specific, one has

$$Q\omega\rho^2 = -QH\rho^2 \quad \Longleftrightarrow \quad Q\omega\Phi\sigma^2\Phi^\dagger = -QH\Phi\sigma^2\Phi^\dagger.$$

Right multiply $\Phi(\sigma^{-1})^2\Phi^\dagger$, one finds that $Q\omega P = -QHP$. Similarly, we obtain $P\omega Q = -PHQ$. The general solution of this system of matrix equations is

$$\omega = -QHP - PHQ - PGP - QGQ,$$

where G is any Hermitian matrix due to the hermiticity of ω. Since $\dot\Phi = \mathrm{i}\omega\Phi$, the equation for Φ can be written as

$$\begin{aligned}\mathrm{i}\dot\Phi = -\omega\Phi &= QHP\Phi + PHQ\Phi + PGP\Phi + QGQ\Phi \\ &= QH\Phi + PG\Phi,\end{aligned}$$

where the fact that $P\Phi = \Phi$ and $Q\Phi = 0$ is used. For the equation of ξ, (4.4.4) yields

$$\begin{aligned}\mathrm{i}\dot\sigma = \mathrm{i}\xi &= \mathrm{i}\Phi^\dagger\left(-\mathrm{i}[H,\rho] - \mathrm{i}[\omega,\rho]\right)\Phi \\ &= [\Phi^\dagger H\Phi,\sigma] - [\Phi^\dagger G\Phi,\sigma].\end{aligned}$$

Finally, we arrive at the dynamics for Φ and σ in a closed form

$$\mathrm{i}\partial_t\Phi = (I - \Phi\Phi^\dagger)H(t,\Phi\sigma\Phi^\dagger)\Phi + \epsilon\Phi\Phi^\dagger G\Phi, \tag{4.4.5}$$

$$\mathrm{i}\partial_t\sigma = \left[\Phi^\dagger\left(H(t,\Phi\sigma\Phi^\dagger) - G\right)\Phi, \sigma\right], \tag{4.4.6}$$

where the Hermitian matrix G is an extra degree of freedom to be chosen.

The next step is to find the optimal choice of G such that the dynamics of Φ changes the slowest, i.e. to find G such that

$$\min \|\dot{\Phi}\|_F^2 = \min \mathrm{Tr}(\dot{\Phi}^\dagger \dot{\Phi}). \tag{4.4.7}$$

The norm can be split into two parts

$$\begin{aligned} \|\dot{\Phi}\|_F^2 &= \|P\dot{\Phi}\|_F^2 + \|Q\dot{\Phi}\|_F^2 \\ &= \|\Phi\Phi^\dagger G\Phi\|_F^2 + \|QH\Phi\|_F^2. \end{aligned}$$

The Frobenius norm of the second term reads

$$\mathrm{Tr}(\Phi^\dagger H^\dagger Q^\dagger Q H \Phi) = \mathrm{Tr}(Q^\dagger Q H \Phi\Phi^\dagger H^\dagger) = \mathrm{Tr}(Q^\dagger Q H(t,\rho) P H^\dagger(t,\rho)),$$

which is independent of the gauge choice. Therefore, to optimize (4.4.7) one can choose $G = 0$. Now we arrive at the parallel transport dynamics

$$\begin{aligned} \mathrm{i}\partial_t \Phi &= (I - \Phi\Phi^\dagger) H(t, \Phi\sigma\Phi^\dagger)\Phi, \\ \mathrm{i}\partial_t \sigma &= \left[\Phi^\dagger H(t, \Phi\sigma\Phi^\dagger)\Phi, \sigma\right], \end{aligned} \tag{4.4.8}$$

which is equivalent to the PT dynamics (4.3.7) derived in Section 4.3.

It is worth pointing out that the Schrödinger gauge in fact corresponds to the choice $G = H$ in (4.4.5) that gives rise to

$$\begin{aligned} \mathrm{i}\partial_t \Phi &= H(t, \Phi\sigma\Phi^\dagger)\Phi, \\ \partial_t \sigma &= 0. \end{aligned} \tag{4.4.9}$$

This immediately implies that the number of occupied orbitals remains unchanged throughout the evolution, which verifies the validity of the solution in the Schrödinger dynamics in Eq. (4.3.3).

4.5 Numerical propagation of the parallel transport dynamics

In order to solve Eq. (4.3.7) numerically, for simplicity we assume that a uniform time discretization $t_n = nh$, and h is the time step size. The numerical values of $\Phi(t), \sigma(t), \rho(t), P(t)$ at time $t = t_n$ are denoted by $\Phi_n, \sigma_n, \rho_n, P_n$, respectively, and we define $H_n = H(t_n, \rho_n)$. Previous studies in Chapter 2 and Chapter 3 suggested that when the spectral radius of H

is large, the PT dynamics should be solved using implicit time integrators. This allows one to use a time step much larger than $\|H\|^{-1}$, and the result from the PT dynamics can be much more accurate than that from the Schrödinger dynamics using the same step size.

In order to discretize the PT dynamics with a mixed initial state, we consider the implicit midpoint (IM) rule (also known as the Gauss-Legendre method of order 2). We introduce the shorthand notations

$$\Phi_{n+\frac{1}{2}} = \frac{1}{2}(\Phi_n + \Phi_{n+1}), \quad \sigma_{n+\frac{1}{2}} = \frac{1}{2}(\sigma_n + \sigma_{n+1}), \tag{4.5.1}$$

and accordingly

$$P_{n+\frac{1}{2}} = \Phi_{n+\frac{1}{2}}(\Phi^\dagger_{n+\frac{1}{2}}\Phi_{n+\frac{1}{2}})^{-1}\Phi^\dagger_{n+\frac{1}{2}}, \quad \rho_{n+\frac{1}{2}} = \Phi_{n+\frac{1}{2}}\sigma_{n+\frac{1}{2}}\Phi^\dagger_{n+\frac{1}{2}}, \quad H_{n+\frac{1}{2}} = H\left(t_{n+\frac{1}{2}}, \rho_{n+\frac{1}{2}}\right). \tag{4.5.2}$$

We remark that $\rho_{n+\frac{1}{2}}$ is only a shorthand notation and may not be an admissible density matrix. In particular, even if $\mathrm{Tr}[\rho_n] = \mathrm{Tr}[\rho_{n+1}] = N_e$ (see Proposition 28), we may not have $\mathrm{Tr}[\rho_{n+\frac{1}{2}}] = N_e$. On the other hand, $P_{n+\frac{1}{2}}$ is still a projector satisfying $P_{n+\frac{1}{2}}\Phi_{n+\frac{1}{2}} = \Phi_{n+\frac{1}{2}}$.

With these notations, the parallel transport-implicit midpoint scheme (PT-IM) reads

$$\mathrm{i}\frac{\Phi_{n+1} - \Phi_n}{h} = (I - P_{n+\frac{1}{2}})H_{n+\frac{1}{2}}\Phi_{n+\frac{1}{2}}, \tag{4.5.3}$$

$$\mathrm{i}\frac{\sigma_{n+1} - \sigma_n}{h} = \left[\Phi^\dagger_{n+\frac{1}{2}}H_{n+\frac{1}{2}}\Phi_{n+\frac{1}{2}}, \sigma_{n+\frac{1}{2}}\right], \tag{4.5.4}$$

which form a set of nonlinear algebraic equations and need to be solved self-consistently. We can rewrite Eqs. (4.5.3) and (4.5.4) as

$$\begin{aligned} \Phi_{n+1} &= \Phi_n + \frac{h}{\mathrm{i}}(I - P_{n+\frac{1}{2}})H_{n+\frac{1}{2}}\Phi_{n+\frac{1}{2}}, \\ \sigma_{n+1} &= \sigma_n + \frac{h}{\mathrm{i}}\left[\Phi^\dagger_{n+\frac{1}{2}}H_{n+\frac{1}{2}}\Phi_{n+\frac{1}{2}}, \sigma_{n+\frac{1}{2}}\right]. \end{aligned} \tag{4.5.5}$$

If we choose $(\Phi_{n+1}, \sigma_{n+1})$ to be the unknowns and identify it with a vector $x \in \mathbb{C}^{N_g N + N^2}$, then Eqs. (4.5.3) and (4.5.4) can be viewed as a fixed point equation in the abstract form

$$x = T(x).$$

The structure of this fixed point problem resembles that of the self-consistent field iterations (SCF) in standard electronic structure calculations [116]. Here we use Anderson's mixing method [8] to solve this fixed problem.

The following proposition shows that PT-IM preserves the orthogonality as well as the trace condition.

Proposition 28. *Assume* $\Phi_n^\dagger \Phi_n = I_N$, $\sigma_n = \sigma_n^\dagger$, $\mathrm{Tr}[\sigma_n] = N_e$, *and that Eqs.* (4.5.3) *and* (4.5.4) *have a unique solution* $(\Phi_{n+1}, \sigma_{n+1})$, *then the solution satisfies*

$$\Phi_{n+1}^\dagger \Phi_{n+1} = I_N, \tag{4.5.6}$$

and

$$\sigma_{n+1}^\dagger = \sigma_{n+1}, \quad \mathrm{Tr}[\sigma_{n+1}] = N_e, \quad \mathrm{Tr}[\sigma_{n+1}^2] = \mathrm{Tr}[\sigma_n^2]. \tag{4.5.7}$$

As a consequence, we have $\mathrm{Tr}[\rho_{n+1}] = \mathrm{Tr}[\rho_n] = N_e$.

Proof. First, use the definition in (4.5.1) and apply $\Phi^\dagger_{n+\frac{1}{2}}$ to both sides of (4.5.3), we obtain

$$\frac{\mathrm{i}}{2h}(\Phi_{n+1}^\dagger \Phi_{n+1} - \Phi_n^\dagger \Phi_n) - \frac{\mathrm{i}}{2h}(\Phi_{n+1}^\dagger \Phi_n - \Phi_n^\dagger \Phi_{n+1}) = \Phi^\dagger_{n+\frac{1}{2}}(I - P_{n+\frac{1}{2}}) H_{n+\frac{1}{2}} \Phi_{n+\frac{1}{2}} = 0.$$

On the left-hand side, the first term is anti-Hermitian and the second term is Hermitian. So both terms must vanish, and

$$\Phi_{n+1}^\dagger \Phi_{n+1} = \Phi_n^\dagger \Phi_n = I_N.$$

This proves Eq. (4.5.6).

Second, denote by $\widetilde{H} := \Phi^\dagger_{n+\frac{1}{2}} H_{n+\frac{1}{2}} \Phi_{n+\frac{1}{2}}$, we may solve the equation

$$\sigma_{n+1} - \sigma_n = -\frac{\mathrm{i}h}{2}[\widetilde{H}, \sigma_n] - \frac{\mathrm{i}h}{2}[\widetilde{H}, \sigma_{n+1}]$$

to obtain σ_{n+1}. Applying the Hermite conjugation to both sides and using that $\widetilde{H}, \sigma_n$ are Hermitian matrices, we have

$$\sigma_{n+1}^\dagger - \sigma_n = -\frac{\mathrm{i}h}{2}[\widetilde{H}, \sigma_n] - \frac{\mathrm{i}h}{2}[\widetilde{H}, \sigma_{n+1}^\dagger].$$

The uniqueness of σ_{n+1} implies $\sigma_{n+1} = \sigma_{n+1}^\dagger$. Moreover, since the right-hand side of Eq. (4.5.4) is traceless, we have $\mathrm{Tr}[\sigma_{n+1}] = \mathrm{Tr}[\sigma_n]$.

Finally, applying $\sigma_{n+\frac{1}{2}}$ from the left to both sides of Eq. (4.5.4), we have

$$\frac{\mathrm{i}}{2h}(\sigma_{n+1}^2 - \sigma_n^2 - \sigma_{n+1}\sigma_n + \sigma_n \sigma_{n+1}) = \sigma_{n+\frac{1}{2}}[\widetilde{H}, \sigma_{n+\frac{1}{2}}].$$

Since the right-hand side is traceless, by taking trace of both sides we obtain

$$\mathrm{Tr}[\sigma_{n+1}^2] = \mathrm{Tr}[\sigma_n^2].$$

This finishes the proof of the equalities in (4.5.7). □

Eq. (4.5.6) can be viewed as a consequence of the general fact that the PT-IM method preserves quadratic invariants, and in particular orthogonality constraints (see e.g. [68, pp 132] for a more general description of orthogonality preserving Runge-Kutta methods). Proposition 28 confirms that the PT-IM scheme preserves orthogonality of $\Phi(t)$, as well as the number of electrons.

4.6 Error analysis

In this section, we consider the numerical error of the PT-IM scheme for concreteness, and compare the form of the error terms with those from the Schrödinger dynamics. The error analysis can also be extended to other Runge-Kutta methods and linear multistep methods.

Error analysis of the PT dynamics

Before proceeding with the detailed error analysis, we first provide some abstract perspectives. Let the local truncation error be defined as $e_k(X) = X(t_k) - \widetilde{X}_k$, where $t_k = kh$ and $\widetilde{X}_k$ represents the numerical solution of X at the k-th step with previous step to be exactly $X(t_{k-1})$, where X is the concatenation of Φ and σ, namely $\begin{pmatrix}\Phi\\ \sigma\end{pmatrix}$. Since IM is a second order method, the local truncation error can be bounded in terms of the third order derivatives [69]

$$\|e_k(X)\| \leqslant C \max_{t\in[t_{k-1},t_k]} \|\partial_t^3 X(t)\| h^3. \tag{4.6.1}$$

Here C is an absolute constant depending only on the choice of the numerical scheme.

Note that P is a rank-N projector, and can be identified with the Grassmann manifold $\mathrm{Gr}(N_g, N; \mathbb{C})$, i.e. the N-dimensional subspace of $\mathbb{C}^{N_g}$. On the other hand, the gauge-transformed wavefunctions Φ belongs to the Stiefel manifold $\mathrm{St}(N_g, N; \mathbb{C})$, which is the set of first N columns of an N_g-dimensional unitary matrices. The Grassmann manifold is the quotient space of $\mathrm{St}(N_g, N; \mathbb{C})$ by $\mathrm{U}(N)$, denoted by

$$\mathrm{Gr}(N_g, N; \mathbb{C}) = \mathrm{St}(N_g, N; \mathbb{C})/\mathrm{U}(N)$$

The projector $P(t)$ can be identified with a curve in $\mathrm{Gr}(N_g, N; \mathbb{C})$, obtained by solving the von Neumann equation. On the other hand, the wavefunctions $\Psi(t), \Phi(t)$ in the Schrödinger and the parallel transport gauge are *lifts* of the curve $P(t)$ from the quotient space to $\mathrm{St}(N_g, N; \mathbb{C})$. In particular, $\Phi(t)$ can be identified as the unique *horizontal lift* [119] of $P(t)$, starting from the initial condition Ψ_0 (which fixes a gauge choice initially). We have demonstrated that the parallel transport gauge yields the slowest dynamics in the sense of minimizing $\|\partial_t \Phi\|_F$. For simplicity, in the following discussions we will consider the operator norm $\|\cdot\|$ for Φ, P and their time derivative. We expect that the size of the k-th order time derivative $\|\partial_t^k \Phi\|$ should also be bounded by that of $\|\partial_t^k P\|$. On the other hand, $\|\partial_t^k \Psi\|$ may not be bounded by $\|\partial_t^k P\|$ due to the gauge matrix.

Recall the relation

$$P\Phi = \Phi, \quad P\partial_t \Phi = 0,$$

and this gives

$$\partial_t \Phi = (\partial_t P)\Phi.$$

Keep differentiating and obtain

$$\partial_t^2\Phi = [\partial_t^2 P + (\partial_t P)^2]\Phi, \quad \partial_t^3\Phi = [\partial_t^3 P + 2(\partial_t^2 P)(\partial_t P) + (\partial_t P)(\partial_t^2 P) + (\partial_t P)^3]\Phi.$$

Using the fact that $\|\Phi\| = 1$, we have

$$\|\partial_t\Phi\| \leqslant \|\partial_t P\|.$$

Similarly

$$\|\partial_t^2\Phi\| \leqslant \|\partial_t^2 P + (\partial_t P)^2\| \leqslant \|\partial_t^2 P\| + \|(\partial_t P)\|^2,$$

and

$$\|\partial_t^3\Phi\| \leqslant \|\partial_t^3 P\| + 3\|\partial_t^2 P\|\|\partial_t P\| + \|\partial_t P\|^3.$$

This implies that $\|\partial_t^k\Phi\|$ is controlled by $\|\partial_t^\ell P\|$ with $\ell \leqslant k$. On the other hand,

$$\sigma = \Phi^\dagger\rho\Phi$$

implies that the time derivative $\|\partial_t^k\sigma(t)\|$ is controlled by $\|\partial_t^\ell\Phi\|$ and $\|\partial_t^\ell\rho\|$ with $\ell \leqslant k$. To be specific, a direct computation gives

$$\begin{aligned}\partial_t^3\sigma =&(\partial_t^3\Phi^\dagger)\rho\Phi + \Phi^\dagger(\partial_t^3\rho)\Phi + \Phi^\dagger\rho(\partial_t^3\Phi) + 6(\partial_t\Phi^\dagger)(\partial_t\rho)(\partial_t\Phi)\\ &+ 3(\partial_t^2\Phi^\dagger)(\partial_t\rho)\Phi + 3(\partial_t^2\Phi^\dagger)\rho(\partial_t\Phi) + 3(\partial_t\Phi^\dagger)(\partial_t^2\rho)\Phi\\ &+ 3(\partial_t\Phi^\dagger)\rho(\partial_t^2\Phi) + 3\Phi^\dagger(\partial_t^2\rho)(\partial_t\Phi) + 3\Phi^\dagger(\partial_t\rho)(\partial_t^2\Phi),\end{aligned}$$

and hence

$$\begin{aligned}\|\partial_t^3\sigma\| \leqslant&2\|\partial_t^3\Phi\| + \|\partial_t^3\rho\| + 6\|\partial_t\Phi^\dagger\|\|\partial_t\rho\|\|\partial_t\Phi\|\\ &+ 6\|\partial_t^2\Phi\|\|\partial_t\rho\| + 6\|\partial_t^2\Phi\|\|\partial_t\Phi\| + 6\|\partial_t\Phi\|\|\partial_t^2\rho\|,\end{aligned}$$

where we used the facts that $\|\Phi\| = 1$ and $\|\rho\| \leqslant 1$.

To summarize, the error analysis of the PT dynamics is reduced to the estimate of $\|\partial_t^k P\|$ and $\|\partial_t^k\rho\|$. In particular, for the analysis of PT-IM, we need $k \leqslant 3$.

Lemma 29. *Suppose $H(t,\rho)$ is continuously differentiable in terms of t and ρ up to second order. Then the derivatives of P satisfy*

$$\|\partial_t P\| \leqslant \|[H,P]\|, \tag{4.6.2}$$

$$\|\partial_t^2 P\| \leqslant \|[H_t,P]\| + \|H_\rho[H,\rho]\| + \|[H,[H,P]]\|, \tag{4.6.3}$$

$$\begin{aligned}\|\partial_t^3 P\| \leqslant&\|[H_{tt},P]\| + 2\|(H_t)_\rho[H,\rho]\| + \|H_{\rho\rho}([H,\rho])^2\| + \|H_\rho[H_t,\rho]\|\\ &+ \|H_\rho[H_\rho[H,\rho],\rho]\| + \|H_\rho[H,[H,\rho]]\| + 2\|[H_t,[H,P]]\| + 2\|[H_\rho[H,\rho],[H,P]]\|\\ &+ \|[H,[H_t,P]]\| + \|[H,[H_\rho[H,\rho],P]]\| + \|[H,[H,[H,P]]]\|,\end{aligned} \tag{4.6.4}$$

where the subscripts denote the partial derivatives.

Proof. The first inequality is trivial. To prepare for the differentiation of P, we start by computing the derivatives of H. For notational simplicity, we use the subscripts to denote the partial derivative and omit the explicit $(t, \rho(t))$ dependence in H. The first order derivative of H reads

$$\dot{H} := \frac{d}{dt} H(t, \rho(t)) = H_t + H_\rho \rho_t = H_t - \mathrm{i} H_\rho [H, \rho], \tag{4.6.5}$$

and the second order derivative of H is given by

$$\begin{aligned} \ddot{H} :=& \frac{d^2}{dt^2} H(t, \rho(t)) = \frac{d}{dt} H_t - \mathrm{i} \frac{d}{dt} (H_\rho [H, \rho]) \\ =& H_{tt} + (H_t)_\rho \rho_t - \mathrm{i} (H_t)_\rho [H, \rho] - \mathrm{i} H_{\rho\rho} \rho_t [H, \rho] - \mathrm{i} H_\rho [\dot{H}, \rho] - \mathrm{i} H_\rho [H, \rho_t]. \end{aligned}$$

It follows from $\mathrm{i}\partial_t \rho = [H, \rho]$ that

$$\begin{aligned} \ddot{H} =& H_{tt} - 2\mathrm{i} (H_t)_\rho [H, \rho] - H_{\rho\rho} ([H, \rho])^2 \\ & - \mathrm{i} H_\rho [H_t, \rho] - H_\rho [H_\rho [H, \rho], \rho] - H_\rho [H, [H, \rho]]. \end{aligned} \tag{4.6.6}$$

The second order derivative of P becomes

$$\begin{aligned} \partial_t^2 P =& -\mathrm{i} \frac{d}{dt} ([H(t, \rho(t)), P(t)]) = -\mathrm{i}[\dot{H}, P] - \mathrm{i}[H, \partial_t P] \\ =& -\mathrm{i}[H_t, P] - [H_\rho [H, \rho], P] - [H, [H, P]], \end{aligned} \tag{4.6.7}$$

together with the fact that $\|P\| \leqslant 1$, we obtain (4.6.3). Similarly, the third order derivative of P can be computed explicitly via

$$\partial_t^3 P = -\mathrm{i}[\ddot{H}, P] - 2\mathrm{i}[\dot{H}, \partial_t P] - \mathrm{i}[H, \partial_t^2 P].$$

Plugging in (4.6.5), (4.6.6) and (4.6.7), one obtains

$$\begin{aligned} \partial_t^3 P =& -\mathrm{i}[H_{tt}, P] - 2[(H_t)_\rho [H, \rho], P] + \mathrm{i}[H_{\rho\rho} ([H, \rho])^2, P] - [H_\rho [H_t, \rho], P] \\ & + \mathrm{i}[H_\rho [H_\rho [H, \rho], \rho], P] + \mathrm{i}[H_\rho [H, [H, \rho]], P] - 2[H_t, [H, P]] \\ & + 2\mathrm{i}[H_\rho [H, \rho], [H, P]] - [H, [H_t, P]] + \mathrm{i}[H, [H_\rho [H, \rho], P]] + \mathrm{i}[H, [H, [H, P]]]. \end{aligned}$$

Taking the norm yields the desired result. □

Lemma 30. *Suppose $H(t, \rho)$ is continuously differentiable in terms of t and ρ up to second order. The derivatives of ρ satisfy*

$$\|\partial_t \rho\| \leqslant \|[H, \rho]\|, \tag{4.6.8}$$

$$\|\partial_t^2 \rho\| \leqslant \|[H_t, \rho]\| + \|[H_\rho [H, \rho], \rho]\| + \|[H, [H, \rho]]\|, \tag{4.6.9}$$

$$\begin{aligned} \|\partial_t^3 \rho\| \leqslant& \|[H_{tt}, \rho]\| + 2\|[(H_t)_\rho [H, \rho], \rho]\| + \|[H_{\rho\rho} ([H, \rho])^2, \rho]\| + \|[H_\rho [H_t, \rho], \rho]\| \\ & + \|[H_\rho [H_\rho [H, \rho], \rho], \rho]\| + \|[H_\rho [H, [H, \rho]], \rho]\| + 2\|[H_t, [H, \rho]]\| + \|[H, [H_t, \rho]]\| \\ & + 2\|[H_\rho [H, \rho], [H, \rho]]\| + \|[H, [H_\rho [H, \rho], \rho]]\| + \|[H, [H, [H, \rho]]]\|, \end{aligned} \tag{4.6.10}$$

where the subscripts denote the partial derivatives.

Proof. The proof is similar as Lemma 29 since ρ satisfies the equation $\mathrm{i}\partial_t\rho = [H,\rho]$, which has the same form of that for P. □

Therefore, the local truncation errors of the PT dynamics can be bounded by terms involving commutators of $[H,P]$, $[H,\rho]$, $[H_t,\rho]$, $[H_{tt},\rho]$, $[H_t,P]$, $[H_{tt},P]$.

Comparison to the Schrödinger dynamics

In this section, we discuss the local truncation error of the Schrödinger dynamics and the global errors of the PT and Schrödinger dynamics. The local truncation error can be summarized in the following lemma. Note that in the bound, we keep the wavefunction Ψ for the terms without commutator structures, such as $\|H^3\Psi\|$, instead of replacing it by the operator norm $\|H^3\|$, because the latter could be significantly larger than the former.

Lemma 31. *For the IM scheme, the local truncation errors of Schrödinger dynamics* (4.2.1) *can be bounded as*

$$\begin{aligned}\|e_k(\Psi)\| \leqslant & C\big(\|H^3\Psi\| + \|HH_t\Psi\| + 2\|H_tH\Psi\| + \|H_{tt}\Psi\| \\ & + \|HH_\rho[H,\rho]\| + 2\|H_\rho[H,\rho]H\| + 2\|(H_t)_\rho[H,\rho]\| + \|H_{\rho\rho}([H,\rho])^2\| \\ & + \|H_\rho[H_t,\rho]\| + \|H_\rho[H_\rho[H,\rho],\rho]\| + \|H_\rho[H,[H,\rho]]\|\big), \end{aligned} \tag{4.6.11}$$

for some constant C that does not depend on t_k, h.

Proof. It suffices to calculate the derivatives of Ψ. The second order derivative is computed as

$$\partial_t^2\Psi = -\mathrm{i}H\partial_t\Psi - \mathrm{i}\dot{H}\Psi = -H^2\Psi - \mathrm{i}\dot{H}\Psi$$

and the third order derivative can be computed as

$$\begin{aligned}\partial_t^3\Psi =& -\mathrm{i}H\ddot{\Psi} - 2\mathrm{i}\dot{H}\dot{\Psi} - \mathrm{i}\ddot{H}\Psi \\ =& \mathrm{i}H^3\Psi - H\dot{H}\Psi - 2\dot{H}H\Psi - \mathrm{i}\ddot{H}\Psi \\ =& \mathrm{i}H^3\Psi - HH_t\Psi + \mathrm{i}HH_\rho[H,\rho]\Psi - 2H_tH\Psi \\ & + 2\mathrm{i}H_\rho[H,\rho]H\Psi - \mathrm{i}H_{tt}\Psi + 2(H_t)_\rho[H,\rho]\Psi + \mathrm{i}H_{\rho\rho}([H,\rho])^2\Psi \\ & - H_\rho[H_t,\rho]\Psi + \mathrm{i}H_\rho[H_\rho[H,\rho],\rho]\Psi + \mathrm{i}H_\rho[H,[H,\rho]]\Psi. \end{aligned}$$

Taking the norm and applying (4.6.1), we obtain the desired result.

□

Lemma 29, Lemma 30 and Lemma 31 give the local truncation error errors of both PT and Schrödinger dynamics. Following the standard stability analysis [102], we obtain the global error bounds.

Theorem 32 (Global error)**.** *For the IM schemes of* (4.2.1) *and* (4.3.7) *up to the time* $t_n = T$*, there exists some constant* C *depending on* T *and* $\|H\|$ *such that*

1. *the errors for the PT dynamics* (4.3.7) *satisfy*

$$\|\Phi(t_n) - \Phi_n\| + \|\sigma(t_n) - \sigma_n\| \leq C f_1(H, \rho, P) h^2, \tag{4.6.12}$$

where f_1 *is a function of* H, ρ, P *that is a linear combination of products of nested commutators up to three layers of the form*

$$\|[A_4, A_3[A_2, A_1 A_0]]\| \tag{4.6.13}$$

with A_0 *being one of the following*

$$[H, P], \quad [H_t, P], \quad [H_{tt}, P], \quad [H, \rho], \quad [H_t, \rho], \quad [H_{tt}, \rho] \tag{4.6.14}$$

and A_i *(*$i = 1, \cdots, 4$*) being the identity matrix* I*, functions of* H, ρ, P *or derivatives of* H*.*

2. *the error for the Schrödinger dynamics* (4.2.1) *satisfies*

$$\|\Psi(t_n) - \Psi_n\| \leq C \left(f_2(H, \rho, P) + \|H^3 \Psi\| + \|H H_t \Psi\| + 2\|H_t H \Psi\| + \|H_{tt} \Psi\| \right) h^2, \tag{4.6.15}$$

where f_2 *has the same form as* f_1*.*

Theorem 32 shows that the error bound of PT dynamics exhibits commutator scaling while that of the Schrödinger equation does not. We remark that the worst-case dependence of the constant C on the norm of H can be very pessimistic, which is due to the standard stability analysis through the Grönwall type estimates. However, the Schrödinger equation inherits a Hamiltonian structure and, together with the fact that IM is a symplectic scheme, this preconstant C may be dramatically improved such that it depends linearly on T and is even possibly independent of $\|H\|$ [68]. In order to formally employ the symplectic properties, however, the PT-IM scheme needs to be slightly modified. This has been demonstrated in Chapter 2 for pure states. Numerical results in Chapter 2 also demonstrate that the performance of the schemes with and without the modification are almost the same, so the modification may only be of theoretical interest. For simplicity, we do not detail such modification here.

Near adiabatic regime

In the near adiabatic regime, we can use commutator structure to demonstrate provable advantage of the PT dynamics over the Schrödinger dynamics. Consider the singular perturbed Schrödinger equation:

$$i\epsilon \partial_t \Psi^\epsilon(t) = H(t, \rho^\epsilon(t)) \Psi^\epsilon(t), \quad \epsilon \ll 1. \tag{4.6.16}$$

Here $\rho^\epsilon(0)$ is a pure state, and $\Psi^\epsilon(0)$ consists of the eigenfunctions of $H(0,\rho^\epsilon(0))$ corresponding to the algebraically lowest N eigenvalues.

Let $\rho^\epsilon = P^\epsilon = \Psi^\epsilon \Psi^{\epsilon\,\dagger}$. Then the PT dynamics become

$$\mathrm{i}\epsilon\partial_t \Phi^\epsilon(t) = H(t,\rho^\epsilon(t))\Phi^\epsilon(t) - \Phi^\epsilon(t)(\Phi^{\epsilon\,\dagger}(t)H(t,\rho^\epsilon(t))\Phi^\epsilon(t)), \quad \Phi^\epsilon(0) = \Psi^\epsilon(0). \tag{4.6.17}$$

In the linear case ($H(t,\rho(t)) = H(t)$ is independent of ρ), if the gap condition is satisfied, i.e. there exists a positive gap between the N-th and $(N+1)$-th smallest eigenvalues of $H(t)$ for all $t \in [0,T]$, The adiabatic theorem (see for example, [147, 67, 90, 91]) for the Schrödinger dynamics (4.6.16) states that

$$\Psi^\epsilon(t) = \Psi_a(t) + \mathcal{O}(\epsilon), \tag{4.6.18}$$

where the columns of $\Psi_a(t)$ are the eigenvectors of the Hamiltonian, namely, there exists a time-dependent diagonal matrix $\Lambda(t)$ whose diagonal entries are eigenvalues of the Hamiltonian such that

$$H(t)\Psi_a(t) = \Psi_a(t)\Lambda(t).$$

The adiabatic theorem can also be generalized to certain linear systems without a gap condition [9, 148], and for some weakly nonlinear systems [57, 61]. A detailed discussion of the technical conditions for the adiabatic approximation is beyond the scope of this chapter. However, when such *a priori* estimate is available, we can evaluate the commutator as

$$[H,\rho^\epsilon] = H\Psi_a\Psi_a^\dagger - \Psi_a\Psi_a^\dagger H + \mathcal{O}(\epsilon) = \Psi_a\Lambda\Psi_a^\dagger - \Psi_a\Lambda\Psi_a^\dagger + \mathcal{O}(\epsilon) = \mathcal{O}(\epsilon). \tag{4.6.19}$$

We now examine the commutator terms in Lemma 30. Note that in the singular perturbed regime, one should replace the H in Lemma 30 by H/ϵ, and hence the leading order terms in ϵ are

$$\begin{aligned}&\epsilon^{-3}\|[H_{\rho^\epsilon\rho^\epsilon}([H,\rho^\epsilon])^2,\rho^\epsilon]\| + \epsilon^{-3}\|[H_{\rho^\epsilon}[H_{\rho^\epsilon}[H,\rho^\epsilon],\rho^\epsilon],\rho^\epsilon]\| + \epsilon^{-3}\|[H_{\rho^\epsilon}[H,[H,\rho^\epsilon]],\rho^\epsilon]\| \\ &+ \epsilon^{-3}\|[H_{\rho^\epsilon}[H,\rho^\epsilon],[H,\rho^\epsilon]]\| + \epsilon^{-3}\|[H,[H_{\rho^\epsilon}[H,\rho^\epsilon],\rho^\epsilon]]\| + \epsilon^{-3}\|[H,[H,[H,\rho^\epsilon]]]\| = \mathcal{O}(\epsilon^{-2}),\end{aligned}$$

thanks to (4.6.19). However, by replacing H in Lemma 31 by H/ϵ, we obtain $\|\partial_t^3\Psi\| = \mathcal{O}(\epsilon^{-3})$ for the Schrödinger dynamics. Finally applying Theorem 32, we find that the numerical schemes for the PT dynamics can gain an order of magnitude in terms of the accuracy in ϵ, which recovers the result in Chapter 2 for the linear case, and generalizes the result to the nonlinear case (provided that adiabatic theorems can be established).

4.7 Numerical Results

In this section, we provide the numerical results of the parallel transport dynamics. We focus on the case of a mixed initial state in this section. In numerical examples, the relative

numerical errors are computed by

$$\sup_{0\leq k\leq n} \frac{\|X_k - X(t_k)\|}{\|X(t_k)\|},$$

where $k = 0, 1, \cdots, n$ and t_n is the final time, and X is the reference values of Φ, σ or Ψ with X_n represents the numerical results of the quantity X at the time t_n.

Our model system is defined by a periodic potential field given by a one-dimensional lattice structure with Hamiltonian

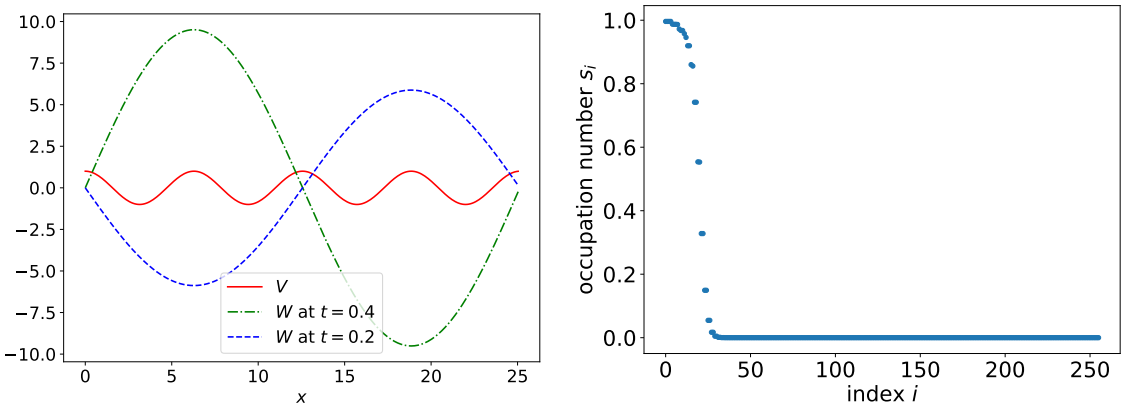
$$H(t) = -\frac{1}{2}\Delta + V(x) + W(x,t). \tag{4.7.1}$$

Here $V(x) = \cos(x)$ is a static potential. The external time-dependent potential with frequency ω is

$$W(x,t) = 10\sin\left(\frac{x}{L}\right)\sin(\omega t), \tag{4.7.2}$$

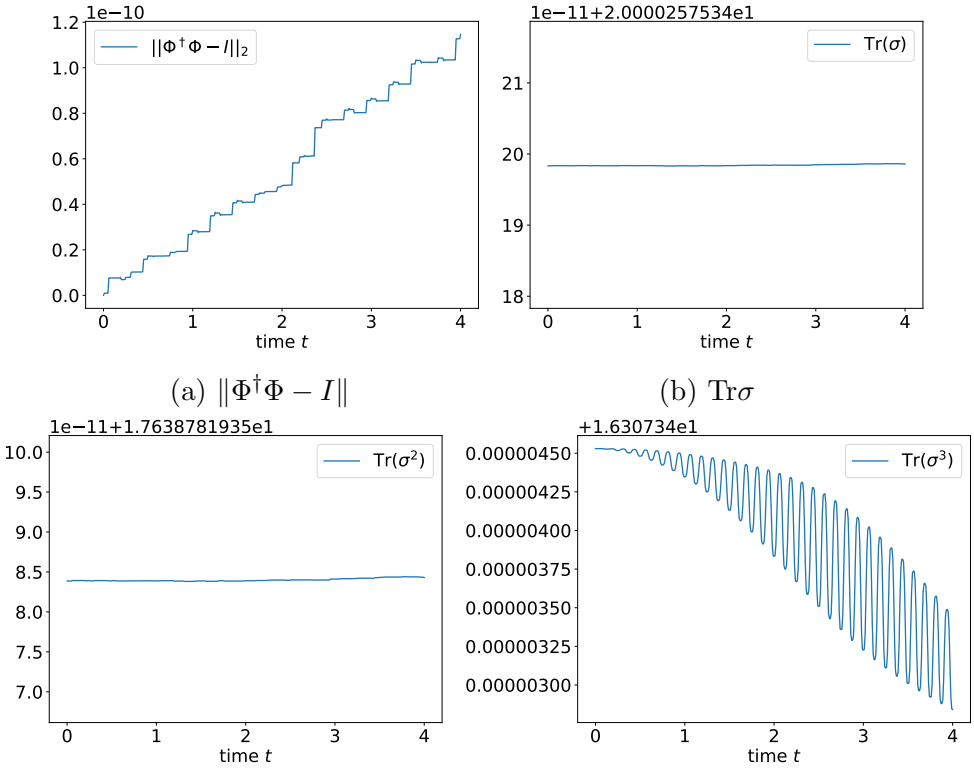
and L denotes the number of unit cells. The length of the lattice (the computational domain) is $2\pi L$. Fig. 4.7.1a shows a typical plot for the two potentials over the lattice cells. The parameters in the system are chosen as $L = 4$, $\beta = 1.453$, $\omega = 16\pi$ and the chemical potential $\mu = 3.299$. The initial occupation number according to the Fermi-Dirac distribution is in Fig. 4.7.1b. Each unit cell is discretized via 64 equidistant grid points, and hence the total number of grid points is $N_g = 64L$.

We first verify the Proposition 28 numerically by simulating the PT dynamics to $T_{\text{final}} = 4$ using the PT-IM scheme with a step size $h = 0.01$. We set $N_e = 20$, and $N = 64$. The norm of $\Phi_{n+1}^{\dagger}\Phi_{n+1}$ and values of $\text{Tr}\sigma_n$ and $\text{Tr}\sigma_n^2$ are plotted for the simulation time in Fig. 4.7.2. It can be seen that the values of all three quantities are constant throughout the simulation, which agrees with Proposition 28. In comparison, we also plot the higher order trace $\text{Tr}\sigma_n^3$, which is not a conserved quantity. Nonetheless, the fluctuation of $\text{Tr}\sigma_n^3$ is still very small and on the order of 10^{-6}.



(a) the potentials (b) the initial occupation

Figure 4.7.1: Left panel: The potentials $V(x)$ (red solid) and $W(t,x)$ at time $t = 0.2$ (blue dashed) and $t = 0.4$ (green dotted), respectively, where W is of time period $1/8$. Right panel: The initial occupation of the Fermi-Dirac statistics. $L = 4$, $\beta = 1.453$, and the chemical potential is chosen such that the initial number of occupation $N_e = \mathrm{Tr}(\rho(0)) = 20$.



(a) $\|\Phi^\dagger\Phi - I\|$ (b) $\mathrm{Tr}\sigma$

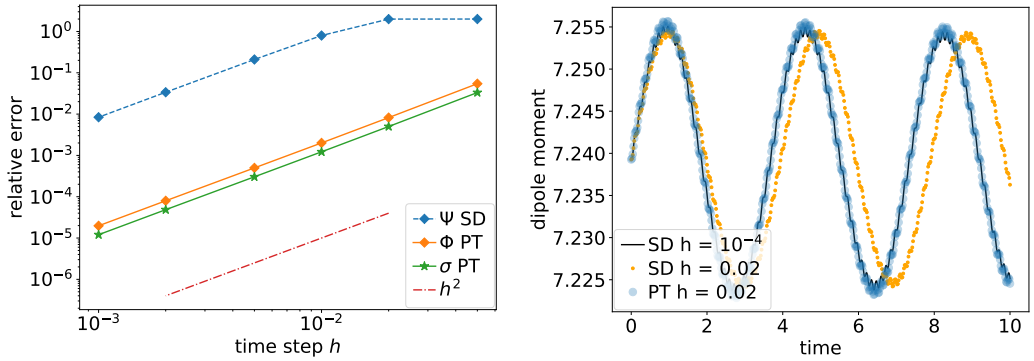
(c) $\mathrm{Tr}\sigma^2$ (d) $\mathrm{Tr}\sigma^3$

Figure 4.7.2: Numerical verification (time step $h = 0.01$) of the orthogonality of Φ and the trace preservation of σ and σ^2, as shown in Proposition 28. On the other hand, the trace of higher powers of σ (e.g. σ^3) may not be preserved in the PT-IM scheme.

Next, we compare the numerical errors in simulating the Schrödinger dynamics (SD) and PT dynamics. Both dynamics are simulated using IM schemes to $T_{\text{final}} = 1$. We set $\mu = 26.893$ (corresponding to $N_e = 60$) and $N = 80$. In order to verify the convergence rate numerically, we set the time steps to be 0.05, 0.02, 0.01, 0.005, 0.002, 0.001. The reference solution is computed using a fine time step of 2×10^{-5}. Fig. 4.7.3a shows that both SD-IM and PT-IM are second order methods, but the preconstant of PT-IM is much smaller. The accuracy of the PT dynamics can also be shown in terms of physical observables, e.g. the dipole moment:

$$\langle x(t) \rangle := \text{Tr}(x\rho(t)).$$

Fig. 4.7.3b compares the dipole moment computed in three different scenarios: PT-IM with $h = 0.02$, SD-IM with $h = 0.02$, and SD-IM using a very small time step $h = 0.0001$. We find that the difference between the time-dependent dipole moment obtained from PT-IM with a large time step $h = 0.02$ is almost the same as that from the reference solution. However, SD-IM with the same time step size is only accurate for a short periodic of time, and its accuracy significantly deteriorates as t increases.



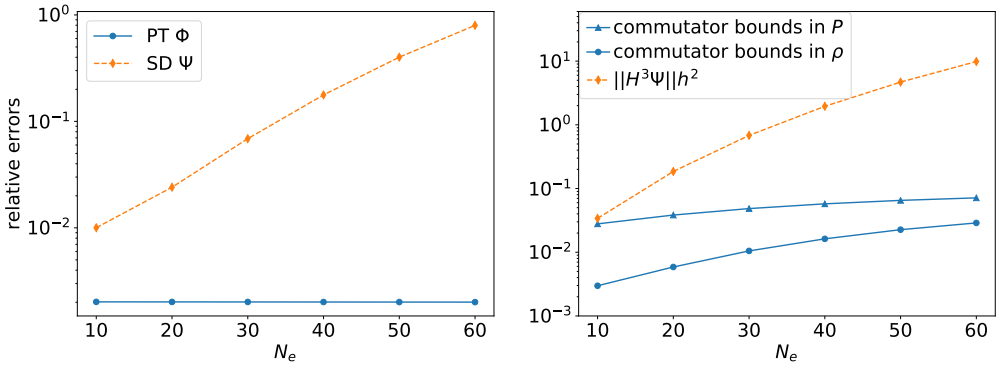
(a) Relative Errors

(b) Dipole moment

Figure 4.7.3: Left Panel: Log-log plot of the relative errors of Φ, σ, Ψ and the density matrix ρ computed via both PT and Schrödinger dynamics (SD). Right Panel: Evolution of the dipole moment for PT-IM with $h = 0.02$, SD-IM with $h = 0.02$, and SD-IM with $h = 0.0001$ (reference solution).

In order to demonstrate that the commutator scaling in Theorem 32, we now vary the number of electrons N_e, and compare the results of PT-IM and SD-IM. The chemical potential μ is set to 3.299, 7.028, 12.291, 18.951, 26.893, and the corresponding N_e are 10, 20, 30, 40, 50, 60, respectively. We also set $N = N_e + 20, h = 0.01$, and $T_{\text{final}} = 1$. The reference solution is computed using a very small step size $h = 2 \times 10^{-5}$.

We plot the relative errors of both the PT and the Schrödinger dynamics in comparison with our theoretical bounds. It can be seen in Fig. 4.7.4 that as N_e increases, the relative error of the wavefunction in the Schrödinger dynamics grows much faster than that in the PT dynamics. Fig. 4.7.4 also plots the terms in the error bounds with or without the commutator structures, respectively. We find that the term without commutator structures can be much larger in magnitude, and the qualitative trend of the growth of the error bound with respect to N_e matches that of the error from the numerical simulation.



(a) errors in wavefunctions (b) Commutator bounds and $\|H^3\Psi\|h^2$

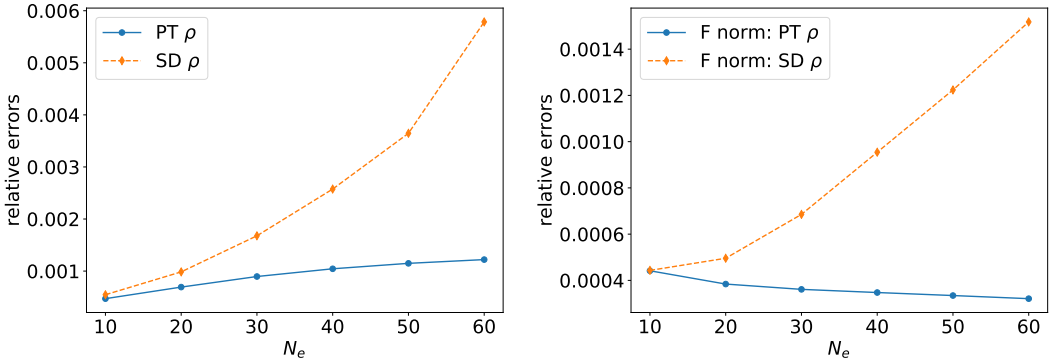
Figure 4.7.4: Relative errors versus the number of occupation N_e in semi-log scale. Left panel: relative errors for the wavefunctions Φ in PT gauge and Ψ in the Schrödinger gauge. Right panel: the commutator bounds on the right-hand-sides of $\partial_t^3 P$ in Lemma 29 and $\partial_t^3 \rho$ in Lemma 30 versus $\|H^3\Psi\|h^2$ that appears in Lemma 31. The commutator bounds (as in PT) are significant smaller than $\|H^3\Psi\|h^2$ term (as in the Schrödinger gauge).

We also plot the relative errors in 2-norm of the density matrix ρ in Fig. 4.7.5a and Fig. 4.7.5b. The errors (measured in both the operator norm and the Frobenius norm) from the PT dynamics is smaller than that from the Schrödinger dynamics. Furthermore, as N_e increases, the relative error in the Frobenius norm from the PT dynamics in fact decreases. This phenomenon can be intuitively explained as follows. Note that the initial σ_0 is a diagonal matrix of the following form

$$\begin{pmatrix} I_{m_0} & 0 & 0 \\ 0 & \sigma_* & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

where m_0 is the number of fully occupied states and σ_* is a diagonal matrix representing the fractional states whose diagonal elements has values in $(0, 1)$. Then we expect that the

fully occupied states are approximately in the near adiabatic regime, and their contribution to the error is much smaller than those from the fractionally occupied ones according to the commutator bound. In this example, m_0 increases with N_e, but the size of σ_* does not change much with respect to N_e. Therefore, the error of the density matrix should be dominated by a small number of orbitals near the Fermi surface. To verify this statement, we plot in Fig. 4.7.6 the histogram of the errors in the vector 2-norm for all orbitals. Indeed, as N_e increases, the errors are dominant by only a few orbitals near the corresponding chemical potential μ, and the number of the orbitals with significant errors does not increase with N_e. On the other hand, the Frobenius norm of the density matrix $\|\rho\|_F = \mathcal{O}(\sqrt{N_e})$. This explains the decay of the relative error of ρ in the PT-dynamics in Fig. 4.7.5b. By comparison, the histogram of the errors in the vector 2-norm for all orbitals in the Schrödinger gauge is provided in Fig. 4.7.7. We find that in the Schrödinger dynamics, the errors are propagated much more widely along the energy spectrum among a larger number of orbitals. It is also interesting to note that the maximal magnitude of the error increases significantly with respect to N_e in the Schrödinger dynamics, but the maximal error is nearly a constant and is much smaller in the PT dynamics.



(a) errors of ρ in 2-norm

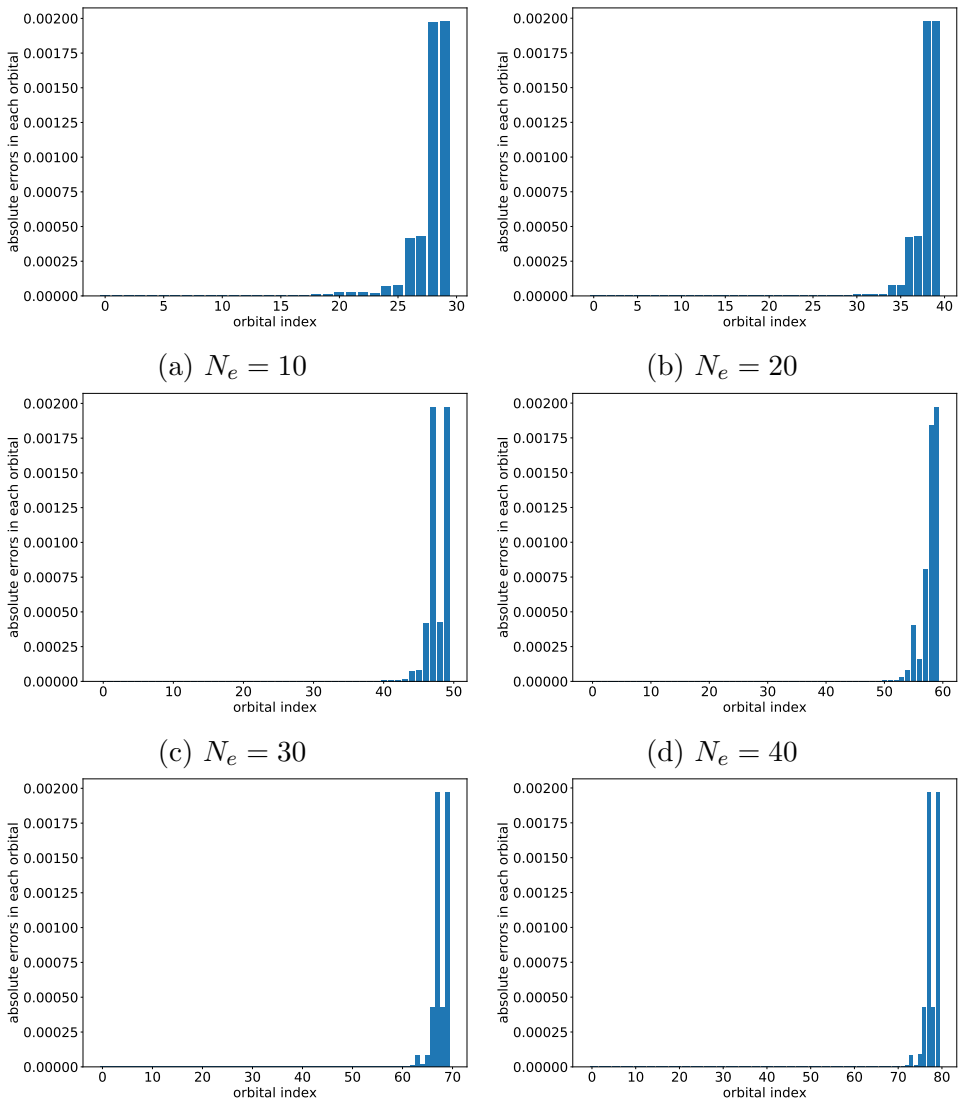
(b) errors of ρ in F-norm

Figure 4.7.5: Plots of the relative errors of ρ versus the number of occupation N_e. Left panel: relative errors in 2-norm. Right panel: the relative errors in the Frobenius norm (F-norm).

Finally, we demonstrate that the PT dynamics remains equally effective in the nonlinear regime. The rt-TDDFT Hamiltonian takes the following general form

$$H(t, \rho(t)) = -\frac{1}{2}\Delta + V_{\text{ext}}(x, t) + V_{\text{Hxc}}[\rho(t)] + V_{\text{X}}[\rho(t)], \tag{4.7.3}$$

where V_{ext} represents the electron-ion interaction and when the external field changes with respect to time, V_{ext} may also depend on time t. V_{Hxc} is the Hartree and local exchange-

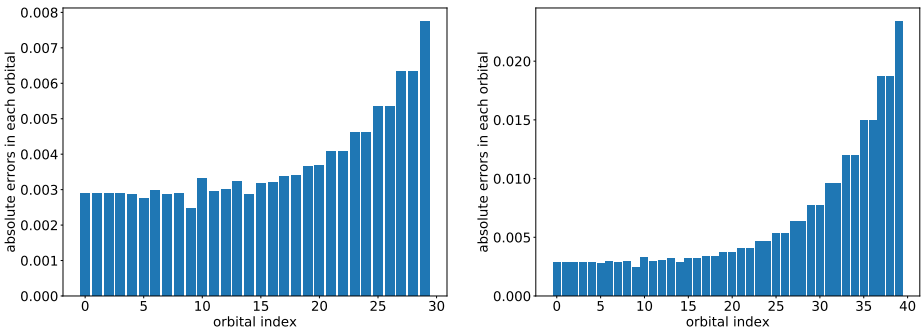


(a) $N_e = 10$ (b) $N_e = 20$

(c) $N_e = 30$ (d) $N_e = 40$

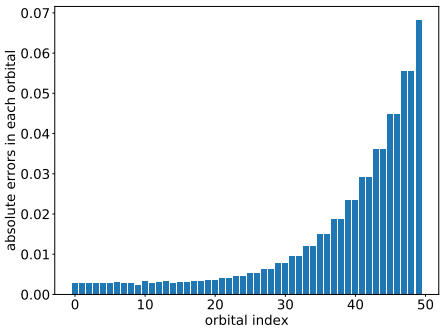
(e) $N_e = 50$ (f) $N_e = 60$

Figure 4.7.6: Plot of the error histogram of all orbitals for various N_e in the PT dynamics.

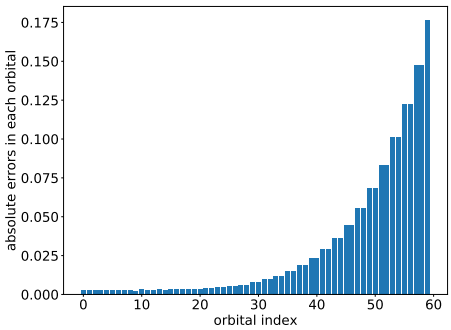


(a) $N_e = 10$

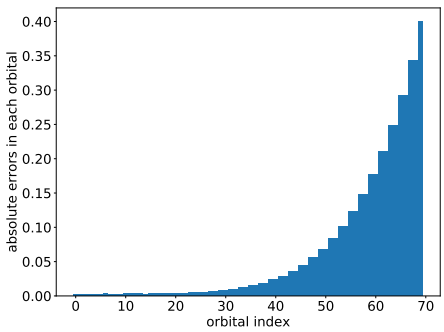
(b) $N_e = 20$



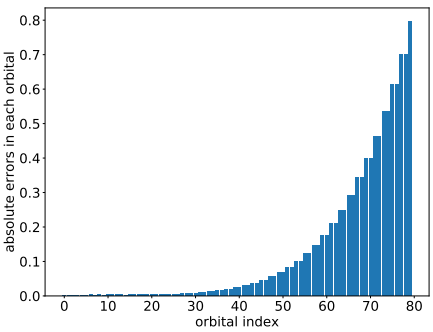
(c) $N_e = 30$



(d) $N_e = 40$



(e) $N_e = 50$



(f) $N_e = 60$

Figure 4.7.7: Plot of the error histogram of all orbitals for various N_e in the Schrödinger dynamics.

correlation contribution and depends only on the diagonal part of $\rho(t)$, and V_{X} is the Fock exchange operator depending on the entire $\rho(t)$. More specifically, V_{X} is an integral operator defined by

$$[(V_{\mathrm{X}}[\rho])\phi](x) = -\int K(x,y)\rho(x,y)\phi(y)\,dy$$

with the kernel $K(x,y)$ represents the electron-electron interaction.

Following Eq. (4.7.3), we consider the following model problem

$$H(t,\rho) = -\frac{1}{2}\Delta + V(x) + W(x,t) + U[\rho], \tag{4.7.4}$$

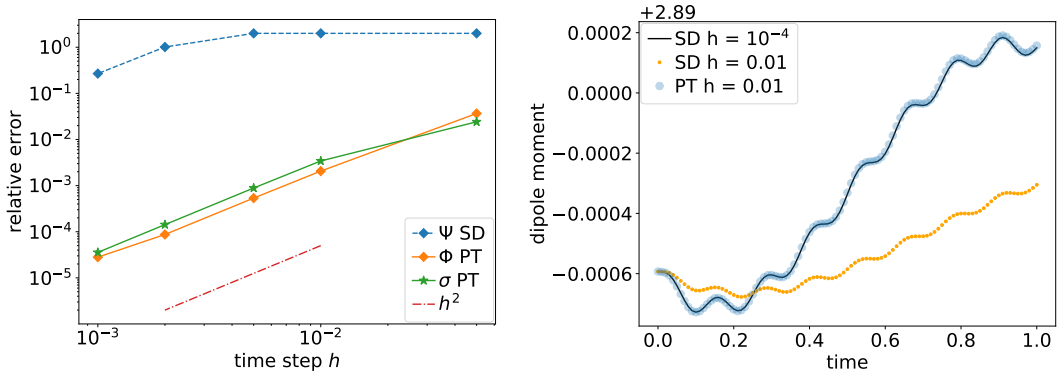
where $V = x^2$, and W is as defined in (4.7.2) and the nonlinear term $U[\rho]$ models $V_X[\rho]$ with the Yukawa kernel

$$K(x,y) = \frac{2\pi}{\kappa\epsilon_0}e^{-\kappa|x-y|}.$$

Note that as $\kappa \to 0$, the Yukawa kernel approaches to the Coulomb interaction that diverges in one dimension and hence is typically used in place of the bare Coulomb interaction for one-dimensional problems. The parameters are chosen as $\epsilon_0 = 100$ and $\kappa = 0.01$ so that the range of the electrostatic interaction is sufficiently long. Here $\mu = 148.99$ so that $N_e = 60$ and we choose $N = 80$. We simulate the system using PT-IM and SD-IM up to $T_{\mathrm{final}} = 0.5$ and compare the relative errors. As shown in Fig. 4.7.8a, the errors from PT-IM are significantly smaller. A comparison of the dipole moment is presented in Fig. 4.7.8b. We also compute the dipole moment using $h = 0.01$ and compare the results with the reference solution obtained using SD-IM with a very fine time step $h = 0.0001$. It can be seen that the result using the PT dynamics agrees well with the reference, which is not the case for the Schrödinger dynamics with the same step size.

4.8 Conclusion

In this chapter, we have introduced the PT dynamics for mixed quantum states, which generalizes the PT dynamics for pure states presented in Chapter 2. Both the PT and Schrödinger dynamics employ the low-rank structure of the density matrix, and can produce the same density matrix and all derived physical observables (such as dipole moments) as those from the von Neumann equation in the continuous time limit. The PT dynamics differ from the Schrödinger dynamics in terms of the choice of the gauge. In particular, the PT gauge yields the slowest possible dynamics for the wavefunctions. This allows us to significantly increase the time step size in the numerical simulation while maintaining accuracy.



(a) Relative Errors

(b) Dipole moment

Figure 4.7.8: Left Panel: The log-log plot of the relative errors of Φ, σ in the model nonlinear rt-TDDFT calculation with Eq. (4.7.4), computed via PT-IM and SD-IM. Right Panel: a comparison of the dipole moment.

As a concrete example, we propose the parallel transport-implicit midpoint (PT-IM) scheme, which is an implicit method suitable for treating Hamiltonians with a large spectral radius. It also preserves certain trace conditions and the orthogonality of the wavefunctions. We establish a new error bound for the PT dynamics, where all terms in the error bounds involve either the commutator of the Hamiltonian (and its derivatives) and the density matrix (or the associated spectral projector). As a comparison, the error analysis of the Schrödinger dynamics is also provided, which does not exhibit such commutator scaling. This new error bound, together with various numerical experiments, justifies the advantage of the PT dynamics for the general mixed states, where the dynamics can be nonlinear, and beyond the near adiabatic regime.

Part IV

Simulating quantum dynamics on quantum computers

This part focuses on simulating linear quantum dynamics on quantum computers, which is referred to as *Hamiltonian simulation* in the quantum computing context. As having been discussed in Part I, quantum simulation can potentially achieve an exponential speedup over classical simulation in storage and computational cost, and thus is promising to be the next generation computing approach to simulate large scale quantum many-body systems without model reduction.

The accuracy of Hamiltonian simulation is usually measured by the error of the unitary evolution operator in the operator norm, which in turn depends on a certain norm of the Hamiltonian. For unbounded operators, after suitable discretization, the norm of the Hamiltonian can be very large, which significantly increases the simulation cost. This is also an alternative interpretation of the computational challenge of the fast oscillatory solution since the fast possible component of the wave function oscillates on the time scale of the inverse norm of the Hamiltonian. However, the operator norm measures the worst-case error of the quantum simulation, while practical simulation concerns the error with respect to a given initial vector at hand. In Chapter 5, we demonstrate that under suitable assumptions of the Hamiltonian and the initial vector, if the error is measured in terms of the vector norm, the computational cost of Trotter type methods may not increase at all as the norm of the Hamiltonian increases. In this sense, our result outperforms all previous error bounds in the quantum simulation literature. Our result extends that of [81] to the time-dependent setting. We also clarify the existence and the importance of commutator scalings of Trotter and generalized Trotter methods for time-dependent Hamiltonian simulations.

As discussed in Part I, Hamiltonian simulation can be applied to solving large-scale eigenvalue problems by constructing a time-dependent Hamiltonian interpolating another simple Hamiltonian and target Hamiltonian with gap condition and performing time-dependent Hamiltonian simulation with sufficiently large physical time. Such a procedure is called adiabatic quantum computing (AQC). In Chapter 6, we study how AQC can be applied to solve large-scale linear system problems, which appear ubiquitously in scientific computing. We demonstrate that with an optimally tuned scheduling function, AQC can readily solve a quantum linear system problem (QLSP) with $\mathcal{O}(\kappa \text{ poly}(\log(\kappa/\epsilon)))$ runtime, where κ is the condition number, and ϵ is the target accuracy. This is near-optimal in both κ and ϵ and is achieved without relying on complicated amplitude amplification procedures that are difficult to implement. Our method applies to general non-Hermitian matrices, and the cost and the number of qubits can be reduced when restricted to Hermitian matrices and further to Hermitian positive definite matrices. The success of the time-optimal AQC implies that the quantum approximate optimization algorithm (QAOA) with an optimal control protocol can also achieve the same complexity in terms of the runtime. Numerical results indicate that QAOA can yield the lowest runtime compared to the time-optimal AQC, vanilla AQC, and the recently proposed randomization method [144].

Please note that in this part, we follow the quantum computing convention in notations

and terminologies. We use the terminology *time-independent Hamiltonian simulation* for simulating dynamics with constant Hamiltonian H, and use *time-dependent Hamiltonian simulation* for simulating dynamics with time-dependent $H(t)$. For the linear algebra notations, we use $|v\rangle$ to denote a normalized (under l^2-norm) vector v, which is also called a quantum state. For a normalized vector v, $\langle v|$ represents its conjugate transpose, and for a matrix A, $A^\dagger$ represents its conjugate transpose. Unless otherwise specified, $\|\cdot\|$ denotes the vector/matrix 2-norm.

Chapter 5

Time-dependent unbounded Hamiltonian simulation with vector norm scaling

5.1 Introduction

Let $H(t)$ be a Hamiltonian defined on the interval $[0, T]$, and $|\psi_0\rangle$ be the initial vector, then the time-dependent Hamiltonian simulation problem aims to find $|\psi(T)\rangle$, which solves the time-dependent Schrödinger equation

$$\mathrm{i}\partial_t |\psi(t)\rangle = H(t) |\psi(t)\rangle , \quad |\psi(0)\rangle = |\psi_0\rangle . \tag{5.1.1}$$

In this chapter, we are concerned with the simulation of a time-dependent and unbounded Hamiltonian $H(t)$, which naturally includes the simulation of a time-independent Hamiltonian $H(t) \equiv H$ as a special case. More precisely, we assume that there is a family of Hamiltonians $H^{(n)}(t)$ such that as $n \to \infty$, the norm of H (e.g. the max of the operator norm or the L^1 norm) also increases towards infinity.

For concreteness, we will consider the bilinear quantum control Hamiltonian of the following form

$$H^{(n)}(t) = f_1(t)H_1^{(n)} + f_2(t)H_2^{(n)}. \tag{5.1.2}$$

Here $H_1^{(n)}$ and $H_2^{(n)}$ are time-independent Hamiltonians, and f_1 and f_2 are two bounded, smooth scalar functions on a time interval $[0, T]$. Without loss of generality we assume that $\lim_{n\to\infty}\|H_1^{(n)}\| = \infty$, while $\lim_{n\to\infty}\|H_2^{(n)}\| < \infty$, i.e. $H_1^{(n)}$ approaches an unbounded operator, while the limit of $H_2^{(n)}$ is a bounded operator. We also assume that $\exp\left(-\mathrm{i}H_1^{(n)}\right)$ and $\exp\left(-\mathrm{i}H_2^{(n)}\right)$ can be efficiently simulated. More specifically, if n also denotes the di-

mension of the Hamiltonian, we assume that the cost of the time-independent simulations depends on at most poly-logarithmically in terms of n and the error ϵ. Such an assumption is standard for $H_2^{(n)}$ because $H_2^{(n)}$ has spectral norm asymptotically independent of n thus can be efficiently simulated, e.g. via the QSP technique [109]. However, the assumption on the effectiveness of simulating $H_1^{(n)}$ is very strong especially when $\|H_1^{(n)}\|$ grows polynomially in terms of n, and the no-fast-forwarding theorem [14, 17] requires roughly $\Omega(\|H_1^{(n)}\|)$ queries for generic quantum algorithms to simulate $\exp\left(-\mathrm{i}H_1^{(n)}\right)$. Nevertheless, for a subset of Hamiltonians with special structures, such a *time-independent* simulation can indeed be fast-forwarded and the query complexity is still poly-logarithmic of n. Typical examples include 1-sparse Hamiltonians [40, 1, 108] and thus unitarily diagonalized Hamiltonians where the diagonalization procedure can be efficiently implemented. We will show later the $H_1^{(n)}$ of interest in this chapter can also be fast-forwarded. The availability of the fast-forwarded time-independent Hamiltonian simulation allows us to measure the cost directly in terms of the number of Trotter steps.

When the context is clear, we will drop the superscript n and assume instead that $\|H_1\|$ is sufficiently large. In particular, we have $\|H_1\| \gg \|H_2\|$. The form of Eq. (5.1.2) allows us to efficiently evaluate terms of the form

$$\int_{t_1}^{t_2} H(t)dt = \left(\int_{t_1}^{t_2} f_1(t)dt\right) H_1 + \left(\int_{t_1}^{t_2} f_2(t)dt\right) H_2,$$

where the coefficients in the parentheses can be precomputed on classical computers when $f_1(t), f_2(t)$ are available.

As an example, consider the following Schrödinger equation with a time-dependent effective mass $M_{\text{eff}}(t)$ (see e.g. [49, 128, 127, 83, 56, 139]) in a domain D with proper boundary conditions as

$$H(t) = -\frac{1}{2M_{\text{eff}}(t)}\Delta + \frac{1}{2}M_{\text{eff}}(t)\omega^2(t)V(x), \quad x \in D. \tag{5.1.3}$$

Here $\omega(t)$ is a frequency parameter. Then we set $f_1(t) = 1/(2M_{\text{eff}}(t))$, $f_2(t) = M_{\text{eff}}(t)\omega^2(t)/2$. When $V(x) \equiv x^2$ the system is a quantum harmonic oscillator with time-dependent effective mass. In general we assume the potential has suitable regularity conditions and is bounded on D. After proper spatial discretization using n degrees of freedom, $H_1^{(n)}$ is the discretized negative Laplacian operator $-\Delta$ which is unbounded, and $H_2^{(n)}$ is the discretized diagonal potential $V(x)$ which is bounded. We notice that the simulation of $H_1^{(n)}$ can be fast-forwarded since it can be diagonalized under the quantum Fourier transform procedure [124]. In order to demonstrate the behavior of the Trotter formulae for unbounded operators, we require n to grow polynomially with respect to ϵ^{-1}, where ϵ is the relative 2-norm error of the solution. This is the case, for instance, when the potential $V(x)$ is of limited regularity. Throughout the

chapter we only require $V(x)$ to be a C^4 function on the domain D.[1] Again for concreteness of discussion about the computational cost, unless otherwise specified, we will assume the system is one-dimensional, $D = [0, 1]$ with the periodic boundary condition, and use the second order finite difference method with n equidistant nodes for spatial discretization.

To the extent of our knowledge, all previous results in the quantum simulation literature (for both time-independent and time-dependent Hamiltonians) measure the error of the evolution operator $\|\widetilde{U}(T) - U(T)\|$, where $U(T) = \exp_{\mathcal{T}}\left(-\mathrm{i}\int_0^T H(t)dt\right)$ is the exact evolution operator expressed in terms of a time-ordered matrix exponential, and $\widetilde{U}(T)$ is an approximate evolution operator obtained via the numerical scheme. We then directly obtain the vector norm error $\||\widetilde{\psi}(T)\rangle - |\psi(T)\rangle\| \leqslant \|\widetilde{U}(T) - U(T)\|$. However, since $\|\widetilde{U}(T) - U(T)\|$ typically depends polynomially on the operator norm $\|H_1\|$, as $\|H_1\|$ increases, if the computational cost does not increase accordingly, then all error bounds of the operator norm $\|\widetilde{U}(T) - U(T)\|$ would increase to $\mathcal{O}(1)$, with the exception of the interaction picture method for time-independent Hamiltonian simulations [108].[2] While the operator norm error provides an upper bound of the error given *any* initial vector, for a *particular* simulation instance, it is the vector norm error $\||\widetilde{\psi}(T)\rangle - |\psi(T)\rangle\|$ that matters. It turns out that for certain unbounded operators and initial vectors, the vector norm bound can be significantly improved. The key reason is that the magnitude of terms such as $\|H_1 |\psi\rangle\|, \|[H_1, H_2] |\psi\rangle\|$ can be much smaller than the corresponding operator norm estimates. In fact the importance of the vector norm estimates has long been recognized in the numerical analysis literature, and the vector norm error bounds have been established for time-independent Hamiltonian simulation using second and higher order Trotter methods of the form $H = -\Delta + V(x)$ [81, 149, 50], and for time-dependent Hamiltonian simulation using Magnus integrators of the form $H = -\Delta + V(t, x)$ [75]. Under suitable discretization and choice of the initial vector, the vector norm error $\||\widetilde{\psi}(T)\rangle - |\psi(T)\rangle\|$ obtained by the standard Trotter method remains small, even as $\|H\| \to \infty$ and the operator norm $\|\widetilde{U}(T) - U(T)\|$ becomes $\mathcal{O}(1)$.

Contribution: The *first contribution* of this chapter is to extend the vector norm estimate [81] to time-dependent unbounded Hamiltonian simulations. For concreteness we focus on the standard first and second-order Trotter methods, as well as a class of generalized Trotter methods proposed in [78], which will be introduced in Section 5.3. Our main result for a given

[1] Here the C^4 regularity is a technical assumption to bound the norm of the nested commutators, which will be detailed in Section 5.6.

[2] In the context of time-independent simulation [44], the error of Trotter methods does not scale directly with respect to the operator norm $\|H_1\|$, but with respect to the norm of the (high-order) commutators $[H_1, H_2], [H_1, [H_1, H_2]], [H_2, [H_2, H_1]]$ and so on. However, this does not change our conclusion here. In principle, the interaction picture method can also be generalized to efficiently simulate time-dependent Hamiltonians. However, its practical performance has not been well understood.

control Hamiltonian Eq. (5.1.2) is Theorem 46. It states that under suitable assumptions, the vector norm error obtained from both standard and generalized Trotter methods depends mainly on $\sup_{t\in[0,T]} \|H_1 |\psi(t)\rangle\|$, which can be significantly smaller than $\|H_1\|$.[3]

In order to simulate the Hamiltonian of the form Eq. (5.1.3), we take both the spatial and temporal discretization into account, and our complexity estimates are given in Theorem 53. Our result compared to existing results are given in Table 5.1, where the complexity for time-independent simulations are obtained by treating $M_{\text{eff}}(t), \omega(t)$ as constants. In particular, the vector norm is asymptotically independent of the spatial discretization parameter n, and complexity in terms of the error matches that of the time-independent Hamiltonian simulation obtained by [81]. Under the same second order spatial discretization, our complexity estimate for second order Trotter formulae outperforms state-of-the-art error bounds using high order Trotter and post-Trotter schemes [16, 108, 18] in terms of the desired level of accuracy, due to their dependence on the spectral norm of H_1 and thus on n.

The effectiveness of the vector norm bound depends on the initial vector $|\psi_0\rangle$. We remark that recently [136] establishes improved error estimates of low-order Trotter methods for time-independent Hamiltonian simulation, when the initial vector is constrained to be within a low energy subspace. Another recent work [143] obtains an improved complexity estimate for simulating a system with η interacting electrons using time-independent Trotter formula, by considering the operator norm constrained on this η-electron sub-manifold. Our vector norm estimate provides a complementary perspective in understanding why such improved estimates are possible. When $\sup_{t\in[0,T]} \|H_1 |\psi(t)\rangle\|$ is indeed comparable to $\|H_1\|$, the operator norm bound still serves as a good indicator of the error.

Given the improved error commutator scaling estimates for time-independent simulations [44], it is natural to ask whether the commutator scaling of the operator norm still holds for time-dependent simulations. The *second contribution* of this chapter is to reveal that for time-dependent simulations, the error of standard Trotter method *does not* exhibit commutator scalings, while the commutator scaling holds for the generalized Trotter method (Theorem 42). Therefore in the context of time-dependent simulations, the use of the generalized Trotter method could reduce the simulation cost. Our proof of the operator norm error bounds mainly follow the procedure proposed in [44], and our results generalize the first and second order time-independent results in [44] in the sense that, when the scalar functions f_1 and f_2 are constant functions, both time-dependent standard Trotter formula and time-dependent generalized Trotter formula degenerate to the same time-independent Trotter formula, and the corresponding operator norm error bound is of commutator scaling.

Yet another twist comes when we ask the question: when H_1 is unbounded, is it clear that

[3]Theorem 46 shows that the number of Trotter steps may not scale with respect to $\|H_1\|$. The mechanism of the improvement is very different from that of the interaction picture approach, where the number of the time steps is still linear in $\|H_1\|$.

the norm of the commutators $\|[H_1, H_2]\|, \|[H_1, [H_1, H_2]]\|, \|[H_2, [H_2, H_1]]\|$ must be smaller than $\|H_1\|$? It turns out that for the Hamiltonian Eq. (5.1.3), we may directly analyze that $\|[H_1, H_2]\|, \|[H_2, [H_2, H_1]]\| \in \mathcal{O}(n)$, while $\|[H_1, [H_1, H_2]]\|, \|H_1\| \in \mathcal{O}(n^2)$ (see Section 5.6). Therefore the first-order generalized Trotter method outperforms the first-order standard Trotter method, but the asymptotic efficiency of the second-order generalized and standard Trotter methods are the same (Lemma 51). Table 5.2 summarizes the performance of Trotter and generalized Trotter methods. Though both second-order schemes share the same asymptotic scaling, the generalized Trotter formula may still be a better choice in practice due to smaller preconstants, which is observed numerically. Moreover, the p-th generalized Trotter scheme depends only on the $(p-1)$-th derivatives of the control functions, while the p-th standard Trotter method on its p-th derivative. Therefore when the control functions have high frequency or limited regularity, the generalized Trotter scheme may significantly outperform the standard one. Such an advantage under first-order schemes has been demonstrated in [78] as well.

All results above are confirmed by numerical experiments for the model Eq. (5.1.3) in Section 5.7, which verifies the sharpness of our estimates.

Organization: The rest of this chapter is organized as follows. In Section 5.2 we introduce several notations and preliminary lemmas used in this chapter, provide a detailed derivation of the results in Table 5.1, and briefly discuss the main ideas of proving our new results. Then in Section 5.3 we show the schemes that will be considered in this chapter and derive their exact error representations. Operator norm error bounds and vector norm error bounds are given in Section 5.4 and Section 5.5, respectively. Section 5.6 shows how the newly obtained vector norm error bounds can be applied to obtaining better complexity estimate of Trotter type methods for solving Schrödinger equation with time-dependent mass and frequency. Numerical results are given in Section 5.7, which verifies our theoretical results.

5.2 Preliminaries

In this section we first introduce several notations and preliminary lemmas used in this chapter. Then we briefly sketch the main ideas for proving the main theorems.

Notations

We refer to a (possibly unnormalized) vector as $\vec{\psi}, \vec{u}$ or $\vec{v}$ depending on the context, and use $|\psi\rangle$ to denote the corresponding quantum state (*i.e.* normalized vector under vector 2-norm). We define two vector norms for a vector $\vec{\psi} = (\psi_0, \cdots, \psi_{n-1})$, namely the standard

	Work/Method	Scaling w. spatial discretization	Overall query complexity
Time-independent 2nd order Trotter	Childs *et al.* [44]	$\mathcal{O}(n)$	$\mathcal{O}(\epsilon^{-1})$
	Jahnke *et al.* [81]	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon^{-0.5})$
Time-independent higher order methods	p-th order Trotter [44]	$\mathcal{O}(n^{2-2/p})$	$\mathcal{O}(\epsilon^{-1})$
	Truncated Taylor series [19, 95]	$\widetilde{\mathcal{O}}(n^2)$	$\widetilde{\mathcal{O}}(\epsilon^{-1})$
	Quantum signal processing [109]	$\mathcal{O}(n^2)$	$\mathcal{O}(\epsilon^{-1})$
	Interaction picture [108]	$\mathcal{O}(\log(n))$	$\mathcal{O}(\text{polylog}(1/\epsilon))$
Time-dependent 2nd order Trotter	Huyghebaert *et al.* [78]	$\mathcal{O}(n)$	$\mathcal{O}(\epsilon^{-1})$
	Wiebe *et al.* [155]	$\mathcal{O}(n^3)$	$\mathcal{O}(\epsilon^{-2})$
	Wecker *et al.* [153]	$\mathcal{O}(n)$	$\mathcal{O}(\epsilon^{-1})$
	This work	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon^{-0.5})$
Time-dependent higher order methods	p-th order Trotter [155]	$\mathcal{O}(n^{2+2/p})$	$\mathcal{O}(\epsilon^{-1-2/p})$
	Truncated Dyson series [16, 108]	$\widetilde{\mathcal{O}}(n^2)$	$\widetilde{\mathcal{O}}(\epsilon^{-1})$
	Rescaled Dyson series [18]	$\widetilde{\mathcal{O}}(n^2)$	$\widetilde{\mathcal{O}}(\epsilon^{-1})$

Table 5.1: Comparison of complexity estimates for simulating the model Eq. (5.1.3) in one-dimension using second order Trotter method or higher order Trotter or post-Trotter method, with C^4 potential function $V(x)$, and time-independent mass and frequency (top 2) or time-dependent mass and frequency (bottom 2). For all the methods, we use a second order finite difference discretization with n degrees of freedom. The third column summarizes the scaling of the cost with respect to n in order to reach constant target accuracy, and the fourth column summarizes the overall query complexity in order to achieve a desired level of relative 2-norm error ϵ. Since we assume the efficiency of time-independent simulation for both H_1 and H_2, the query complexity is measured by the number of required Trotter steps for Trotter-type methods, or the query complexity under standard query model for post-Trotter methods. The simulation time T is $\mathcal{O}(1)$. 'This work' refers to the vector norm error bound using the second order standard or generalized Trotter formula. Throughout the chapter $f = \widetilde{\mathcal{O}}(g)$ if $f = \mathcal{O}(g\,\text{polylog}(g))$. See Section 5.2 for details of the derivation of the scalings.

	Method & Error type	Scaling w. spatial discretization	Overall number of Trotter steps
First-order Trotter	standard, operator norm	$\mathcal{O}(n^2)$	$\mathcal{O}(\epsilon^{-2})$
	generalized, operator norm	$\mathcal{O}(n)$	$\mathcal{O}(\epsilon^{-1.5})$
	standard, vector norm	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon^{-1})$
	generalized, vector norm	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon^{-1})$
Second-order Trotter	standard, operator norm	$\mathcal{O}(n)$	$\mathcal{O}(\epsilon^{-1})$
	generalized, operator norm	$\mathcal{O}(n)$	$\mathcal{O}(\epsilon^{-1})$
	standard, vector norm	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon^{-0.5})$
	generalized, vector norm	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon^{-0.5})$

Table 5.2: Summary of results for first and second order Trotter formulae applied to simulating the model Eq. (5.1.3) in one-dimension with time-dependent effective mass and frequency. For all the methods, we use a second order finite difference discretization with n degrees of freedom. The third column summarizes the scaling of the cost with respect to n in order to reach constant target accuracy (Lemma 51), and the fourth column summarizes the overall number of required Trotter steps to achieve a desired level of relative 2-norm error ϵ estimated from error bounds in different norms (Theorem 53). The simulation time T is $\mathcal{O}(1)$.

2-norm

$$\|\vec{\psi}\| = \sqrt{\sum_{k=0}^{n-1} |\psi_k|^2},$$

and the rescaled 2-norm

$$\|\vec{\psi}\|_\star = \frac{1}{\sqrt{n}}\|\vec{\psi}\|.$$

The rescaled 2-norm is directly motivated by the discretization of the continuous L^2 norm [102, 150]. Specifically, for a real-space function $u(x)$ discretized in the real space using n equidistant nodes, we apply the trapezoidal rule and obtain

$$\int_0^1 |u(x)|^2 dx \approx \sum_{k=0}^{n-1} \left(|u(k/n)|^2 \frac{1}{n} \right) = \frac{1}{n}\|(u(k/n))_{k=0}^{n-1}\|^2 = \|(u(k/n))_{k=0}^{n-1}\|_\star^2. \tag{5.2.1}$$

Since the $\|\cdot\|_\star$ simply rescales the standard vector 2-norm, the estimates that we will derive for 2-norm also hold for this rescaled 2-norm. Furthermore, the corresponding matrix norm induced by the rescaled 2-norm is still the standard matrix 2-norm without any rescaling factor, as

$$\|A\| = \sup_{\|\vec{u}\|\neq 0} \|A\vec{u}\|/\|\vec{u}\| = \sup_{\|\vec{u}\|_\star\neq 0} \|A\vec{u}\|_\star/\|\vec{u}\|_\star.$$

We remark that it is equivalent to use either 2-norm or rescaled 2-norm if we wish to bound the relative error of the numerical solutions.

For two matrices A, B, define the adjoint mapping ad_A as

$$\mathrm{ad}_A(B) = [A, B] = AB - BA, \tag{5.2.2}$$

and then the conjugation of matrix exponentials of the form $\exp(A)B\exp(-A)$ can be simply expressed as

$$\exp(\mathrm{ad}_A)B = \exp(A)B\exp(-A). \tag{5.2.3}$$

The following conjugation of matrix exponentials will be commonly used for a scale-valued function f and matrices A, B

$$\exp\left(\mathrm{ad}_{\mathrm{i}\int_{t_1}^{t_2} f(s)dsA}\right) B = \exp\left(\mathrm{i}\int_{t_1}^{t_2} f(s)dsA\right) B \exp\left(-\mathrm{i}\int_{t_1}^{t_2} f(s)dsA\right). \tag{5.2.4}$$

For a scalar-valued continuous function $f(t)$ defined on time domain $t \in [0, T]$, we use $\|f\|_\infty$ to denote the supremum of the function in this time interval, *i.e.*

$$\|f\|_\infty = \sup_{t\in[0,T]} |f(t)|.$$

Elementary lemmas

We review two elementary lemmas to be used in the proof of the chapter. Proofs of the results can be found in, *e.g.* [69, 96].

Lemma 33 (Taylor's theorem)**.** *For any k-th order continuously differentiable function f (scale-valued or matrix-valued) defined on an interval $[a, t]$, we have*

$$f(t) = \sum_{j=0}^{k-1} \frac{f^{(j)}(a)}{j!}(t-a)^j + \int_a^t \frac{f^{(k)}(s)}{(k-1)!}(t-s)^{k-1}ds. \tag{5.2.5}$$

Lemma 34 (Variation of parameters formula)**.** *Assume $U(t,s)$ solves the differential equation*

$$\partial_t U(t,s) = H(t)U(t,s), \quad U(s,s) = I. \tag{5.2.6}$$

Then

1. *For any matrix-valued continuous function $R(t)$, the solution of the differential equation*

$$\partial_t \widetilde{U}(t,0) = H(t)\widetilde{U}(t,0) + R(t), \quad \widetilde{U}(0,0) = I \tag{5.2.7}$$

can be represented as

$$\widetilde{U}(t,0) = U(t,0) + \int_0^t U(t,s)R(s)ds. \tag{5.2.8}$$

2. *For any vector-valued continuous function $\vec{r}(t)$, the solution of the differential equation*

$$\partial_t \vec{\widetilde{u}}(t) = H(t)\vec{\widetilde{u}}(t) + \vec{r}(t), \quad \vec{\widetilde{u}}(0) = \vec{u}_0 \tag{5.2.9}$$

can be represented as

$$\vec{\widetilde{u}}(t) = U(t,0)\vec{u}_0 + \int_0^t U(t,s)\vec{r}(s)ds. \tag{5.2.10}$$

Derivation of results in Table 5.1

In this section we show explicitly how to derive the results in Table 5.1. Throughout this section we are considering the setup in Section 5.6 with $T = \mathcal{O}(1)$.

To obtain Table 5.1, we first restate all the complexity estimates for different methods proved in existing literature and show how they depend on ϵ as well as the scale of the Hamiltonians H_1 and H_2. The dependence on H_1 naturally gives rise to the dependence on n, by noticing that

$$\|H_1\| = \mathcal{O}(n^2), \quad \|H_2\| = \mathcal{O}(1), \quad \|[H_1, H_2]\| = \mathcal{O}(n),$$

$$\|[H_1, [H_1, H_2]]\| = \mathcal{O}(n^2), \quad \|[H_2, [H_2, H_1]]\| = \mathcal{O}(n),$$

as is discussed in Lemma 48. Then, under second order finite difference spatial discretization, Lemma 50 and Eq. (5.6.31) tell that n should be chosen as large as $\mathcal{O}(\epsilon^{-1/2})$. Plugging this back into the complexity estimates leads to the overall scaling in terms of ϵ, as shown in the last column of Table 5.1.

Time-independent schemes

Time-independent second order Trotter formula [44, Proposition 16] gives an operator norm error bound for time-independent second order Trotter formula that the one-step local Trotter error is bounded by

$$\frac{h^3}{12}\|[H_2,[H_2,H_1]]\|+\frac{h^3}{24}\|[H_1,[H_1,H_2]]\|,$$

thus the global Trotter error is bounded by

$$\left(\frac{1}{12}\|[H_2,[H_2,H_1]]\|+\frac{1}{24}\|[H_1,[H_1,H_2]]\|\right)\frac{T^3}{L^2}=\mathcal{O}\left(\frac{n^2}{L^2}\right).$$

To bound this by ϵ, it suffices to choose

$$L=\mathcal{O}\left(\frac{n}{\epsilon^{1/2}}\right)=\mathcal{O}\left(\frac{1}{\epsilon}\right).$$

[81, Theorem 3.2] provides a vector norm error bound for time-independent second order Trotter formula that the global Trotter error is bounded by

$$\mathcal{O}\left(h^2\left(\|H_1\vec{v}\|_\star+\|D_1\vec{v}\|_\star+\|\vec{v}\|_\star\right)\right)=\mathcal{O}\left(\frac{1}{L^2}\left(\|H_1\vec{v}\|_\star+\|\vec{v}\|_\star\right)\right).$$

We remark that [81] does not track explicitly the dependence on T. Noticing that $\|H_1\vec{v}\|_\star$ and $\|\vec{v}\|_\star$ are independent of ϵ and n (shown in the proof of Lemma 51), the number of required Trotter steps scales

$$L=\mathcal{O}\left(\frac{1}{\epsilon^{1/2}}\right).$$

Time-independent high order Trotter formula [44, Corollary 12] shows that for a p-th order time-independent Trotter formula, the number of required Trotter steps to obtain an ϵ-approximation of the exact evolution operator is

$$L=\mathcal{O}\left(\frac{\widetilde{\alpha}_{\text{comm}}^{1/p}T^{1+1/p}}{\epsilon^{1/p}}\right),$$

where

$$\widetilde{\alpha}_{\text{comm}}=\sum_{\gamma_1,\cdots,\gamma_{p+1}=1}^{2}\|[H_{\gamma_{p+1}},\cdots[H_{\gamma_2},H_{\gamma_1}]]\|.$$

Straightforward bounds for these p-th nested commutators are that

$$\|[H_{\gamma_{p+1}}, \cdots [H_{\gamma_2}, H_{\gamma_1}]]\| = \mathcal{O}(\|H_1\|^{p-2}\|[H_1, [H_1, H_2]]\|) = \mathcal{O}(n^{2p-2}),$$

which results in

$$L = \mathcal{O}\left(\frac{n^{2-2/p}}{\epsilon^{1/p}}\right) = \mathcal{O}\left(\frac{1}{\epsilon}\right).$$

Notice that the scaling of ϵ is not improved by higher order Trotter formula. This is because such an estimate is made under the assumption that the potential $V(x)$ is a C^4 function, therefore we only have better scaling for nested commutator up to second order. If the potential $V(x)$ has higher regularity, we expect better bounds to exist for general nested commutators, just like the case of $[H_1, H_2]$ and $[H_1, [H_1, H_2]]$. In particular, although we do not present complete proof in this chapter, a continuous analog as well as discretization under Fourier basis suggests that the norm of p-th order nested commutator $\|[H_1, \cdots, [H_1, H_2]]\|$ is bounded by $\mathcal{O}(\|D_1^p\|)$ if $V(x)$ is $(2p)$-th order continuously differentiable. In that case the complexity can be improved to $L = \mathcal{O}(n/\epsilon^{1/p})$, although there is still a linear dependence on n.

Truncated Taylor series [19, Theorem 1] shows that to obtain an ϵ-approximation of the exact evolution operator using truncated Taylor series, the query complexity is

$$\mathcal{O}\left(d^2\|H\|_{\max}\frac{\log(d\|H\|_{\max}/\epsilon)}{\log\log(d\|H\|_{\max}/\epsilon)}\right).$$

Here d is the sparsity of the Hamiltonian, $\|H\|_{\max}$ denotes the largest matrix element of H in absolute value. Notice that $\|H_1\|_{\max} = \mathcal{O}(n^2)$ since every non-zero entry of H_1 is either n^2 or $(-2n^2)$, and $\|H_2\|_{\max} = \mathcal{O}(1)$, we have $\|H\|_{\max} = \mathcal{O}(n^2)$. Therefore the query complexity becomes

$$\mathcal{O}\left(n^2\frac{\log(n^2/\epsilon)}{\log\log(n^2/\epsilon)}\right) = \widetilde{\mathcal{O}}\left(n^2\right) = \widetilde{\mathcal{O}}\left(\frac{1}{\epsilon}\right).$$

We remark that the work [95] studies further the complexity of simulating time-independent many-body Hamiltonian and discusses carefully the errors from both time and space discretization. In this work, the authors use truncated Taylor series as well for time discretization, and use high order finite difference formula for spatial discretization. However, they only assume that the potential $V(x)$ is first-order continuous differentiable thus the high order finite difference formula does not offer improved scaling of n than $\mathcal{O}(1/\epsilon)$ [95, Theorem 4], which results in a total complexity $\widetilde{\mathcal{O}}(1/\epsilon^2)$ [95, Theorem 3 & 4]. The scaling can be improved if $V(x)$ becomes smoother.

Quantum signal processing [109, Theorem 3] proposes a quantum signal processing approach for time-independent Hamiltonian simulation with optimal query complexity in all parameters, which is

$$\mathcal{O}\left(d\|H\|_{\max}+\frac{\log(1/\epsilon)}{\log\log(1/\epsilon)}\right).$$

Here d is the sparsity of the Hamiltonian, $\|H\|_{\max}$ denotes the largest matrix element of H in absolute value. Notice that $\|H_1\|_{\max}=\mathcal{O}(n^2)$ since every non-zero entry of H_1 is either n^2 or $(-2n^2)$, and $\|H_2\|_{\max}=\mathcal{O}(1)$, we have $\|H\|_{\max}=\mathcal{O}(n^2)$. Therefore the query complexity becomes

$$\mathcal{O}\left(n^2+\frac{\log(n^2/\epsilon)}{\log\log(n^2/\epsilon)}\right)=\mathcal{O}\left(n^2+\log(n^2/\epsilon)\right)=\mathcal{O}\left(\frac{1}{\epsilon}+\log(1/\epsilon)\right)=\mathcal{O}\left(\frac{1}{\epsilon}\right).$$

Interaction picture [108, Theorem 7] shows that by applying truncated Dyson series to simulate time-independent Hamiltonian H_1+H_2 in the interaction picture rather than the Schrödinger picture, it requires

$$\mathcal{O}\left(\|H_2\|\frac{\log(\|H_2\|/\epsilon)}{\log\log(\|H_2\|/\epsilon)}\right)$$

queries to H_2 and

$$\mathcal{O}\left(\|H_2\|\frac{\log(\|H_2\|/\epsilon)}{\log\log(\|H_2\|/\epsilon)}\log\left(\frac{\|H_1\|+\|H_2\|}{\epsilon}\right)\right)$$

queries to the unitary time evolution $e^{-\mathrm{i}sH_1}$. Therefore the query complexity is logarithmic in n and thus the scaling in terms of ϵ is still poly-logarithmic. Note that the number of time steps is included in the oracle HAM-T and scales as $\mathcal{O}(\|H_1\|)$ [108, Lemma 6].

Time-dependent schemes

Time-dependent second order Trotter formulae [78, Eq. (A12-A14)] show that for generalized second-order Trotter formula applied to the model Eq. (5.1.3) with time-independent mass and time-dependent frequency (in particular, $f_2(t)H_2$ and $f_2(s)H_2$ commute for any t and s), the one-step local Trotter error scales as

$$\mathcal{O}\left(h^3\left(\|[H_1,H_2]\|+\|[H_1,[H_1,H_2]]\|+\|[H_2,[H_2,H_1]]\|\right)\right),$$

thus the global error scales

$$\mathcal{O}\left(h^2\left(\|[H_1,H_2]\|+\|[H_1,[H_1,H_2]]\|+\|[H_2,[H_2,H_1]]\|\right)\right)=\mathcal{O}\left(\frac{n^2}{L^2}\right).$$

To bound this by ϵ, it suffices to choose

$$L = \mathcal{O}\left(\frac{n}{\epsilon^{1/2}}\right) = \mathcal{O}\left(\frac{1}{\epsilon}\right).$$

The second order complexity estimate from [155] is a special case of their general high order result. We will show the general case later.

[153, Appendix A] proves an improved operator norm error bound for the second order standard Trotter formula. The one-step local Trotter error is bounded by

$$\left(\frac{1}{24}\sup\|H''(s)\| + \sup\|[H_1(s),[H_1(s),H_2(s)]\| \right.$$
$$\left. + \frac{1}{12}\sup\|[H'(s),H(s)]\| + \|[H_2(s),[H_2(s),H_1(s)]\|\right)h^3,$$

thus the global error scales

$$\mathcal{O}\left(h^2\left(\|H_1\| + \|[H_1,H_2]\| + \|[H_1,[H_1,H_2]]\| + \|[H_2,[H_2,H_1]]\|\right)\right) = \mathcal{O}\left(\frac{n^2}{L^2}\right).$$

To bound this by ϵ, it suffices to choose

$$L = \mathcal{O}\left(\frac{n}{\epsilon^{1/2}}\right) = \mathcal{O}\left(\frac{1}{\epsilon}\right).$$

Time-dependent high order Trotter formula [155, Theorem 1] proves that, to simulate a system with Hamiltonian $H(t) = \sum_{j=1}^{m} H_j(t)$ within operator spectral norm error ϵ using a $2k$-th order standard Trotter formula, the total number of exponentials is

$$2m5^{k-1}\left\lceil 5k\Lambda T\left(\frac{5}{3}\right)^k\left(\frac{\Lambda T}{\epsilon}\right)^{1/(2k)}\right\rceil$$

where

$$\Lambda = \sup_{p=0,1,\cdots,2k}\left(\sup_t\left(\sum_{j=1}^{m}\|\partial_t^p H_j(t)\|\right)^{1/(p+1)}\right).$$

We first notice that the total number of exponentials only differ from the total number of Trotter steps by a factor of $2m5^{k-1}$. After absorbing all the terms independent of n and ϵ into the big-$\mathcal{O}$ notation, in the case of the Schrödinger equation with a time-dependent effective mass, the total number of Trotter steps becomes

$$\mathcal{O}\left(\Lambda\left(\frac{\Lambda}{\epsilon}\right)^{1/2k}\right).$$

It remains to estimate the scaling of Λ. By noticing $\partial_t^p H_j(t) = f_j^{(p)}(t) H_j$, we obtain that $\left(\sum_{j=1}^m \|\partial_t^p H_j(t)\|\right)$ is dominated by $H_1 = \mathcal{O}(n^2)$, and

$$\Lambda = \mathcal{O}\left(\sup_{p=0,1,\cdots,2k} (n^2)^{1/(p+1)}\right) = \mathcal{O}\left(n^2\right).$$

Therefore the total number of Trotter steps becomes

$$\mathcal{O}\left(\frac{n^{2+1/k}}{\epsilon^{1/(2k)}}\right) = \mathcal{O}\left(\frac{1}{\epsilon^{1+1/k}}\right).$$

Truncated Dyson series [108, Theorem 9] shows that to obtain an ϵ approximation of the exact evolution operator with success probability at least $(1-\epsilon)$ using truncated Dyson series method, the query complexity is

$$\mathcal{O}\left(d\|H\|_{\max,\infty} T \frac{\log(d\|H\|_{\max,\infty} T/\epsilon)}{\log\log(d\|H\|_{\max,\infty} T/\epsilon)}\right).$$

Here d is the sparsity of the Hamiltonian, and $\|H\|_{\max,\infty} = \sup_{t\in[0,T]} \|H(t)\|_{\max}$, where $\|A\|_{\max}$ denotes the largest matrix element of A in absolute value. In the case of the model Eq. (5.1.3), noticing that $\|H_1\|_{\max} = \mathcal{O}(n^2)$ because every non-zero entry of H_1 is either n^2 or $(-2n^2)$, we have $\|H(t)\|_{\max,\infty} = \mathcal{O}(\|H_1\|_{\max}) = \mathcal{O}(n^2)$, then the query complexity becomes

$$\mathcal{O}\left(n^2 \frac{\log(n^2/\epsilon)}{\log\log(n^2/\epsilon)}\right) = \widetilde{\mathcal{O}}\left(n^2\right) = \widetilde{\mathcal{O}}\left(\frac{1}{\epsilon}\right).$$

Rescaled Dyson series [18, Theorem 10] shows that to obtain an ϵ approximation of the exact evolution operator using rescaled Dyson series method, the query complexity is

$$\mathcal{O}\left(d\|H\|_{\max,1} \frac{\log(d\|H\|_{\max,1}/\epsilon)}{\log\log(d\|H\|_{\max,1}/\epsilon)}\right).$$

Here d is the sparsity of the Hamiltonian, $\|H\|_{\max,1} = \int_0^T \|H(t)\|_{\max} dt$ where $\|A\|_{\max}$ denotes the largest matrix element of A in absolute value. In the case of the model Eq. (5.1.3), noticing that $\|H_1\|_{\max} = \mathcal{O}(n^2)$ because every non-zero entry of H_1 is either n^2 or $(-2n^2)$, we have $\|H\|_{\max,1} = \mathcal{O}(n^2)$. Therefore the query complexity becomes

$$\mathcal{O}\left(n^2 \frac{\log(n^2/\epsilon)}{\log\log(n^2/\epsilon)}\right) = \widetilde{\mathcal{O}}\left(n^2\right) = \widetilde{\mathcal{O}}\left(\frac{1}{\epsilon}\right).$$

We mention that in [18] another method called continuous qDRIFT is also proposed to successfully achieve L^1 scaling of the Hamiltonian. However, continuous qDRIFT is a first order method, and its complexity dependence on $\|H\|_{\max,1}$ is quadratic, which is worse than that of rescaled Dyson series. Hence we only include the rescaled Dyson series method in our table for comparison.

Main ideas

Here we discuss the main ideas for our operator and vector norm error bounds, and they are applicable to both standard and generalized Trotter formulae to be introduced in Section 5.3. For simplicity, we only discuss first order formulae here, and the ideas for second order formulae are similar. First, in Section 5.3, we derive the error representations between the exact evolution operator $U(h,0)$ and the Trotterized evolution operator $U_s(h,0)$ or $U_g(h,0)$, by establishing the differential equations that these unitary operators satisfy and by using variation of parameters. Such error representations are exact. Furthermore, although full error representations can be technically complicated, they are simply linear combinations of integrals with integrand of the form

$$\left(\prod_{j=1}^{J} g_j\right)\left[\prod_{k=1}^{K}\exp\left(\mathrm{i}h\xi_k H_{l_k}\right)\right] A \left[\prod_{k'=1}^{K'}\exp\left(\mathrm{i}h\xi'_{k'} H_{l'_{k'}}\right)\right]. \tag{5.2.11}$$

Here functions g_j's can be f_1, f_2, or their derivatives; H_{l_k} and $H_{l'_{k'}}$ are either H_1 or H_2; and ξ_k and $\xi'_{k'}$ are some bounded real numbers. The matrix A is in the set $\{H_1, H_2, [H_1, H_2]\}$ for standard Trotter formula and can only be $[H_1, H_2]$ for generalized Trotter formula. Therefore it suffices to focus on each term in the form of Eq. (5.2.11) to obtain error bounds.

The operator norm error bounds directly follow the error representations. Under the assumption that H_1 and H_2 are bounded operators, we can simply bound all the unitaries by 1 and the (local) operator norm error bounds become αh^2 where α can be expressed in terms of $\|H_1\|, \|H_2\|$ and $\|[H_1, H_2]\|$ for the standard Trotter formula, and of $\|[H_1, H_2]\|$ for the generalized Trotter formula, respectively. Notice that the local errors for both formulae are $\mathcal{O}(h^2)$ for first order schemes, which agrees with the order condition. Furthermore, the preconstant for the generalized Trotter formula only consists of commutators, while the preconstant for standard Trotter formula still includes norms of H_1 and H_2 themselves.

Next we focus on the situation when $\|H_1\|$ is very large, and we still would like to obtain a well approximated quantum state of the exact wavefunction. In this case, the operator norm error bounds do not offer useful performance guarantees. To obtain a vector norm error bound, the starting point of our approach is still the exact error representation. Notice that the error between the exact state $|\psi\rangle$ and the approximate state $|\widetilde{\psi}\rangle$ obtained by Trotter

formulae can be expressed as $\||\widetilde{\psi}(h)\rangle - |\psi(h)\rangle\| \leqslant \|\left(\widetilde{U}(h,0) - U(h,0)\right)|\psi(0)\rangle\|$. Therefore the vector norm error bounds should be a linear combination of the terms of the form

$$\left\| \left[\prod_{k=1}^{K} \exp\left(\mathrm{i}h\xi_k H_{l_k}\right)\right] A \left[\prod_{k'=1}^{K'} \exp\left(\mathrm{i}h\xi'_{k'} H_{l'_{k'}}\right)\right] \vec{v} \right\| = \left\| A \left[\prod_{k'=1}^{K'} \exp\left(\mathrm{i}h\xi'_{k'} H_{l'_{k'}}\right)\right] \vec{v} \right\|. \quad (5.2.12)$$

This can be obtained by applying the operator Eq. (5.2.11) to some vector $\vec{v}$, and $\vec{v}$ is related to the initial condition as well as the exact Schrödinger wavefunctions.

To further bound Eq. (5.2.12), the key observation is as follows. When A is H_1 (or commutators involving H_1), although $\|A\|$ can be very large, it is possible for $\|A\vec{v}\|$ to be small for certain vectors $\vec{v}$. As an example, let us consider the continuous case in one dimension and we take $H_1 = -\Delta$ and H_2 being a bounded, smooth potential function $V(x)$. The direct computation shows that

$$H_1\psi = -\partial_x^2\psi,$$

$$[H_1, H_2]\psi = [-\Delta, V]\psi = -(\partial_x^2 V)\psi - 2(\partial_x V)\partial_x\psi.$$

Both of the terms on the right hand side depend on the spatial derivatives of the wavefunctions, of which the norm can be small if ψ is a smooth function. Then according to Eq. (5.2.12), we only need to somehow exchange the order between A and the exponentials without introducing much overhead (Lemma 44). Combining all previous arguments, we can obtain the desired vector norm error bounds.

5.3 Trotter type algorithms and error representations

In this section, we consider two different types of Trotter algorithms – the standard and generalized Trotter formulae – in simulating time dependent Hamiltonian Eq. (5.1.2), and derive their error representations explicitly. Here for simplicity we restrict ourselves to the first-order and second-order cases. We point out that such schemes and results regarding the (non-)existence of commutator scaling can be generalized to their higher order counterparts.

The standard and generalized Trotter formulae

Both the standard and the generalized Trotter formulae (proposed in [78]) belong to the class of splitting methods [68].

The first-order standard Trotter algorithm is

$$U_{s,1}(t+h,t) = \exp\left(-\mathrm{i}f_2(t+h)H_2 h\right)\exp\left(-\mathrm{i}f_1(t+h)H_1 h\right). \quad (5.3.1)$$

The first-order generalized Trotter formula is

$$U_{g,1}(t+h,t) = \exp\left(-\mathrm{i}\int_t^{t+h} f_2(s)ds H_2\right)\exp\left(-\mathrm{i}\int_t^{t+h} f_1(s)ds H_1\right). \tag{5.3.2}$$

The second-order standard Trotter formula is

$$U_{s,2}(t+h,t) = \exp\left(-\frac{\mathrm{i}h}{2}f_1(t+h/2)H_1\right)\exp\left(-\mathrm{i}hf_2(t+h/2)H_2\right)\exp\left(-\frac{\mathrm{i}h}{2}f_1(t+h/2)H_1\right). \tag{5.3.3}$$

The second-order generalized Trotter formula is

$$\begin{aligned} U_{g,2}(t+h,t) = &\exp\left(-\mathrm{i}\int_{t+h/2}^{t+h} f_1(s)ds H_1\right)\exp\left(-\mathrm{i}\int_t^{t+h} f_2(s)ds H_2\right) \\ &\times \exp\left(-\mathrm{i}\int_t^{t+h/2} f_1(s)ds H_1\right). \end{aligned} \tag{5.3.4}$$

It is clear that the difference between the standard and the generalized Trotter formulae lies in the temporal treatment of f_1 and f_2, and the standard Trotter formula can be viewed as applying certain quadrature rules in representing the integrals of f_1 and f_2. From now on we assume that $\int_a^b f(s)ds$ can be accurately computed with negligible extra cost for any scalar-valued smooth function $f(s)$. Furthermore, we remark that although in our definitions of the schemes we perform evolution governed by H_1 at first and then by H_2, the order of H_1 and H_2 only affects the absolute preconstants in the error bounds and will not lead to any difference in the asymptotic scalings.

Error representations

For the time-independent Trotter formula, the work of [44, 50] prove a commutator type of error of any order by writing down an explicit error representation via variation of parameters formula. Here we follow the procedure in [44] to write down the corresponding error representations for standard and generalized time-dependent Trotter formulae, which turn out to be the starting point for proving both the operator norm and the vector norm error bounds. Although we only present the error representation on the interval $[0,h]$, this is just for notation simplicity and with minor modifications the results naturally hold on $[t,t+h]$ for any t. The proofs are given in the next subsection.

Lemma 35 (Error representation of the first-order standard Trotter formula)**.**

$$U_{s,1}(h,0) - U(h,0) = \int_0^h U(h,s)\exp\left(-\mathrm{i}sf_2(s)H_2\right)E_{s,1}(s)\exp\left(-\mathrm{i}sf_1(s)H_1\right)ds \tag{5.3.5}$$

where

$$E_{s,1}(h) = \int_0^h f_1(h) f_2(s) \left(\exp\left(\mathrm{ad}_{\mathrm{i}sf_2(s)H_2}\right)([H_1, H_2])\right) ds - \mathrm{i}hf_1'(h)H_1 - \mathrm{i}hf_2'(h)H_2 + \int_0^h sf_1(h)f_2'(s) \left(\exp\left(\mathrm{ad}_{\mathrm{i}sf_2(s)H_2}\right)([H_1, H_2])\right) ds. \tag{5.3.6}$$

Lemma 36 (Error representation of the first-order generalized Trotter formula)**.**

$$U_{g,1}(h,0) - U(h,0) = \int_0^h U(h,s) \exp\left(-\mathrm{i}\int_0^s f_2(s')ds'H_2\right) \times E_{g,1}(s) \exp\left(-\mathrm{i}\int_0^s f_1(s')ds'H_1\right) ds \tag{5.3.7}$$

where

$$E_{g,1}(h) = \int_0^h f_1(h) f_2(s) \left(\exp\left(\mathrm{ad}_{\mathrm{i}\int_0^h f_2(s')ds'H_2}\right)([H_1, H_2])\right) ds. \tag{5.3.8}$$

Lemma 37 (Error representation of the second-order standard Trotter formula)**.**

$$U_{s,2}(h,0) - U(h,0) = \int_0^h U(h,s) \exp\left(-\frac{\mathrm{i}s}{2}f_1(s/2)H_1\right) E_{s,2}(s) \exp\left(-\mathrm{i}sf_2(s/2)H_2\right) \exp\left(-\frac{\mathrm{i}s}{2}f_1(s/2)H_1\right) ds \tag{5.3.9}$$

where $E_{s,2}$ is defined in Eq. (5.3.21)*.*

Lemma 38 (Error representation of the second-order generalized Trotter formula)**.**

$$U_{g,2}(h,0) - U(h,0) = \int_0^h U(h,s) \exp\left(-\mathrm{i}\int_{s/2}^s f_1(s')ds'H_1\right) E_{g,2}(s) \times \exp\left(-\mathrm{i}\int_0^s f_2(s')ds'H_2\right) \exp\left(-\mathrm{i}\int_0^{s/2} f_1(s')ds'H_1\right) ds \tag{5.3.10}$$

where $E_{g,2}$ is defined in Eq. (5.3.22)*.*

The expressions of these exact error representations are somewhat complicated, but the structures for all the representations are the same. As introduced in Section 5.2, the error

representations for both standard and generalized Trotter formulae of first and second-order are linear combinations of integrals with integrands in the form of Eq. (5.2.11), which can be expressed as the multiplication of matrix exponentials, Hamiltonians, and commutators of Hamiltonians.

Before we proceed, we remark that in the error representations of standard Trotter formulae of first and second-order, besides the $\mathcal{O}(h^{p+1})$ terms, we also include higher order terms $\mathcal{O}(h^{p+2})$ and/or $\mathcal{O}(h^{p+3})$. This is because we aim at writing down the exact error terms, and the Taylor expansion of terms like $\exp(-\mathrm{i}hf_1(h)H_1)$ will naturally involve higher order term even though we only expand it up to the desired lower order. For example, if we look at the first-order derivative of $\exp(-\mathrm{i}hf_1(h)H_1)$, then

$$\frac{d}{dh}(\exp(-\mathrm{i}hf_1(h)H_1)) = -\mathrm{i}f_1(h)H_1\exp(-\mathrm{i}hf_1(h)H_1) - \mathrm{i}hf_1'(h)H_1\exp(-\mathrm{i}hf_1(h)H_1).$$

We can observe that the first term is $\mathcal{O}(1)$ and the second term is $\mathcal{O}(h)$, which are on different scales. Therefore, the same order term in the Taylor expansion of the unitaries does not necessarily have the same scaling in terms of h.

Remark 39 (Exact error representation)**.** *It is possible to derive a simpler error bound by considering only the lowest order term and discarding all the higher order terms in h. However, such an error bound would not reveal the commutator scaling in the higher order remainder terms. For instance, for the time-independent Hamiltonian simulation, [149] deduces an error bound for the p-th order Trotter formula, in which the $\mathcal{O}(h^p)$ term has a commutator structure, but the higher order terms do not. This leads to complexity overhead when the spectral norms of the Hamiltonians become large. The work [44] fixes this issue by deriving an exact error representation, demonstrating the validity of the commutator scaling for high order terms as well. Therefore for time-dependent simulation, we also preserve all the terms in the error representation (at least for now). We can observe that all the terms in the exact representation, regardless of the order in h, are in the form of Eq.* (5.2.11)*, thus no overhead will be introduced by higher order terms and it is safe to bound them by the lowest order term later in estimating complexity.*

Proof of error representations

In this part, we derive the error representations of the first-order and second-order Trotter formulae, as presented in Lemma 35 - Lemma 38. All of the proofs consisting of the following two steps: One first compares the derivatives

$$\partial_h U(h,0) = (-\mathrm{i}f_1(h)H_1 - \mathrm{i}f_2(h)H_2)U(h,0), \tag{5.3.11}$$

and its numerical analogs $\partial_h U_{m,p}(h,0)$ ($m = g, s$ and $p = 1, 2$), and apply the variation of parameter formula (Lemma 34); Then the Taylor theorem (Lemma 33) is applied to further simplify the terms.

We first present the proof of Lemma 36, since its error representation contains fewest terms. The rest of the error representations, Lemma 35, Lemma 37 and Lemma 38, follow the exact same idea of proof, just involving more calculations.

Proof of Lemma 36. By taking derivative of $U_{g,1}(h,0)$ with respect to h, one has

$$\begin{aligned}
\partial_h U_{g,1}(h,0) &= -\mathrm{i} f_2(h) H_2 U_{g,1}(h,0) \\
&\quad + \exp\left(-\mathrm{i}\int_0^h f_2(s)ds H_2\right)(-\mathrm{i} f_1(h) H_1)\exp\left(-\mathrm{i}\int_0^h f_1(s)ds H_1\right) \\
&= (-\mathrm{i} f_1(h) H_1 - \mathrm{i} f_2(h) H_2)\, U_{g,1}(h,0) \\
&\quad + \exp\left(-\mathrm{i}\int_0^h f_2(s)ds H_2\right) E_{g,1}(h) \exp\left(-\mathrm{i}\int_0^h f_1(s)ds H_1\right) \qquad (5.3.12)
\end{aligned}$$

where $E_{g,1}(h)$ is defined as

$$E_{g,1}(h) = \mathrm{i} f_1(h)\left[\exp\left(\mathrm{ad}_{\mathrm{i}\int_0^h f_2(s')ds' H_2}\right) H_1 - H_1\right]. \qquad (5.3.13)$$

By applying Lemma 34 to Eq. (5.3.11) and Eq. (5.3.12), one obtains

$$U_{g,1}(h,0) = U(h,0) + \int_0^h U(h,s)\exp\left(-\mathrm{i}\int_0^s f_2(s')ds' H_2\right) E_{g,1}(s)\exp\left(-\mathrm{i}\int_0^s f_1(s')ds' H_1\right) ds. \qquad (5.3.14)$$

The representation of $E_{g,1}$, by Taylor's theorem (Lemma 33), reads

$$E_{g,1}(h) = \int_0^h f_1(h) f_2(s)\left(\exp\left(\mathrm{ad}_{\mathrm{i}\int_0^h f_2(s')ds' H_2}\right)([H_1, H_2])\right) ds. \qquad (5.3.15)$$

□

Proof of Lemma 35. One starts by taking derivative of $U_{s,1}(h,0)$ with respect to h, which reads

$$\begin{aligned}
\partial_h U_{s,1}(h,0) &= (-\mathrm{i} f_2(h) H_2 - \mathrm{i} h f_2'(h) H_2)\exp(-\mathrm{i} h f_2(h) H_2)\exp(-\mathrm{i} h f_1(h) H_1) \\
&\quad + \exp(-\mathrm{i} h f_2(h) H_2)(-\mathrm{i} f_1(h) H_1 - \mathrm{i} h f_1'(h) H_1)\exp(-\mathrm{i} h f_1(h) H_1) \\
&= (-\mathrm{i} f_2(h) H_2 - \mathrm{i} f_1(h) H_1)\, U_{s,1}(h,0) \\
&\quad + \exp(-\mathrm{i} h f_2(h) H_2)\, E_{s,1}(h)\exp(-\mathrm{i} h f_1(h) H_1), \qquad (5.3.16)
\end{aligned}$$

where $E_{s,1}(h)$ is defined as

$$E_{s,1}(h) = \mathrm{i}f_1(h)\left[\exp\left(\mathrm{ad}_{\mathrm{i}hf_2(h)H_2}\right)H_1 - H_1\right] - \mathrm{i}hf_1'(h)H_1 - \mathrm{i}hf_2'(h)H_2. \tag{5.3.17}$$

By applying Lemma 34 to Eq. (5.3.11) and Eq. (5.3.16), one has

$$U_{s,1}(h,0) = U(h,0) + \int_0^h U(h,s)\exp\left(-\mathrm{i}sf_2(s)H_2\right)E_{s,1}(s)\exp\left(-\mathrm{i}sf_1(s)H_1\right)ds. \tag{5.3.18}$$

It remains to derive the representation of $E_{s,1}$. The representation of $E_{s,1}$ can be derived from Taylor's theorem up to first-order. By Lemma 33

$$\begin{aligned}&\exp\left(\mathrm{ad}_{\mathrm{i}hf_2(h)H_2}\right)H_1 - H_1\\ &= \int_0^h \mathrm{i}f_2(s)\left(\exp\left(\mathrm{ad}_{\mathrm{i}sf_2(s)H_2}\right)([H_2,H_1])\right)ds + \int_0^h \mathrm{i}sf_2'(s)\left(\exp\left(\mathrm{ad}_{\mathrm{i}sf_2(s)H_2}\right)([H_2,H_1])\right)ds.\end{aligned} \tag{5.3.19}$$

Therefore, one has

$$\begin{aligned}E_{s,1}(h) = &\int_0^h f_1(h)f_2(s)\left(\exp\left(\mathrm{ad}_{\mathrm{i}sf_2(s)H_2}\right)([H_1,H_2])\right)ds - \mathrm{i}hf_1'(h)H_1 - \mathrm{i}hf_2'(h)H_2\\ &+ \int_0^h sf_1(h)f_2'(s)\left(\exp\left(\mathrm{ad}_{\mathrm{i}sf_2(s)H_2}\right)([H_1,H_2])\right)ds.\end{aligned} \tag{5.3.20}$$

□

Before proceeding, we first define the following quantities needed in the error represen-

tations of the second order standard and generalized Trotter formulae

$$
\begin{aligned}
E_{s,2}(h) = {} & \mathrm{i}\int_0^h f_1''(s)(h-s)H_1 ds - \frac{\mathrm{i}}{8}\int_0^h f_1''(s/2)(2h-s)H_1 ds \\
& - \frac{\mathrm{i}}{4}\int_0^h f_2''(s/2)(2h-s)H_2 ds \\
& - \frac{\mathrm{i}}{8}\int_0^h \left[f_1''(s/2)\exp\left(\mathrm{ad}_{-\mathrm{i}sf_2(s/2)H_2}\right)H_1\right](2h-s)ds \\
& + \frac{1}{4}\int_0^h \left[f_1'(s/2)f_2(s/2)\left(\exp\left(\mathrm{ad}_{-\mathrm{i}sf_2(s/2)H_2}\right)H_1\right)\right]hds \\
& + \mathrm{i}\int_0^h \left[f_2''(s)\exp\left(\mathrm{ad}_{i\frac{s}{2}f_1(s/2)H_1}\right)H_2\right](h-s)ds \\
& + \frac{1}{2}\int_0^h (f_1'(s/2)f_2(s/2)+f_1(s/2)f_2'(s/2))\left(\exp\left(\mathrm{ad}_{-\mathrm{i}sf_2(s/2)H_2}\right)[H_1,H_2]\right)(h-s)ds \\
& - \frac{1}{2}\int_0^h (f_2'(s)f_1(s/2)+f_2(s)f_1'(s/2))\left(\exp\left(\mathrm{ad}_{i\frac{s}{2}f_1(s/2)H_1}\right)[H_1,H_2]\right)(h-s)ds \\
& + \frac{\mathrm{i}}{2}\int_0^h \left[f_1(s/2)f_2^2(s/2)\left(\exp\left(\mathrm{ad}_{-\mathrm{i}sf_2(s/2)H_2}\right)[H_2,[H_1,H_2]]\right)\right](h-s)ds \\
& - \frac{\mathrm{i}}{4}\int_0^h \left[f_2(s)f_1^2(s/2)\left(\exp\left(\mathrm{ad}_{i\frac{s}{2}f_1(s/2)H_1}\right)[H_1,[H_1,H_2]]\right)\right](h-s)ds \\
& + \frac{1}{8}\int_0^h \left[f_1'(s/2)f_2'(s/2)\left(\exp\left(\mathrm{ad}_{-\mathrm{i}sf_2(s/2)H_2}\right)H_1\right)\right]shds \\
& + \frac{1}{4}\int_0^h \left(f_1'(s/2)f_2'(s/2)+\frac{1}{2}f_1(s/2)f_2''(s/2)\right) \\
& \qquad \times \left(\exp\left(\mathrm{ad}_{-\mathrm{i}sf_2(s/2)H_2}\right)[H_1,H_2]\right)s(h-s)ds \\
& - \frac{1}{4}\int_0^h \left(f_2'(s)f_1'(s/2)+\frac{1}{2}f_2(s)f_1''(s/2)\right)\exp\left(\mathrm{ad}_{i\frac{s}{2}f_1(s/2)H_1}\right)[H_1,H_2]s(h-s)ds \\
& + \frac{\mathrm{i}}{2}\int_0^h \left[f_1(s/2)f_2(s/2)f_2'(s/2)\left(\exp\left(\mathrm{ad}_{-\mathrm{i}sf_2(s/2)H_2}\right)[H_2,[H_1,H_2]]\right)\right]s(h-s)ds \\
& - \frac{\mathrm{i}}{4}\int_0^h \left[f_2(s)f_1(s/2)f_1'(s/2)\left(\exp\left(\mathrm{ad}_{i\frac{s}{2}f_1(s/2)H_1}\right)[H_1,[H_1,H_2]]\right)\right]s(h-s)ds \\
& + \frac{\mathrm{i}}{8}\int_0^h \left[f_1(s/2)f_2'^2(s/2)\left(\exp\left(\mathrm{ad}_{-\mathrm{i}sf_2(s/2)H_2}\right)[H_2,[H_1,H_2]]\right)\right]s^2(h-s)ds \\
& - \frac{\mathrm{i}}{16}\int_0^h \left[f_2(s)f_1'^2(s/2)\left(\exp\left(\mathrm{ad}_{i\frac{s}{2}f_1(s/2)H_1}\right)[H_1,[H_1,H_2]]\right)\right]s^2(h-s)ds, \quad (5.3.21)
\end{aligned}
$$

and

$$
\begin{aligned}
E_{g,2}(h) = &-\frac{h}{2} f_1(0) \int_0^h f_2'(s) ds [H_1, H_2] + \frac{h}{4} f_2(0) \int_0^h f_1'(s/2) ds [H_1, H_2] \\
&- f_2(h) \int_0^h (f_1'(s) - \frac{1}{4} f_1'(s/2)) \left(\exp\left(\mathrm{ad}_{\mathrm{i} \int_{s/2}^s f_1(s') ds' H_1} \right) [H_1, H_2] \right) (h-s) ds \\
&+ \frac{1}{2} f_1(h/2) \int_0^h f_2'(s) \left(\exp\left(\mathrm{ad}_{-\mathrm{i} \int_0^s f_2(s') ds' H_2} \right) [H_1, H_2] \right) (h-s) ds \\
&- \mathrm{i} f_2(h) \int_0^h \left(f_1(s) - \frac{1}{2} f_1(s/2) \right)^2 \left(\exp\left(\mathrm{ad}_{\mathrm{i} \int_{s/2}^s f_1(s') ds' H_1} \right) [H_1, [H_1, H_2]] \right) (h-s) ds \\
&+ \frac{\mathrm{i}}{2} f_1(h/2) \int_0^h f_2^2(s) \left(\exp\left(\mathrm{ad}_{-\mathrm{i} \int_0^s f_2(s') ds' H_2} \right) [H_2, [H_2, H_1]] \right) (h-s) ds. \qquad (5.3.22)
\end{aligned}
$$

Proof of Lemma 37. One first compute the derivative with respect to h of $U_{s,2}$

$$
\begin{aligned}
\partial_h U_{s,2} = &\left(-\mathrm{i} \frac{1}{2} f_1(h/2) - \mathrm{i} \frac{h}{4} f_1'(h/2) \right) H_1 \exp\left(-\mathrm{i} \frac{h}{2} f_1(h/2) H_1 \right) \\
&\quad \times \exp\left(-\mathrm{i} h f_2(h/2) H_2 \right) \exp\left(-\mathrm{i} \frac{h}{2} f_1(h/2) H_1 \right) \\
&+ \exp\left(-\mathrm{i} \frac{h}{2} f_1(h/2) H_1 \right) \left(-\mathrm{i} f_2(h/2) - \mathrm{i} \frac{h}{2} f_2'(h/2) \right) \\
&\quad \times H_2 \exp\left(-\mathrm{i} h f_2(h/2) H_2 \right) \exp\left(-\mathrm{i} \frac{h}{2} f_1(h/2) H_1 \right) \\
&+ \exp\left(-\mathrm{i} \frac{h}{2} f_1(h/2) H_1 \right) \exp\left(-\mathrm{i} h f_2(h/2) H_2 \right) \\
&\quad \times \left(-\mathrm{i} \frac{1}{2} f_1(h/2) - \mathrm{i} \frac{h}{4} f_1'(h/2) \right) H_1 \exp\left(-\mathrm{i} \frac{h}{2} f_1(h/2) H_1 \right) \\
= &\left(-\mathrm{i} f_1(h) H_1 - \mathrm{i} f_2(h) H_2 \right) U_{s,2} \\
&+ \exp\left(-\mathrm{i} \frac{h}{2} f_1(h/2) H_1 \right) E_{s,2}(h) \exp\left(-\mathrm{i} h f_2(h/2) H_2 \right) \exp\left(-\mathrm{i} \frac{h}{2} f_1(h/2) H_1 \right), \qquad (5.3.23)
\end{aligned}
$$

where $E_{s,2}(h)$ is defined as

$$\begin{aligned}E_{s,2}(h) = &\mathrm{i}f_1(h)H_1 - \mathrm{i}\frac{1}{2}f_1(h/2)H_1 - \mathrm{i}\frac{h}{4}f_1'(h/2)H_1 \\ &- \left(\mathrm{i}\frac{1}{2}f_1(h/2) + \mathrm{i}\frac{h}{4}f_1'(h/2)\right)\exp\left(\mathrm{ad}_{-\mathrm{i}hf_2(h/2)H_2}\right)H_1 \\ &+ \mathrm{i}f_2(h)\exp\left(\mathrm{ad}_{i\frac{h}{2}f_1(h/2)H_1}\right)H_2 - \left(\mathrm{i}f_2(h/2) + \frac{\mathrm{i}h}{2}f_2'(h/2)\right)H_2. \end{aligned} \tag{5.3.24}$$

Similar as the proofs for first-order formulae, applying Lemma 34 gives

$$\begin{aligned}U_{s,2}(h,0) = U(h,0) + \int_0^h U(h,s)\exp\left(-\mathrm{i}\frac{s}{2}f_1(s/2)H_1\right)E_{s,2}(s) \\ \times \exp\left(-\mathrm{i}sf_2(s/2)H_2\right)\exp\left(-\mathrm{i}\frac{s}{2}f_1(s/2)H_1\right)ds. \end{aligned} \tag{5.3.25}$$

The rest of the proof follows straightforward calculations. To be exact, one then applies the Taylor's theorem (Lemma 33) to expand each term in $E_{s,2}$ to second-order in terms of h with respect to 0. The first three terms can be expressed as

$$\mathrm{i}f_1(h)H_1 = \mathrm{i}f_1(0)H_1 + \mathrm{i}hf_1'(0)H_1 + \mathrm{i}\int_0^h f_1''(s)(h-s)H_1 ds, \tag{5.3.26}$$

$$-\mathrm{i}\frac{1}{2}f_1(h/2)H_1 = -\mathrm{i}\frac{1}{2}f_1(0)H_1 - \mathrm{i}\frac{h}{4}f_1'(0)H_1 - \mathrm{i}\frac{1}{8}\int_0^h f_1''(s/2)(h-s)H_1 ds, \tag{5.3.27}$$

$$-\mathrm{i}\frac{h}{4}f_1'(h/2)H_1 = -\mathrm{i}\frac{h}{4}f_1'(0)H_1 - \mathrm{i}\frac{h}{8}\int_0^h f_1''(s/2)H_1 ds, \tag{5.3.28}$$

Similarly, let us apply the Taylor theorem to the fourth term in $E_{s,2}$, which is the sum of

$$\begin{aligned}&-\mathrm{i}\frac{1}{2}f_1(h/2)\exp\left(\mathrm{ad}_{-\mathrm{i}hf_2(h/2)H_2}\right)H_1 \\ =&-\frac{\mathrm{i}}{2}f_1(0)H_1 - \frac{\mathrm{i}h}{4}f_1'(0)H_1 + \frac{h}{2}f_1(0)f_2(0)[H_1,H_2] \\ &-\frac{\mathrm{i}}{2}\int_0^h ds\frac{1}{4}f_1''(s/2)\exp\left(\mathrm{ad}_{-\mathrm{i}sf_2(s/2)H_2}\right)H_1(h-s) \\ &-\frac{\mathrm{i}}{2}\int_0^h ds f_1(s/2)\left(\mathrm{i}f_2(s/2) + \frac{\mathrm{i}s}{2}f_2'(s/2)\right)^2\exp\left(\mathrm{ad}_{-\mathrm{i}sf_2(s/2)H_2}\right)[H_2,[H_1,H_2]](h-s) \\ &-\frac{\mathrm{i}}{2}\int_0^h ds\left(\mathrm{i}(f_1'(s/2)f_2(s/2) + f_1(s/2)f_2'(s/2)) + \frac{\mathrm{i}s}{2}f_1'(s/2)f_2'(s/2) + \frac{\mathrm{i}s}{4}f_1(s/2)f_2''(s/2)\right) \\ &\qquad\times\exp\left(\mathrm{ad}_{-\mathrm{i}sf_2(s/2)H_2}\right)[H_1,H_2](h-s), \end{aligned} \tag{5.3.29}$$

and

$$
\begin{aligned}
&-\mathrm{i}\frac{h}{4}f_1'(h/2)\exp\left(\mathrm{ad}_{-\mathrm{i}hf_2(h/2)H_2}\right)H_1\\
=&-\frac{\mathrm{i}h}{4}f_1'(0)H_1-\frac{\mathrm{i}h}{4}\int_0^h ds\frac{1}{2}f_1''(s/2)\exp\left(\mathrm{ad}_{-\mathrm{i}sf_2(s/2)H_2}\right)H_1\\
&-\frac{\mathrm{i}h}{4}\int_0^h ds\mathrm{i}f_1'(s/2)\left(f_2(s/2)+\frac{s}{2}f_2'(s/2)\right)\exp\left(\mathrm{ad}_{-\mathrm{i}sf_2(s/2)H_2}\right)H_1, \qquad (5.3.30)
\end{aligned}
$$

The fifth term in $E_{s,2}$ reads

$$
\begin{aligned}
&\mathrm{i}f_2(h)\exp\left(\mathrm{ad}_{i\frac{h}{2}f_1(h/2)H_1}\right)H_2\\
=&\,\mathrm{i}f_2(0)H_2+\mathrm{i}hf_2'(0)H_2-\frac{h}{2}f_2(0)f_1(0)[H_1,H_2]\\
&+\mathrm{i}\int_0^h ds(h-s)f_2''(s)\exp\left(\mathrm{ad}_{i\frac{s}{2}f_1(s/2)H_1}\right)H_2(h-s)\\
&+\mathrm{i}\int_0^h ds(h-s)f_2(s)\left(\frac{\mathrm{i}}{2}f_1(s/2)+\frac{\mathrm{i}s}{4}f_1'(s/2)\right)^2\exp\left(\mathrm{ad}_{i\frac{s}{2}f_1(s/2)H_1}\right)\Big[H_1,[H_1,H_2]]\Big](h-s)\\
&+\mathrm{i}\int_0^h ds(h-s)\left(\frac{\mathrm{i}}{2}f_2'(s)f_1(s/2)+\frac{\mathrm{i}}{2}f_2(s)f_1'(s/2)+\frac{\mathrm{i}s}{4}f_2'(s)f_1'(s/2)+\frac{\mathrm{i}s}{8}f_2(s)f_1''(s/2)\right)\\
&\qquad\qquad\times\exp\left(\mathrm{ad}_{i\frac{s}{2}f_1(s/2)H_1}\right)[H_1,H_2](h-s), \qquad (5.3.31)
\end{aligned}
$$

The last term in $E_{s,2}$ is the sum of

$$
-\mathrm{i}f_2(h/2)H_2=-\mathrm{i}f_2(0)H_2-\mathrm{i}\frac{h}{2}f_2'(0)H_2-\mathrm{i}\int_0^h\frac{1}{4}f_2''(s/2)(h-s)ds, \qquad (5.3.32)
$$

and

$$
-\frac{\mathrm{i}h}{2}f_2'(h/2)H_2=-\frac{\mathrm{i}h}{2}f_2'(0)H_2-\frac{\mathrm{i}h}{4}\int_0^h f_2''(s/2)H_2ds. \qquad (5.3.33)
$$

Notice that, if we add the above eight equations together, all the zeroth-order and first-order terms of h cancel, then the desired expression of $E_{s,2}(h)$ is achieved.

□

Proof of Lemma 38. The strategy for proving Lemma 38 is the same as that for Lemma 37. By taking derivatives with respect to h in both $U(h,0)$ and $U_{g,2}(h,0)$, we have Eq. (5.3.11)

and

$$\begin{aligned}
\partial_h U_{g,2}(h,0) &= \left(-\mathrm{i}f_1(h) + \frac{\mathrm{i}}{2}f_1(h/2)\right) H_1 U_{g,2}(h,0) \\
&\quad + \exp\left(-\mathrm{i}\int_{h/2}^{h} f_1(s)dsH_1\right)(-\mathrm{i}f_2(h)H_2) \\
&\quad\quad \times \exp\left(-\mathrm{i}\int_0^h f_2(s)dsH_2\right)\exp\left(-\mathrm{i}\int_0^{h/2} f_1(s)dsH_1\right) \\
&\quad + \exp\left(-\mathrm{i}\int_{h/2}^{h} f_1(s)dsH_1\right)\exp\left(-\mathrm{i}\int_0^h f_2(s)dsH_2\right) \\
&\quad\quad \times \left(-\frac{\mathrm{i}}{2}f_1(h/2)/H_1\right)\exp\left(-\mathrm{i}\int_0^{h/2} f_1(s)dsH_1\right) \\
&= -(\mathrm{i}f_1(h)H_1 + \mathrm{i}f_2(h)H_2)U_{g,2}(h,0) \\
&\quad + \exp\left(-\mathrm{i}\int_{h/2}^{h} f_1(s)dsH_1\right)E_{g,2}(h) \\
&\quad\quad \times \exp\left(-\mathrm{i}\int_0^h f_2(s)dsH_2\right)\exp\left(-\mathrm{i}\int_0^{h/2} f_1(s)dsH_1\right), \qquad (5.3.34)
\end{aligned}$$

where $E_{g,2}(h)$ denotes

$$\begin{aligned}
E_{g,2}(h) &= if_2(h)\exp(\mathrm{ad}_{\mathrm{i}\int_{h/2}^{h} f_1(s)dsH_1})H_2 + \frac{\mathrm{i}}{2}f_1(h/2)H_1 \\
&\quad - \mathrm{i}f_2(h)H_2 - \frac{\mathrm{i}}{2}f_1(h/2)\exp\left(\mathrm{ad}_{-\mathrm{i}\int_0^h f_2(s)dsH_2}\right)H_1 \\
&= \mathrm{i}f_2(h)\left[\exp\left(\mathrm{ad}_{\mathrm{i}\int_{h/2}^{h} f_1(s)dsH_1}\right)H_2 - H_2\right] \\
&\quad - \frac{\mathrm{i}}{2}f_1(h/2)\left[\exp\left(\mathrm{ad}_{-\mathrm{i}\int_0^h f_2(s)dsH_2}\right)H_1 - H_1\right]. \qquad (5.3.35)
\end{aligned}$$

By applying Lemma 34, we have

$$\begin{aligned}
U_{g,2}(h,0) = U(h,0) + \int_0^h U(h,s)\exp\left(-\mathrm{i}\int_{s/2}^{s} f_1(s')ds'H_1\right)E_{g,2}(s) \\
\times \exp\left(-\mathrm{i}\int_0^s f_2(s')ds'H_2\right)\exp\left(-\mathrm{i}\int_0^{s/2} f_1(s')ds'H_1\right)ds. \quad (5.3.36)
\end{aligned}$$

It remains to derive the representation of $E_{g,2}$. It follows from the Taylor's theorem (Lemma 33) that

$$\begin{aligned}
&\exp\left(\mathrm{ad}_{\mathrm{i}\int_{h/2}^{h} f_1(s)dsH_1}\right) H_2 - H_2 \\
&= \frac{\mathrm{i}h}{2} f_1(0)[H_1, H_2] + \int_0^h \left(\mathrm{i}f_1'(s) - \frac{1}{4}\mathrm{i}f_1'(s/2)\right) \left(\exp\left(\mathrm{ad}_{\mathrm{i}\int_{s/2}^{s} f_1(s')ds'H_1}\right)[H_1, H_2]\right)(h-s)ds \\
&\quad + \int_0^h \left(\mathrm{i}f_1(s) - \frac{1}{2}\mathrm{i}f_1(s/2)\right)^2 \left(\exp\left(\mathrm{ad}_{\mathrm{i}\int_{s/2}^{s} f_1(s')ds'H_1}\right)[H_1, [H_1, H_2]]\right)(h-s)ds,
\end{aligned}$$

and

$$\begin{aligned}
&\exp\left(\mathrm{ad}_{-\mathrm{i}\int_{0}^{h} f_2(s)dsH_2}\right) H_1 - H_1 \\
&= -\mathrm{i}hf_2(0)[H_2, H_1] - \int_0^h \mathrm{i}f_2'(s) \left(\exp\left(\mathrm{ad}_{-\mathrm{i}\int_{0}^{s} f_2(s')ds'H_2}\right)[H_2, H_1]\right)(h-s)ds \\
&\quad + \int_0^h (\mathrm{i}f_2(s))^2 \left(\exp\left(\mathrm{ad}_{-\mathrm{i}\int_{0}^{s} f_2(s')ds'H_2}\right)[H_2, [H_2, H_1]]\right)(h-s)ds.
\end{aligned}$$

Thus we have

$$
\begin{aligned}
E_{g,2}(h) = {} & \mathrm{i}f_2(h)\frac{\mathrm{i}h}{2}f_1(0)[H_1,H_2] + \frac{\mathrm{i}}{2}f_1(h/2)\mathrm{i}hf_2(0)[H_2,H_1] \\
& + \mathrm{i}f_2(h)\int_0^h \left(\mathrm{i}f_1'(s) - \frac{\mathrm{i}}{4}f_1'(s/2)\right)\left(\exp\left(\mathrm{ad}_{\mathrm{i}\int_{s/2}^s f_1(s')ds'H_1}\right)[H_1,H_2]\right)(h-s)ds \\
& + \mathrm{i}f_2(h)\int_0^h \left(\mathrm{i}f_1(s) - \frac{\mathrm{i}}{2}f_1(s/2)\right)^2\left(\exp\left(\mathrm{ad}_{\mathrm{i}\int_{s/2}^s f_1(s')ds'H_1}\right)[H_1,[H_1,H_2]]\right)(h-s)ds \\
& + \frac{\mathrm{i}}{2}f_1(h/2)\int_0^h \mathrm{i}f_2'(s)\left(\exp\left(\mathrm{ad}_{-\mathrm{i}\int_0^s f_2(s')ds'H_2}\right)[H_2,H_1]\right)(h-s)ds \\
& - \frac{\mathrm{i}}{2}f_1(h/2)\int_0^h (\mathrm{i}f_2(s))^2\left(\exp\left(\mathrm{ad}_{-\mathrm{i}\int_0^s f_2(s')ds'H_2}\right)[H_2,[H_2,H_1]]\right)(h-s)ds \\
= {} & -\frac{h}{2}f_1(0)\int_0^h f_2'(s)ds[H_1,H_2] + \frac{h}{4}f_2(0)\int_0^h f_1'(s/2)ds[H_1,H_2] \\
& - f_2(h)\int_0^h (f_1'(s) - \frac{1}{4}f_1'(s/2))\left(\exp\left(\mathrm{ad}_{\mathrm{i}\int_{s/2}^s f_1(s')ds'H_1}\right)[H_1,H_2]\right)(h-s)ds \\
& + \frac{1}{2}f_1(h/2)\int_0^h f_2'(s)\left(\exp\left(\mathrm{ad}_{-\mathrm{i}\int_0^s f_2(s')ds'H_2}\right)[H_1,H_2]\right)(h-s)ds \\
& - \mathrm{i}f_2(h)\int_0^h \left(f_1(s) - \frac{1}{2}f_1(s/2)\right)^2\left(\exp\left(\mathrm{ad}_{\mathrm{i}\int_{s/2}^s f_1(s')ds'H_1}\right)[H_1,[H_1,H_2]]\right)(h-s)ds \\
& + \frac{\mathrm{i}}{2}f_1(h/2)\int_0^h f_2^2(s)\left(\exp\left(\mathrm{ad}_{-\mathrm{i}\int_0^s f_2(s')ds'H_2}\right)[H_2,[H_2,H_1]]\right)(h-s)ds.
\end{aligned}
$$

□

5.4 Operator norm error bounds

We first establish the operator norm error bounds. In this section we assume that H_1 and H_2 are two bounded operators. The operator norm error bounds can be directly obtained from the error representations by bounding the operator norms of all the unitaries by 1.

Theorem 40. *The error of each standard/generalized Trotter step measured in the operator norm is as follows:*

1. *First-order standard Trotter formula:*

$$\|U_{s,1}(h,0) - U(h,0)\| \leq \alpha_{s,1}h^2 + \beta_{s,1}h^3, \tag{5.4.1}$$

where

$$\alpha_{s,1} = \frac{1}{2}\|f_1'\|_\infty\|H_1\| + \frac{1}{2}\|f_2'\|_\infty\|H_2\| + \frac{1}{2}\|f_1\|_\infty\|f_2\|_\infty\|[H_1, H_2]\| \tag{5.4.2}$$

and

$$\beta_{s,1} = \frac{1}{6}\|f_1\|_\infty\|f_2'\|_\infty\|[H_1, H_2]\|. \tag{5.4.3}$$

2. *First-order generalized Trotter formula:*

$$\|U_{g,1}(h,0) - U(h,0)\| \leq \alpha_{g,1}h^2 \tag{5.4.4}$$

where

$$\alpha_{g,1} = \frac{1}{2}\|f_1\|_\infty\|f_2\|_\infty\|[H_1, H_2]\|. \tag{5.4.5}$$

3. *Second-order standard Trotter formula:*

$$\|U_{s,2}(h,0) - U(h,0)\| \leq \alpha_{s,2}h^3 + \beta_{s,2}h^4 + \gamma_{s,2}h^5, \tag{5.4.6}$$

where

$$\begin{aligned}\alpha_{s,2} = {} & \frac{7}{24}\|f_1''\|_\infty\|H_1\| + \frac{1}{12}\|f_1'\|_\infty\|f_2\|_\infty\|H_1\| + \frac{7}{24}\|f_2''\|_\infty\|H_2\| \\ & + \frac{1}{6}(\|f_1'\|_\infty\|f_2\|_\infty + \|f_1\|_\infty\|f_2'\|_\infty)\|[H_1, H_2]\| \\ & + \frac{1}{24}\|f_1\|_\infty^2\|f_2\|_\infty\|[H_1,[H_1, H_2]]\| + \frac{1}{12}\|f_1\|_\infty\|f_2\|_\infty^2\|[H_2,[H_1, H_2]]\|,\end{aligned} \tag{5.4.7}$$

$$\begin{aligned}\beta_{s,2} = {} & \frac{1}{64}\|f_1'\|_\infty\|f_2'\|_\infty\|H_1\| \\ & + \left(\frac{1}{192}\|f_1\|_\infty\|f_2''\|_\infty + \frac{1}{192}\|f_1''\|_\infty\|f_2\|_\infty + \frac{1}{48}\|f_1'\|_\infty\|f_2'\|_\infty\right)\|[H_1, H_2]\| \\ & + \frac{1}{96}\|f_1\|_\infty\|f_1'\|_\infty\|f_2\|_\infty\|[H_1,[H_1, H_2]]\| + \frac{1}{48}\|f_1\|_\infty\|f_2\|_\infty\|f_2'\|_\infty\|[H_2,[H_1, H_2]]\|,\end{aligned} \tag{5.4.8}$$

and

$$\gamma_{s,2} = \frac{1}{960}\|f_1'\|_\infty^2\|f_2\|_\infty\|[H_1,[H_1, H_2]]\| + \frac{1}{480}\|f_1\|_\infty\|f_2'\|_\infty^2\|[H_2,[H_1, H_2]]\|. \tag{5.4.9}$$

4. Second-order generalized Trotter formula:

$$\|U_{g,2}(h,0) - U(h,0)\| \leq \alpha_{g,2} h^3, \tag{5.4.10}$$

where

$$\begin{aligned}\alpha_{g,2} = &\left(\frac{7}{12}\|f_1\|_\infty \|f_2'\|_\infty + \frac{11}{24}\|f_1'\|_\infty \|f_2\|_\infty\right) \|[H_1, H_2]\| \\ &+ \frac{3}{8}\|f_1\|_\infty^2 \|f_2\|_\infty \|[H_1,[H_1,H_2]]\| + \frac{1}{12}\|f_1\|_\infty \|f_2\|_\infty^2 \|[H_2,[H_2,H_1]]\|.\end{aligned} \tag{5.4.11}$$

Remark 41 (The preconstants in Theorem 40)**.** *First, the preconstants of the standard Trotter formula involve the norms of both the Hamiltonians as well as their commutators, while the preconstants of the generalized Trotter formula of the first order only involve the commutator $[H_1, H_2]$, and those of the second order involve further nested commutators. Second, the p-th order standard Trotter scheme ($p = 1, 2$) depends on the p-th order derivatives of the control functions while the p-th generalized Trotter scheme depends only on $(p-1)$-th order derivatives. In this sense, when the time derivatives of f_1 and f_2 are large (or when the regularity of f_1, f_2 are limited), the generalized Trotter formula can further outperform the standard Trotter method.*

Now we move on to the global error bounds. To obtain an approximation of the exact unitary evolution up to time T, we can divide the time interval $[0, T]$ into L equilength segments and implement Trotter discretization on each segment. Since the evolutions for both continuous and discretized cases are unitary, the global error is simply a linear accumulation of local errors at each time step. For sufficiently large L, the total error can be controlled to be arbitrarily small.

Theorem 42. *Let $T > 0$ be the evolution time, and the dynamics Eq.* (5.1.1) *is discretized via standard and generalized Trotter formulae with L equidistant time steps (thus the time step size $h = T/L$). Then*

$$\left\|\prod_{l=1}^{L} U_{s,1}\left(\frac{lT}{L}, \frac{(l-1)T}{L}\right) - U(T,0)\right\| \leq \alpha_{s,1}\frac{T^2}{L} + \beta_{s,1}\frac{T^3}{L^2}, \tag{5.4.12}$$

$$\left\|\prod_{l=1}^{L} U_{g,1}\left(\frac{lT}{L}, \frac{(l-1)T}{L}\right) - U(T,0)\right\| \leq \alpha_{g,1}\frac{T^2}{L}, \tag{5.4.13}$$

$$\left\|\prod_{l=1}^{L} U_{s,2}\left(\frac{lT}{L}, \frac{(l-1)T}{L}\right) - U(T,0)\right\| \leq \alpha_{s,2}\frac{T^3}{L^2} + \beta_{s,2}\frac{T^4}{L^3} + \gamma_{s,2}\frac{T^5}{L^4}, \tag{5.4.14}$$

$$\left\|\prod_{l=1}^{L} U_{g,2}\left(\frac{lT}{L}, \frac{(l-1)T}{L}\right) - U(T,0)\right\| \leq \alpha_{g,2}\frac{T^3}{L^2}, \tag{5.4.15}$$

where preconstants α and β are defined in Theorem 40.

5.5 Vector norm error bounds

Now we consider the case when H_1 is an approximation of an unbounded operator, while H_2 remains reasonably bounded. In this section, we assume that $\|H_1\| \gg \|H_2\|$, and the functions f_1, f_2, as well as their first and second-order derivatives are bounded. To simplify the proof and emphasize our focus on overcoming the difficulty brought by H_1, throughout this section we will not track the explicit dependence on H_2, f_1 and f_2. We use the notation $\widetilde{C}$ with a tilde above to denote preconstants (with possibly varying sizes in different inequalities) which can depend polynomially on H_2, $\|f_1^{(k)}\|_\infty$ and $\|f_2^{(k)}\|_\infty$ but do not depend on H_1. Furthermore, we make the following assumptions.

Assumption 43 (Bounds of commutators)**.** *We assume H_1 is a positive semidefinite operator, and there exists an operator D_1 such that $H_1 = D_1^\dagger D_1$. Furthermore, we assume for any vector $\vec{v}$, there exist constants $\widetilde{C}_1, \widetilde{C}_2$ such that*

$$\|[H_1, H_2]\vec{v}\| \leq \widetilde{C}_1(\|D_1\vec{v}\| + \|\vec{v}\|), \tag{5.5.1}$$

and

$$\|[H_1, [H_1, H_2]]\vec{v}\| \leq \widetilde{C}_2(\|H_1\vec{v}\| + \|\vec{v}\|). \tag{5.5.2}$$

Assumption 43 has been used in previous works [81, 75] related to the vector norm error bounds of time-independent Trotter formula and exponential integrators for $H = -\Delta + V(x)$, where $H_1 = -\Delta$ is positive semidefinite and $H_2 = V(x)$ is a bounded operator. It will be helpful to understand Assumption 43 from a continuous analog, in which $H_1 = D_1^\dagger D_1$ and D_1 is a first-order differential operator. A direct calculation shows that the operator

$$[-\Delta, V] = -\partial_x^2 V - 2(\partial_x V)\partial_x$$

is a first-order differential operator, and

$$[-\Delta, [-\Delta, V]] = \partial_x^4 V + 4\partial_x^3 V \partial_x + 4\partial_x^2 V \partial_x^2 \tag{5.5.3}$$

is a second-order differential operator, given that $V(x)$ is a C^4 function. These are exactly what Eqs. (5.5.1) and (5.5.2) are addressing. Furthermore, if the wavefunction v is smooth enough with bounded derivatives, then the right hand sides of these inequalities are bounded, which provides the key motivation and possibility to establish vector norm error bounds and obtain improvement in complexity estimates. In the context of quantum simulation, since all matrices and vectors are finite dimensional, we omit the explicit statements of regularity assumptions of v below.

As we have briefly discussed before, starting from the error representations, the vector norm error bounds are just linear combinations of the terms in the form of Eq. (5.2.12), and the key step to prove vector norm error bounds is to exchange the order of a Hamiltonian or an commutator with matrix exponentials. We find that such an exchange of order will not introduce any overhead in the error bounds. This is established by the following lemma.

Lemma 44. *Under Assumption 43, we have the following:*

1. *For any vector $\vec{v}$,*
$$\|D_1\vec{v}\| \leq \|H_1\vec{v}\| + \|\vec{v}\|. \tag{5.5.4}$$

2. *Let ξ be any real number such that $(\widetilde{C}_1 + \|H_2\|)|\xi| \leq 1/2$. Then for any vector $\vec{v}$,*
$$\|H_1 \exp(\mathrm{i}\xi H_2)\,\vec{v}\| \leq 2(\|H_1\vec{v}\| + \|\vec{v}\|). \tag{5.5.5}$$

3. *Let K be a positive integer, and H_{l_k} be either H_1 or H_2, and ξ_k be some real numbers for $1 \leqslant k \leqslant K$. Assume that $(\widetilde{C}_1 + \|H_2\|)|\xi_k|h \leq 1/2$, then for any vector $\vec{v}$, all the following inequalities hold:*
$$\begin{aligned}
\left\| H_1 \left[\prod_{k=1}^{K} \exp(\mathrm{i}h\xi_k H_{l_k})\right] \vec{v} \right\| &\leq \widetilde{C}(\|H_1\vec{v}\| + \|\vec{v}\|),\\
\left\| H_2 \left[\prod_{k=1}^{K} \exp(\mathrm{i}h\xi_k H_{l_k})\right] \vec{v} \right\| &\leq \widetilde{C}\|\vec{v}\|,\\
\left\| [H_1, H_2] \left[\prod_{k=1}^{K} \exp(\mathrm{i}h\xi_k H_{l_k})\right] \vec{v} \right\| &\leq \widetilde{C}\left(\sqrt{\|\vec{v}\|\|H_1\vec{v}\|} + \|\vec{v}\|\right),\\
\left\| [H_1, [H_1, H_2]] \left[\prod_{k=1}^{K} \exp(\mathrm{i}h\xi_k H_{l_k})\right] \vec{v} \right\| &\leq \widetilde{C}(\|H_1\vec{v}\| + \|\vec{v}\|),\\
\left\| [H_2, [H_2, H_1]] \left[\prod_{k=1}^{K} \exp(\mathrm{i}h\xi_k H_{l_k})\right] \vec{v} \right\| &\leq \widetilde{C}(\|H_1\vec{v}\| + \|\vec{v}\|).
\end{aligned} \tag{5.5.6}$$
for some constant $\widetilde{C} > 0$ depending only on $\|H_2\|$.

Proof. 1. By the definition of D_1 and the Cauchy-Schwarz inequality,
$$\|D_1\vec{v}\|^2 = (D_1\vec{v})^\dagger D_1\vec{v} = \vec{v}^\dagger H_1\vec{v} \leq \|\vec{v}\|\|H_1\vec{v}\| \leq (\|H_1\vec{v}\| + \|\vec{v}\|)^2. \tag{5.5.7}$$

2. We start with Taylor's theorem of $\exp(\mathrm{i}t\xi H_2)$ up to first-order, then

$$\begin{aligned} H_1 \exp(\mathrm{i}t\xi H_2)\vec{v} &= H_1\vec{v} + \int_0^t \mathrm{i}\xi H_1 H_2 \exp(\mathrm{i}\alpha\xi H_2)\,\vec{v} d\alpha \\ &= H_1\vec{v} + \int_0^t \mathrm{i}\xi [H_1, H_2] \exp(\mathrm{i}\alpha\xi H_2)\,\vec{v} d\alpha + \int_0^t \mathrm{i}\xi H_2 H_1 \exp(\mathrm{i}\alpha\xi H_2)\,\vec{v} d\alpha. \end{aligned}$$

The norm can be estimated using Eq. (5.5.1) as

$$\begin{aligned} \|H_1 \exp(\mathrm{i}t\xi H_2)\vec{v}\| &\le \|H_1\vec{v}\| + \int_0^t |\xi| \|[H_1, H_2] \exp(\mathrm{i}\alpha\xi H_2)\,\vec{v}\| d\alpha \\ &\quad + \int_0^t |\xi| \|H_2\| \, \|H_1 \exp(\mathrm{i}\alpha\xi H_2)\,\vec{v}\| \, d\alpha \\ &\le \|H_1\vec{v}\| + \widetilde{C}_1 |\xi| \|\vec{v}\| t + \int_0^t \widetilde{C}_1 |\xi| \|D_1 \exp(\mathrm{i}\alpha\xi H_2)\,\vec{v}\| d\alpha \\ &\quad + \int_0^t |\xi| \|H_2\| \, \|H_1 \exp(\mathrm{i}\alpha\xi H_2)\,\vec{v}\| \, d\alpha. \end{aligned} \tag{5.5.8}$$

Define $M(t) := \|H_1 \exp(\mathrm{i}t\xi H_2)\vec{v}\|$, and it follows from Eq. (5.5.4) that

$$\|D_1 \exp(\mathrm{i}t\xi H_2)\,\vec{v}\| \le M(t) + \|\vec{v}\|.$$

Thus Eq. (5.5.8) can be rewritten as

$$M(t) \le \|H_1\vec{v}\| + 2\widetilde{C}_1 |\xi| \|\vec{v}\| t + \int_0^t |\xi| \left(\widetilde{C}_1 + \|H_2\|\right) M(\alpha)\, d\alpha.$$

Since $\|H_1\vec{v}\| + 2\widetilde{C}_1|\xi|\|\vec{v}\|t$ is non-decreasing with respect to t, applying Gronwall's inequality yields the bound

$$M(t) \le \left(\|H_1\vec{v}\| + 2\widetilde{C}_1|\xi|\|\vec{v}\|t\right) \exp\left(|\xi|(\widetilde{C}_1 + \|H_2\|)t\right).$$

Finally taking $t = 1$ and applying the condition on ξ, the desired result is achieved

$$\|H_1 \exp(\mathrm{i}\xi H_2)\vec{v}\| \le \left(\|H_1\vec{v}\| + 2\widetilde{C}_1|\xi|\|\vec{v}\|\right) \exp\left(|\xi|(\widetilde{C}_1 + \|H_2\|)\right) \le 2(\|H_1\vec{v}\| + \|\vec{v}\|).$$

3. We first show that it suffices to only prove the first inequality in Eq. (5.5.6). Let $\vec{w} = \left[\prod_{k=1}^{K} \exp(\mathrm{i}h\xi_k H_{l_k})\right] \vec{v}$. Since H_2 is bounded, $\|H_2\vec{w}\|$ is directly bounded by $\widetilde{C}\|\vec{v}\|$. By Eqs. (5.5.1) and (5.5.7), and the fact $\|\vec{w}\| = \|\vec{v}\|$, we have

$$\|[H_1, H_2]\vec{w}\| \le \widetilde{C}_1(\|D_1\vec{w}\| + \|\vec{w}\|) \le \widetilde{C}\left(\sqrt{\|\vec{w}\|\|H_1\vec{w}\|} + \|\vec{w}\|\right) = \widetilde{C}\left(\sqrt{\|\vec{v}\|\|H_1\vec{w}\|} + \|\vec{v}\|\right). \tag{5.5.9}$$

Similarly Eq. (5.5.2) gives

$$\|[H_1,[H_1,H_2]]\vec{w}\| \le \widetilde{C}_2(\|H_1\vec{w}\| + \|\vec{w}\|) = \widetilde{C}_2(\|H_1\vec{w}\| + \|\vec{v}\|). \tag{5.5.10}$$

Furthermore,

$$\begin{aligned}
\|[H_2,[H_2,H_1]]\vec{w}\| &\le \|H_2[H_2,H_1]\vec{w}\| + \|[H_2,H_1]H_2\vec{w}\| \\
&\le \widetilde{C}\|[H_2,H_1]\vec{w}\| + \widetilde{C}_1(\|D_1H_2\vec{w}\| + \|H_2\vec{w}\|) \\
&\le \widetilde{C}(\|D_1\vec{w}\| + \|\vec{w}\|) + \widetilde{C}(\|H_1H_2\vec{w}\| + \|H_2\vec{w}\|) \\
&\le \widetilde{C}(\|H_1\vec{w}\| + \|\vec{w}\|) + \widetilde{C}(\|[H_1,H_2]\vec{w}\| + \|H_2H_1\vec{w}\| + \|\vec{w}\|) \\
&\le \widetilde{C}(\|H_1\vec{w}\| + \|\vec{w}\|) + \widetilde{C}(\|D_1\vec{w}\| + \|\vec{w}\|) \\
&\le \widetilde{C}(\|H_1\vec{w}\| + \|\vec{w}\|) = \widetilde{C}(\|H_1\vec{w}\| + \|\vec{v}\|).
\end{aligned} \tag{5.5.11}$$

Therefore we only need to bound $\|H_1\vec{w}\|$ further by $\widetilde{C}(\|H_1\vec{v}\| + \|\vec{v}\|)$.

Notice that $\|H_1\exp(\mathrm{i}\xi H_1)\vec{v}\| = \|\exp(\mathrm{i}\xi H_1)H_1\vec{v}\| = \|H_1\vec{v}\|$, together with Eq. (5.5.5),

$$\|H_1\exp(\mathrm{i}\xi H_l)\vec{v}\| \le 2(\|H_1\vec{v}\| + \|\vec{v}\|) \tag{5.5.12}$$

for H_l being either H_1 or H_2. Then we have the recursive relation

$$\begin{aligned}
\left\| H_1\left[\prod_{k=1}^{K}\exp\left(\mathrm{i}h\xi_k H_{l_k}\right)\right]\vec{v}\right\| &\le 2\left\| H_1\left[\prod_{k=1}^{K-1}\exp\left(\mathrm{i}h\xi_k H_{l_k}\right)\right]\vec{v}\right\| + 2\left\|\left[\prod_{k=1}^{K-1}\exp\left(\mathrm{i}h\xi_k H_{l_k}\right)\right]\vec{v}\right\| \\
&= 2\left\| H_1\left[\prod_{k=1}^{K-1}\exp\left(\mathrm{i}h\xi_k H_{l_k}\right)\right]\vec{v}\right\| + 2\|\vec{v}\|.
\end{aligned} \tag{5.5.13}$$

By applying this estimation K times, we obtain the desired result

$$\left\| H_1\left[\prod_{k=1}^{K}\exp\left(\mathrm{i}h\xi_k H_{l_k}\right)\right]\vec{v}\right\| \le \widetilde{C}\left(\|H_1\vec{v}\| + \|\vec{v}\|\right). \tag{5.5.14}$$

□

Now we are ready to state our main theorems for the vector norm estimates.

Theorem 45. *For any vector v and time step size $h \le (\|f_1\|_\infty + \|f_2\|_\infty)^{-1}(\widetilde{C}_1 + \|H_2\|)^{-1}/2$, there exists a constant $\widetilde{C}$ such that*

$$\begin{aligned}
&\|U_{s,1}(h,0)\vec{v} - U(h,0)\vec{v}\| \le \widetilde{C}h^2(\|H_1\vec{v}\| + \|\vec{v}\|), \\
&\|U_{g,1}(h,0)\vec{v} - U(h,0)\vec{v}\| \le \widetilde{C}h^2(\sqrt{\|\vec{v}\|\|H_1\vec{v}\|} + \|\vec{v}\|), \\
&\|U_{s,2}(h,0)\vec{v} - U(h,0)\vec{v}\| \le \widetilde{C}h^3(\|H_1\vec{v}\| + \|\vec{v}\|), \\
&\|U_{g,2}(h,0)\vec{v} - U(h,0)\vec{v}\| \le \widetilde{C}h^3(\|H_1\vec{v}\| + \|\vec{v}\|).
\end{aligned}$$

Proof. We start with the error representations (**??** 35–38). By multiplying a vector $\vec{v}$ on the right of the error representations, bounding all the $f_j^{(k)}$ by its supremum, bounding all the higher order terms involving the (nested) commutators by second-order terms, and bounding all the unitaries multiplied on the left by 1, we obtain

$$\begin{aligned}
\|U_{s,1}(h,0)\vec{v} - U(h,0)\vec{v}\| &\leq \widetilde{C}\theta_{s,1}h^2,\\
\|U_{g,1}(h,0)\vec{v} - U(h,0)\vec{v}\| &\leq \widetilde{C}\theta_{g,1}h^2,\\
\|U_{s,2}(h,0)\vec{v} - U(h,0)\vec{v}\| &\leq \widetilde{C}\theta_{s,2}h^3,\\
\|U_{g,2}(h,0)\vec{v} - U(h,0)\vec{v}\| &\leq \widetilde{C}\theta_{g,2}h^3,
\end{aligned}$$

where θ_s and θ_g are linear combinations of the terms in the form of Eq. (5.5.6), with an exception that $\theta_{g,1}$ only consists terms like $\|[H_1, H_2]\vec{w}\|$. Then part 3 of Lemma 44 completes the proof. □

Similar to the operator norm error bound case, from Theorem 45, we can establish the global error bound and estimate the total number of time steps we need to achieve a desired accuracy using standard and generalized Trotter formulae. The proof of the global error bound is essentially the same as the standard argument for error accumulation in quantum computing, that is, to replace the exact evolution operator by numerical evolution operator step by step and bound each step by local error bound. In the operator norm case, it does not matter whether we replace the local evolution operator in a forward or backward fashion. However, in the vector norm case, the order of the replacements indeed matters. In particular, we would like to obtain an error bound that depends on the exact, instead of the numerical solution of the dynamics. We state our vector norm global error bound in Theorem 46, and provide a complete proof for the second-order generalized Trotter formula.

Theorem 46. *Let $T > 0$ be the evolution time, $\vec{\psi}(t)$ be the exact solution of the dynamics Eq.* (5.1.1)*, and the dynamics Eq.* (5.1.1) *is discretized via standard and generalized Trotter formulae with L time steps such that the time step size $h = T/L$ is bounded by $(\|f_1\|_\infty +$*

$\|f_2\|_\infty)^{-1}(\widetilde{C}_1 + \|H_2\|)^{-1}/2$. *Then there exists a constant* $\widetilde{C}$ *such that*

$$\left\| \left(\prod_{l=1}^{L} U_{s,1}\left(\frac{lT}{L}, \frac{(l-1)T}{L}\right)\right)\vec{\psi}(0) - U(T,0)\vec{\psi}(0)\right\| \leq \widetilde{C}\frac{T^2}{L}\left(\sup_{t\in[0,T]}\|H_1\vec{\psi}(t)\| + \|\vec{\psi}(0)\|\right), \tag{5.5.15}$$

$$\begin{aligned}&\left\| \left(\prod_{l=1}^{L} U_{g,1}\left(\frac{lT}{L}, \frac{(l-1)T}{L}\right)\right)\vec{\psi}(0) - U(T,0)\vec{\psi}(0)\right\| \\ &\leq \widetilde{C}\frac{T^2}{L}\left(\sup_{t\in[0,T]}\sqrt{\|\vec{\psi}(0)\|\|H_1\vec{\psi}(t)\|} + \|\vec{\psi}(0)\|\right),\end{aligned} \tag{5.5.16}$$

$$\left\| \left(\prod_{l=1}^{L} U_{s,2}\left(\frac{lT}{L}, \frac{(l-1)T}{L}\right)\right)\vec{\psi}(0) - U(T,0)\vec{\psi}(0)\right\| \leq \widetilde{C}\frac{T^3}{L^2}\left(\sup_{t\in[0,T]}\|H_1\vec{\psi}(t)\| + \|\vec{\psi}(0)\|\right), \tag{5.5.17}$$

$$\left\| \left(\prod_{l=1}^{L} U_{g,2}\left(\frac{lT}{L}, \frac{(l-1)T}{L}\right)\right)\vec{\psi}(0) - U(T,0)\vec{\psi}(0)\right\| \leq \widetilde{C}\frac{T^3}{L^2}\left(\sup_{t\in[0,T]}\|H_1\vec{\psi}(t)\| + \|\vec{\psi}(0)\|\right). \tag{5.5.18}$$

Remark Although the error for the first-order generalized Trotter formula can also be bounded by $\sup_{t\in[0,T]}\|H_1\vec{\psi}(t)\| + \|\vec{\psi}(0)\|$ as the other schemes by a direct application of the Cauchy-Schwarz inequality, we keep it as $\sup_{t\in[0,T]}\sqrt{\|\vec{\psi}(0)\|\|H_1\vec{\psi}(t)\|} + \|\vec{\psi}(0)\|$ which promotes a better dependence on $\|H_1\vec{\psi}(t)\|$. This improvement is achievable only for the first-order generalized Trotter formula since its error only contains terms depending on $[H_1, H_2]$ which process a better bound as in Eq. (5.5.6), while the other schemes also contains terms depending on H_1 and/or the nested commutators $[H_1, [H_1, H_2]]$ and $[H_1, [H_2, H_1]]$.

Proof. Here we only present the proof for second-order generalized Trotter formula Eq. (5.5.18). The other three cases can be proved using the same approach.

For the second-order generalized Trotter formula, according to Theorem 45 and notice

that $\|\vec{\psi}(t)\| = \|\vec{\psi}(0)\|$ for all $t \in [0, T]$, we obtain

$$
\begin{aligned}
&\left\| \left(\prod_{l=1}^{L} U_{g,2}\left(\frac{lT}{L}, \frac{(l-1)T}{L}\right)\right) \vec{\psi}(0) - U(T,0)\vec{\psi}(0) \right\| \\
&= \left\| \left(\prod_{l=1}^{L} U_{g,2}\left(\frac{lT}{L}, \frac{(l-1)T}{L}\right)\right) \vec{\psi}(0) - \left(\prod_{l=1}^{L} U\left(\frac{lT}{L}, \frac{(l-1)T}{L}\right)\right) \vec{\psi}(0) \right\| \\
&\leq \sum_{k=1}^{L} \left\| \left(\prod_{l=k+1}^{L} U_{g,2}\left(\frac{lT}{L}, \frac{(l-1)T}{L}\right)\right) \left(\prod_{l=1}^{k} U\left(\frac{lT}{L}, \frac{(l-1)T}{L}\right)\right) \vec{\psi}(0) \right. \\
&\qquad \left. - \left(\prod_{l=k}^{L} U_{g,2}\left(\frac{lT}{L}, \frac{(l-1)T}{L}\right)\right) \left(\prod_{l=1}^{k-1} U\left(\frac{lT}{L}, \frac{(l-1)T}{L}\right)\right) \vec{\psi}(0) \right\| \\
&\leq \sum_{k=1}^{L} \left\| \left(\prod_{l=1}^{k} U\left(\frac{lT}{L}, \frac{(l-1)T}{L}\right)\right) \vec{\psi}(0) \right. \\
&\qquad \left. - U_{g,2}\left(\frac{kT}{L}, \frac{(k-1)T}{L}\right) \left(\prod_{l=1}^{k-1} U\left(\frac{lT}{L}, \frac{(l-1)T}{L}\right)\right) \vec{\psi}(0) \right\| \\
&= \sum_{k=1}^{L} \left\| \left(U\left(\frac{kT}{L}, \frac{(k-1)T}{L}\right) - U_{g,2}\left(\frac{kT}{L}, \frac{(k-1)T}{L}\right)\right) \vec{\psi}((k-1)T/L) \right\| \\
&\leq \widetilde{C}\frac{T^3}{L^3} \sum_{k=1}^{L} \left(\left\| H_1 \vec{\psi}((k-1)T/L) \right\| + \left\| \vec{\psi}((k-1)T/L) \right\| \right) \\
&\leq \widetilde{C}\frac{T^3}{L^2} \left(\sup_{t\in[0,T]} \left\| H_1 \vec{\psi}(t) \right\| + \left\| \vec{\psi}(0) \right\| \right). \qquad (5.5.19)
\end{aligned}
$$

□

Now we compare the error bounds in terms of operator norm (Theorem 42) with those in terms of vector norm (Theorem 46). We notice that the scalings with respect to T and L are the same for schemes of the same order, and the difference is in the dependence of the preconstants on H_1 and H_2. More precisely, the operator norm error bounds still depend on $\|H_1\|$ for standard Trotter formula and depend on norms of commutators like $\|[H_1, [H_1, H_2]]\|$ for generalized Trotter formula. On the other hand, the vector norm error bounds only depend on $\|H_2\|$, and the dependence on H_1 only appears in the form of $\|H_1\vec{\psi}\|$. Such a difference implies that the vector norm bounds can be much sharper than the operator norm bounds when $\|H_1\|$ is very large, but $\|H_1\vec{\psi}\|$ and $\|H_2\|$ are relatively small. We will show later that this is indeed the case for the model of interest in Eq. (5.1.3). Furthermore,

the difference in the error bounds can influence the scaling of total required Trotter steps with respect to the accuracy ϵ.

5.6 Application to Schrödinger equation with time-dependent effective mass and frequency

The model of the Schrödinger equation with a time-dependent effective mass and frequency in Eq. (5.1.3) has been studied in many works [49, 128, 127, 83, 56, 139]. Our goal is to study the complexity to obtain an ϵ-approximation of the wavefunction at time $T \sim \mathcal{O}(1)$, where $D = [0, 1]$ with periodic boundary conditions. Throughout the section we make the following assumptions.

1. $M_{\text{eff}}(t)$ is positive function, and is uniformly bounded from below.
2. $M_{\text{eff}}(t), \omega(t)$ are second-order continuously differentiable functions in t with uniformly bounded function and derivative values up to second order.
3. $V(x)$ is a fourth-order continuously differentiable function in x with bounded function and derivative values up to fourth order.

Here the fourth order derivative of $V(x)$ is required when estimating the errors of the second-order formulae in the operator norm. To be specific, it guarantees the nested commutator $[H_1, [H_1, H_2]]$ in its spatial discretization with n spatial grids to have an operator norm bounded by n^2 (instead of n^4). Without going into details of the discretization, which will be presented in the proofs, here we provide an intuition of this requirement on the continuous level – the presence of the fourth derivative of V in $[-\Delta, [-\Delta, V]]$ as in Eq. (5.5.3). Since the control functions and potential are bounded, throughout we will not track explicitly the dependence on them and absorb them into the preconstant denoted by $\widetilde{C}$ or the big-O notation $\mathcal{O}$ in our estimates.

We discretize the dynamics Eq. (5.1.3) as follows. First we perform spatial discretization using a central finite difference scheme with n equidistant nodes $x_k = k/n, 0 \leq k \leq n-1$. Then the semi-discretized dynamics becomes $\mathrm{i}\partial_t \vec{\psi}(t) = H(t)\vec{\psi}(t)$. Here the k-th entry of $\vec{\psi}(t)$ will be an approximation of the exact wavefunction evaluated at t and $x = (k-1)/n$. $H(t) = f_1(t)H_1 + f_2(t)H_2$ with

$$H_1 = n^2 \begin{pmatrix} 2 & -1 & & & -1 \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ -1 & & & -1 & 2 \end{pmatrix}, \tag{5.6.1}$$

and

$$H_2 = \mathrm{diag}(V(0), V(1/n), \cdots, V((n-1)/n)). \tag{5.6.2}$$

The standard or generalized Trotter formulae are used to discretize the dynamics in time with equidistant time steps and obtain numerical approximation of the wavefunction.

Furthermore, the H_1 and H_2 under central finite difference scheme using n equidistant nodes satisfy Assumption 43 with $H_1 = D_1^\dagger D_1$,

$$D_1 = n \begin{pmatrix} -1 & & & & 1 \\ 1 & -1 & & & \\ & \ddots & \ddots & & \\ & & 1 & -1 & \\ & & & 1 & -1 \end{pmatrix}, \tag{5.6.3}$$

and therefore the vector norm error bounds proved in Section 5.5 can be applied. This can be verified by straightforward but somewhat tedious matrix computations. We formally state the result in Lemma 47.

Lemma 47. *Consider H_1 and H_2 defined in Eqs.* (5.6.1) *and* (5.6.2)*, then Assumption 43 is satisfied under the rescaled 2-norm $\|\cdot\|_\star$ with D_1 defined in Eq.* (5.6.3)*.*

Proof. Let $V_k^{(0)} = V_k = V(x_k)$ for $0 \le k \le n-1$ and $V_{k+n} = V_k = V_{k-n}$ defined in a cyclic manner. Recursively we define $V_k^{(j+1)} = n(V_{k+1}^{(j)} - V_k^{(j)})$. Notice that $V_k^{(j)}$ is an approximation of the j-th order derivative of $V(x)$ evaluated at $x = x_k$. By Taylor's theorem and the assumption that $V(x)$ has bounded derivatives up to fourth order, we obtain that $V_k^{(j)}$ is bounded for any k and $0 \le j \le 4$. The equality $H_1 = D_1^\dagger D_1$ directly follows from the definition. We focus on the proof of the commutator bounds.

We start with the calculation of an explicit expression of $[H_1, H_2]$,

$$H_1 H_2 = n^2 \begin{pmatrix} 2V_0 & -V_1 & & & -V_{n-1} \\ -V_0 & 2V_1 & -V_2 & & \\ & \ddots & \ddots & \ddots & \\ & & -V_{n-3} & 2V_{n-2} & -V_{n-1} \\ -V_0 & & & -V_{n-2} & 2V_{n-1} \end{pmatrix}, \tag{5.6.4}$$

$$H_2 H_1 = n^2 \begin{pmatrix} 2V_0 & -V_0 & & & -V_0 \\ -V_1 & 2V_1 & -V_1 & & \\ & \ddots & \ddots & \ddots & \\ & & -V_{n-2} & 2V_{n-2} & -V_{n-2} \\ -V_{n-1} & & & -V_{n-1} & 2V_{n-1} \end{pmatrix}. \tag{5.6.5}$$

Then

$$[H_1, H_2] = n^2 \begin{pmatrix} 0 & V_0 - V_1 & & & V_0 - V_{n-1} \\ V_1 - V_0 & 0 & V_1 - V_2 & & \\ & \ddots & \ddots & \ddots & \\ & & V_{n-2} - V_{n-3} & 0 & V_{n-2} - V_{n-1} \\ V_{n-1} - V_0 & & & V_{n-1} - V_{n-2} & 0 \end{pmatrix}$$

$$= n \begin{pmatrix} 0 & -V_0^{(1)} & & & V_{n-1}^{(1)} \\ V_0^{(1)} & 0 & -V_1^{(1)} & & \\ & \ddots & \ddots & \ddots & \\ & & V_{n-3}^{(1)} & 0 & -V_{n-2}^{(1)} \\ -V_{n-1}^{(1)} & & & V_{n-2}^{(1)} & 0 \end{pmatrix}. \tag{5.6.6}$$

We further split $[H_1, H_2] = D_L + D_R + S$ where

$$D_L = n \begin{pmatrix} -V_{n-1}^{(1)} & & & & V_{n-1}^{(1)} \\ V_0^{(1)} & -V_0^{(1)} & & & \\ & \ddots & \ddots & & \\ & & V_{n-3}^{(1)} & -V_{n-3}^{(1)} & \\ & & & V_{n-2}^{(1)} & -V_{n-2}^{(1)} \end{pmatrix}, \tag{5.6.7}$$

$$D_R = n \begin{pmatrix} V_0^{(1)} & -V_0^{(1)} & & & \\ & V_1^{(1)} & -V_1^{(1)} & & \\ & & \ddots & \ddots & \\ & & & V_{n-2}^{(1)} & -V_{n-2}^{(1)} \\ -V_{n-1}^{(1)} & & & & V_{n-1}^{(1)} \end{pmatrix}. \tag{5.6.8}$$

and $S = -\mathrm{diag}(V_{n-1}^{(2)}, V_0^{(2)}, V_1^{(2)}, \cdots, V_{n-2}^{(2)})$. Notice that for any vector $\vec{v} = (v_k)_{k=0}^{n-1}$,

$$\begin{aligned} \|D_L \vec{v}\|_\star^2 &= n \sum_{k=0}^{n-1} |V_{k-1}^{(1)}|^2 |v_{k-1} - v_k|^2 \\ &\leq n \sup |V_k^{(1)}|^2 \sum_{k=0}^{n-1} |v_{k-1} - v_k|^2 = n \sup |V_k^{(1)}|^2 \left(\vec{v}^\dagger \frac{H_1}{n^2} \vec{v} \right) \\ &= \frac{1}{n} \sup |V_k^{(1)}|^2 \left(\vec{v}^\dagger H_1 \vec{v} \right) = \sup |V_k^{(1)}|^2 \|D_1 \vec{v}\|_\star^2, \end{aligned} \tag{5.6.9}$$

and similarly $\|D_R\vec{v}\|_\star^2 \le \sup|V_k^{(1)}|^2\|D_1\vec{v}\|_\star^2$. Furthermore we have $\|S\vec{v}\|_\star \le \sup|V_k^{(2)}|\|\vec{v}\|_\star$, thus there exists $\widetilde{C}$ such that

$$\|[H_1,H_2]\vec{v}\|_\star \le \|D_L\vec{v}\|_\star + \|D_R\vec{v}\|_\star + \|S\vec{v}\|_\star \le \widetilde{C}(\|D_1\vec{v}\|_\star + \|\vec{v}\|_\star). \tag{5.6.10}$$

To bound $\|[H_1,[H_1,H_2]]\vec{v}\|_\star$, we first compute $[H_1,[H_1,H_2]]$ and it gives

$$[H_1,[H_1,H_2]] = n^2\begin{pmatrix} -2V_{n-1}^{(2)} & & V_0^{(2)} & & \cdot & V_{n-2}^{(2)} & \\ & -2V_0^{(2)} & & V_1^{(2)} & & & V_{n-1}^{(2)} \\ V_0^{(2)} & & -2V_2^{(2)} & & & & \\ & V_1^{(2)} & & -2V_3^{(2)} & & & \\ & & \ddots & \ddots & \ddots & \ddots & V_{n-3}^{(2)} \\ V_{n-2}^{(2)} & & & & & -2V_{n-3}^{(2)} & \\ & V_{n-1}^{(2)} & & & V_{n-3}^{(2)} & & -2V_{n-2}^{(2)} \end{pmatrix}, \tag{5.6.11}$$

i.e. the only non-zero entries are

$$[H_1,[H_1,H_2]]_{k+1,k+1} = -2V_k^{(2)}, \quad [H_1,[H_1,H_2]]_{k,k+2} = [H_1,[H_1,H_2]]_{k+2,k} = V_k^{(2)}.$$

Then we split $[H_1,[H_1,H_2]] = H_L + H_R + 2H_C + 2D_{DL} + 2D_{DR} + W$ where

$$H_L = n^2\begin{pmatrix} V_{n-2}^{(2)} & & & V_{n-2}^{(2)} & -2V_{n-2}^{(2)} \\ -2V_{n-1}^{(2)} & V_{n-1}^{(2)} & & & V_{n-1}^{(2)} \\ & & \ddots & \ddots & \\ & & & \ddots & \\ V_{n-3}^{(2)} & & & -2V_{n-3}^{(2)} & V_{n-3}^{(2)} \end{pmatrix}, \tag{5.6.12}$$

$$H_R = n^2\begin{pmatrix} V_0^{(2)} & -2V_0^{(2)} & V_0^{(2)} & & \\ & V_1^{(2)} & -2V_1^{(2)} & V_1^{(2)} & \\ & & \ddots & \ddots & \\ & & & \ddots & \\ -2V_{n-1}^{(2)} & V_{n-1}^{(2)} & & & V_{n-1}^{(2)} \end{pmatrix}, \tag{5.6.13}$$

$$H_C = n^2\begin{pmatrix} -2V_{n-1}^{(2)} & V_{n-1}^{(2)} & & & V_{n-1}^{(2)} \\ V_0^{(2)} & -2V_0^{(2)} & V_0^{(2)} & & \\ & & \ddots & \ddots & \\ & & & \ddots & \\ V_{n-2}^{(2)} & & & V_{n-2}^{(2)} & -2V_{n-2}^{(2)} \end{pmatrix}, \tag{5.6.14}$$

$$D_{DL} = n \begin{pmatrix} V_{n-2}^{(3)} & & & & -V_{n-2}^{(3)} \\ -V_{n-1}^{(3)} & V_{n-1}^{(3)} & & & \\ & \ddots & \ddots & & \\ & & -V_{n-4}^{(3)} & V_{n-4}^{(3)} & \\ & & & -V_{n-3}^{(3)} & V_{n-3}^{(3)} \end{pmatrix}, \tag{5.6.15}$$

$$D_{DR} = n \begin{pmatrix} -V_{n-1}^{(3)} & V_{n-1}^{(3)} & & & \\ & -V_0^{(3)} & V_0^{(3)} & & \\ & & \ddots & \ddots & \\ & & & -V_{n-3}^{(3)} & V_{n-3}^{(3)} \\ V_{n-2}^{(3)} & & & & -V_{n-2}^{(3)} \end{pmatrix}, \tag{5.6.16}$$

and $W = n^2 \mathrm{diag}(V_k^{(2)} + V_{k-2}^{(2)} - 2V_{k-1}^{(2)})_{k=0}^{n-1}$. Notice that H_L, H_R, H_C are modifications from H_1, with the center of the central formula to be on the diagonal or subdiagonal and multiplying each row by different bounded parameters. D_{DL} and D_{DR} are very similar to D_L and D_R, with higher order potential. Furthermore, we have

$$\begin{aligned} \|H_L \vec{v}\|_\star^2 &= n^3 \sum_{k=0}^{n-1} |V_{k-2}^{(2)}|^2 |v_k - 2v_{k-1} + v_{k-2}|^2 \\ &\leq n^3 \sup |V_k^{(2)}|^2 \sum_{k=0}^{n-1} |v_k - 2v_{k-1} + v_{k-2}|^2 \\ &= \sup |V_k^{(2)}|^2 \|H_1 \vec{v}\|_\star^2, \end{aligned} \tag{5.6.17}$$

and similarly

$$\|H_R \vec{v}\|_\star \leq \widetilde{C} \|H_1 \vec{v}\|_\star, \quad \|H_C \vec{v}\|_\star \leq \widetilde{C} \|H_1 \vec{v}\|_\star.$$

For D_{DL} and D_{DR}, they can be bounded by the same way we bound D_L and D_R before, and thus

$$\|D_{DL} \vec{v}\|_\star \leq \widetilde{C} \|D_1 \vec{v}\|_\star \leq \widetilde{C} (\|H_1 \vec{v}\|_\star + \|\vec{v}\|_\star)$$

and

$$\|D_{DR} \vec{v}\|_\star \leq \widetilde{C} (\|H_1 \vec{v}\|_\star + \|\vec{v}\|_\star).$$

Finally since V has bounded fourth order derivative, the term $n^2(V_k^{(2)} + V_{k-2}^{(2)} - 2V_{k-1}^{(2)})$ is bounded. Therefore

$$\|W \vec{v}\|_\star \leq \widetilde{C} \|\vec{v}\|_\star.$$

Combining all the estimates together, we have

$$\begin{aligned}\|[H_1,[H_1,H_2]]\vec{v}\|_\star &\leq \|H_L\vec{v}\|_\star + \|H_R\vec{v}\|_\star + 2\|H_C\vec{v}\|_\star + 2\|D_{DL}\vec{v}\|_\star + 2\|D_{DR}\vec{v}\|_\star + \|W\vec{v}\|_\star \\ &\leq \widetilde{C}(\|H_1\vec{v}\|_\star + \|\vec{v}\|_\star). \end{aligned} \tag{5.6.18}$$

□

Lemma 47 can be directly used to provide the scaling of the Hamiltonians and their commutators in terms of n. Using the fact that the matrix 2-norms can be estimated by considering its 1-norm (the maximum absolute column sum) and ∞-norm (the maximum absolute row sum) by virtue of the fact that $\|M\|_2 \leq \sqrt{\|M\|_1\|M\|_\infty}$, we obtain the following Lemma. Notice that this result is consistent with the continuous picture that their continuous analogs $[-\Delta, V]$ and $[V,[-\Delta,V]]$ are differential operators of the first order and $[-\Delta,[-\Delta,V]]$ of second order.

Lemma 48. *Consider H_1, H_2 and D_1 defined in Eqs.* (5.6.1) *to* (5.6.3)*, then*

$$\|H_1\| = \mathcal{O}(n^2), \quad \|H_2\| = \mathcal{O}(1), \quad \|D_1\| = \mathcal{O}(n), \tag{5.6.19}$$

and

$$\|[H_1,H_2]\| = \mathcal{O}(n), \|[H_2,[H_1,H_2]]\| \leqslant 2\|[H_2\|\|[H_1,H_2]]\| = \mathcal{O}(n), \|[H_1,[H_1,H_2]]\| = \mathcal{O}(n^2). \tag{5.6.20}$$

Remark 49. *Although $\|H_1\|$ depends quadratically in n, the time-independent simulations $\exp(-\mathrm{i}H_1)$ and $\exp(-\mathrm{i}H_2)$ can still be performed efficiently. For H_1, it can be diagonalized under Fourier transform. Specifically, let $F = (\omega^{jk}/\sqrt{n})_{j,k=0}^{n-1}$ be the Fourier transform unitary matrix with $\omega = \exp(2\pi\mathrm{i}/n)$, then $H_1 = F\Lambda F^\dagger$ where $\Lambda = diag(2n^2(1-\cos(2\pi\mathrm{i}j/n)))_{j=0}^{n-1}$. Therefore $\exp(-\mathrm{i}H_1) = F\exp(-\mathrm{i}\Lambda)F^\dagger$ can be simulated efficiently, by first applying inverse quantum Fourier transform $F^\dagger$, then applying fast-forwarding techniques [40, 1] for $\exp(-\mathrm{i}\Lambda)$, and finally applying quantum Fourier transform F. For H_2, it can be implemented via either QSP technique [109] due to the boundedness of $\|H_2\|$, or fast-forwarding techniques since H_2 is a diagonal matrix as well. Due to the efficiency of simulating $\exp(-\mathrm{i}H_1)$ and $\exp(-\mathrm{i}H_2)$, it is reasonable to estimate the query complexity by counting the total number of Trotter steps.*

Now we are ready to analyze the errors. We measure the discretization errors using rescaled 2-norm, *i.e.* $\|\vec{\widetilde{\psi}}(t) - (\phi(t,k/n))_{k=0}^{n-1}\|_\star$, where $\vec{\widetilde{\psi}}(t)$ is the numerical solution after spatial and time discretization at time t, and $\phi(t,x)$ denotes the exact solution. As discussed before, the reason why we use rescaled 2-norm, rather than regular 2-norm, to measure

the error is because the exact solution $(\phi(t,k/n))_{k=0}^{n-1}$ is a discrete representation of the function ϕ, which is normalized under continuous L^2 norm rather than discrete 2-norm. Furthermore, if we encode the spatial discretized solution $\vec{\psi}$ into a quantum state, then the normalized condition requires $|\psi\rangle \sim \frac{1}{\sqrt{n}}\vec{\psi}$, thus $\| |\psi\rangle \| \sim \frac{1}{\sqrt{n}}\|\vec{\psi}\| = \|\vec{\psi}\|_\star$, that is, under correct normalization in each scenario, bounding regular 2-norm error for quantum states is equivalent to bounding rescaled 2-norm error for spatial discretized vectors. We remark that if we would like to control the relative error, then it does not matter whether the rescaled 2-norm or the 2-norm is used.

The errors are from two sources: spatial discretization of the Laplacian and potential operator, and the time discretization by Trotter formulae. We first bound the error from spatial discretization in Lemma 50.

Lemma 50. *Let the exact solution of the Schrödinger equation with Hamiltonian Eq.* (5.1.3) *be $\phi(t,x)$. Then*

1. *for any $0 \leqslant t \leqslant T, x \in [0,1]$,*

$$\begin{aligned} &|n^2(\phi(t,x+1/n) - 2\phi(t,x) + \phi(t,x-1/n)) - \Delta\phi(t,x)| \\ \leq &\frac{1}{3n^2} \sup_{y\in[0,1]} \left| \frac{\partial^4}{\partial x^4}\phi(t,y) \right|. \end{aligned} \tag{5.6.21}$$

2. *Let $\vec{\psi}(t)$ denote the solution of the dynamics*

$$\mathrm{i}\partial_t \vec{\psi}(t) = (f_1(t)H_1 + f_2(t)H_2)\vec{\psi}(t) \tag{5.6.22}$$

where H_1 and H_2 are the discretized Hamiltonian defined in Eq. (5.6.1) *and Eq.* (5.6.2)*, then for any $0 \leqslant t \in T$,*

$$\|(\phi(t,k/n))_{k=0}^{n-1} - \vec{\psi}(t)\|_\star \leq \frac{t}{3n^2}\|f_1\|_\infty \sup_{s\in[0,t],y\in[0,1]} \left| \frac{\partial^4}{\partial x^4}\phi(s,y) \right|. \tag{5.6.23}$$

Proof. 1. From Taylor's theorem,

$$\begin{aligned} &n^2(\phi(t,x+1/n) - 2\phi(t,x) + \phi(t,x-1/n)) - \Delta\phi(t,x) \\ = &\frac{n^2}{6}\left[\int_{x-1/n}^{x} (y-(x-1/n))^3 \frac{\partial^4}{\partial x^4}\phi(t,y)dy + \int_{x}^{x+1/n} (x+1/n-y)^3 \frac{\partial^4}{\partial x^4}\phi(t,y)dy \right]. \end{aligned} \tag{5.6.24}$$

Therefore

$$\begin{aligned}
&|n^2(\phi(t,x+1/n)-2\phi(t,x)+\phi(t,x-1/n))-\Delta\phi(t,x)| \\
&\le \frac{1}{6n}\left[\int_{x-1/n}^{x}\left|\frac{\partial^4}{\partial x^4}\phi(t,y)\right|dy+\int_{x}^{x+1/n}\left|\frac{\partial^4}{\partial x^4}\phi(t,y)\right|dy\right] \\
&\le \frac{1}{3n^2}\sup_{y\in[0,1]}\left|\frac{\partial^4}{\partial x^4}\phi(t,y)\right|. \qquad (5.6.25)
\end{aligned}$$

2. Since $\phi(t,x)$ satisfies the equation

$$\begin{aligned}
\mathrm{i}\partial_t\phi(t,x) &= H(t)\phi(t,x) \\
&= -f_1(t)n^2(\phi(t,x+1/n)-2\phi(t,x)+\phi(t,x-1/n)) \\
&\quad + f_2(t)V(x)\phi(t,x)+f_1(t)r(t,x) \qquad (5.6.26)
\end{aligned}$$

where $r(t,x)=n^2(\phi(t,x+1/n)-2\phi(t,x)+\phi(t,x-1/n))-\Delta\phi(t,x)$, the vector $\vec{\phi}(t)=(\phi(t,k/n))_{k=0}^{n-1}$ satisfies the ordinary differential equation

$$\mathrm{i}\partial_t\vec{\phi}(t)=(f_1(t)H_1+f_2(t)H_2)\vec{\phi}(t)+f_1(t)\vec{R}(t) \qquad (5.6.27)$$

where $\vec{R}(t)=(r(t,k/n))_{k=0}^{n-1}$. Same as our previous notations, let $U(t,s)$ denote the evolution operator from time s to t of the dynamics Eq. (5.1.1) with Hamiltonian Eq. (5.1.3). By the variation of parameters formula (Lemma 34),

$$\vec{\phi}(t)=\vec{\psi}(t)+\int_0^t U(t,s)f_1(s)\vec{R}(s)ds, \qquad (5.6.28)$$

and thus

$$\|\vec{\phi}(t)-\vec{\psi}(t)\|_\star \le t\|f_1\|_\infty \sup_{s\in[0,t]}\|\vec{R}(s)\|_\star. \qquad (5.6.29)$$

It remains to bound $\|\vec{R}(s)\|_\star$.

By the definition of $\vec{R}$ and the first part of this lemma, for any s,

$$\begin{aligned}
\|\vec{R}(s)\|_\star^2 &= \frac{1}{n}\sum_{k=0}^{n-1}|r(s,k/n)|^2 \\
&\le \frac{1}{n}\sum_{k=0}^{n-1}\left(\frac{1}{3n^2}\sup_{y\in[0,1]}\left|\frac{\partial^4}{\partial x^4}\phi(s,y)\right|\right)^2 \\
&\le \frac{1}{9n^4}\left(\sup_{y\in[0,1]}\left|\frac{\partial^4}{\partial x^4}\phi(s,y)\right|\right)^2. \qquad (5.6.30)
\end{aligned}$$

Plug this estimate back to Eq. (5.6.29), we complete the proof.

□

Lemma 50 shows that in order to make the vector error induced by the spatial discretization bounded by ϵ, it suffices to choose

$$n = \mathcal{O}\left(\frac{T^{1/2}}{\epsilon^{1/2}}\left(\sup\left|\frac{\partial^4\phi}{\partial x^4}\right|\right)^{1/2}\right). \tag{5.6.31}$$

The second source of the error is the time discretization using standard or generalized Trotter formula by applying the aforementioned Theorems for the errors in the operator and vector norm. The following lemma makes explicit the scaling in n.

Lemma 51. *Let $\vec{\psi}(t)$ be the solution of spatially discretized Schrödinger equation Eq. (5.1.3) using finite difference with n grid points, and $U(t,0)$ be the evolution operator. Let $\vec{\psi}_{s,p}(t)$ and $\vec{\psi}_{g,p}(t)$ be the corresponding numerical solution from p-th order standard and generalized Trotter formula with L equidistant time steps, respectively, and $U_{s,p}(t,0)$, $U_{g,p}(t,0)$ be the corresponding evolution operators. Assume that L is sufficiently large, then there exists a constant $\widetilde{C} > 0$, independent of $T, L, H_1, n, \vec{\psi}, \phi$, such that*

1.

$$\begin{aligned}
\|U_{s,1}(T,0) - U(T,0)\| &\leq \widetilde{C}\frac{n^2T^2}{L},\\
\|U_{g,1}(T,0) - U(T,0)\| &\leq \widetilde{C}\frac{nT^2}{L},\\
\|U_{s,2}(T,0) - U(T,0)\| &\leq \widetilde{C}\frac{n^2T^3}{L^2},\\
\|U_{g,2}(T,0) - U(T,0)\| &\leq \widetilde{C}\frac{n^2T^3}{L^2}.
\end{aligned} \tag{5.6.32}$$

2.

$$
\begin{aligned}
\|\vec{\psi}_{s,1}(T)-\vec{\psi}(T)\|_\star &\le \widetilde{C}\frac{T^2}{L}\left((T+1)\sup\left|\frac{\partial^4\phi}{\partial x^4}\right|+\sup\left|\frac{\partial^2\phi}{\partial x^2}\right|+\sup_{y\in[0,1]}|\phi(0,y)|\right),\\
\|\vec{\psi}_{g,1}(T)-\vec{\psi}(T)\|_\star &\le \widetilde{C}\frac{T^2}{L}\left[\left((T+1)\sup\left|\frac{\partial^4\phi}{\partial x^4}\right|+\sup\left|\frac{\partial^2\phi}{\partial x^2}\right|\right)^{1/2}\left(\sup_{y\in[0,1]}|\phi(0,y)|\right)^{1/2}\right]\\
&\quad+\widetilde{C}\frac{T^2}{L}\sup_{y\in[0,1]}|\phi(0,y)|,\\
\|\vec{\psi}_{s,2}(T)-\vec{\psi}(T)\|_\star &\le \widetilde{C}\frac{T^3}{L^2}\left((T+1)\sup\left|\frac{\partial^4\phi}{\partial x^4}\right|+\sup\left|\frac{\partial^2\phi}{\partial x^2}\right|+\sup_{y\in[0,1]}|\phi(0,y)|\right),\\
\|\vec{\psi}_{g,2}(T)-\vec{\psi}(T)\|_\star &\le \widetilde{C}\frac{T^3}{L^2}\left((T+1)\sup\left|\frac{\partial^4\phi}{\partial x^4}\right|+\sup\left|\frac{\partial^2\phi}{\partial x^2}\right|+\sup_{y\in[0,1]}|\phi(0,y)|\right),
\end{aligned}
\tag{5.6.33}
$$

where the notation sup *without any subscript should be interpreted as* $\sup_{t\in[0,T],x\in[0,1]}$.

Remark 52. *The condition that L should be sufficiently large is to ensure that the lowest order term in the error bounds are dominant and to allow us to discard the higher order terms. This can be guaranteed by requiring the desired level of error ϵ to be sufficiently small in the complexity estimate later.*

Proof. 1. The result follows by combining Theorem 42 and the scaling of the matrix norms provided in Lemma 48.

2. According to Theorem 46 and the fact that $\|\vec{\psi}(0)\|_\star \le \sup_{y\in[0,1]}|\psi(0,y)|$, we only need to bound $\|H_1\vec{\psi}(t)\|_\star$ for any $t\in[0,T]$. Let $r(t,x)=n^2(\phi(t,x+1/n)-2\phi(t,x)+\phi(t,x-1/n))-\Delta\phi(t,x)$ where $\phi(t,x)$ is the exact solution before any discretization. By Lemma 50,

$$
\begin{aligned}
\|H_1\vec{\psi}(t)\|_\star &\le \|H_1(\vec{\psi}(t)-(\phi(t,k/n))_{k=0}^{n-1})\|_\star+\|H_1(\phi(t,k/n))_{k=0}^{n-1}\|_\star\\
&\le \|H_1(\vec{\psi}(t)-(\phi(t,k/n))_{k=0}^{n-1}\|_\star+\|((\Delta\phi(t,k/n))_{k=0}^{n-1}\|_\star+\|(r(t,k/n))_{k=0}^{n-1}\|_\star\\
&\le \widetilde{C}\left(t\sup_{s\in[0,T],y\in[0,1]}\left|\frac{\partial^4}{\partial x^4}\phi(s,y)\right|+\sup_{y\in[0,1]}\left|\frac{\partial^2}{\partial x^2}\phi(t,y)\right|+\frac{1}{n^2}\sup_{s\in[0,T],y\in[0,1]}\left|\frac{\partial^4}{\partial x^4}\phi(s,y)\right|\right)\\
&\le \widetilde{C}\left((T+1)\sup_{s\in[0,T],y\in[0,1]}\left|\frac{\partial^4}{\partial x^4}\phi(s,y)\right|+\sup_{s\in[0,T],y\in[0,1]}\left|\frac{\partial^2}{\partial x^2}\phi(s,y)\right|\right)
\end{aligned}
\tag{5.6.34}
$$

□

Finally, we combine both spatial and temporal errors. It is not possible to obtain an operator norm error bound between the evolution operator of an unbounded operator and that of a finite dimensional matrix. Hence the operator norm bounds below are obtained by plugging in the estimate of n that achieves the vector norm error with precision ϵ. In particular, the operator norm error bound involves the derivatives of the exact solution of interest ϕ. Combining Eq. (5.6.31) and Lemma 51, we obtain the total complexity estimates.

Theorem 53. *We use central finite difference for spatial discretization and Trotter formulae for time discretization to obtain an ϵ-approximation in rescaled 2-norm of the solution $\phi(t,x)$. Let $L_{ope,s,1}$ and $L_{ope,g,1}$ denote the total number of required time steps of first-order standard and generalized Trotter estimated from operator norm error bounds, respectively, $L_{vec,s,1}$ and $L_{vec,g,1}$ denote the estimates from vector norm error bounds, and $L_{ope,s,2}$, $L_{ope,g,2}$, $L_{vec,s,2}$, $L_{vec,g,2}$ are the corresponding estimates for second-order schemes. Then for sufficiently small ϵ,*

$$
\begin{aligned}
L_{ope,s,1} &= \mathcal{O}\left(\frac{T^3}{\epsilon^2}\left(\sup\left|\frac{\partial^4\phi}{\partial x^4}\right|\right)\right),\\
L_{ope,g,1} &= \mathcal{O}\left(\frac{T^{5/2}}{\epsilon^{3/2}}\left(\sup\left|\frac{\partial^4\phi}{\partial x^4}\right|\right)^{1/2}\right),\\
L_{vec,s,1} &= \mathcal{O}\left(\frac{T^2}{\epsilon}\left((T+1)\sup\left|\frac{\partial^4\phi}{\partial x^4}\right| + \sup\left|\frac{\partial^2\phi}{\partial x^2}\right| + \sup_{y\in[0,1]}|\phi(0,y)|\right)\right),\\
&L_{vec,g,1}\\
&= \mathcal{O}\left(\frac{T^2}{\epsilon}\left[\left((T+1)\sup\left|\frac{\partial^4\phi}{\partial x^4}\right| + \sup\left|\frac{\partial^2\phi}{\partial x^2}\right|\right)^{1/2}\left(\sup_{y\in[0,1]}|\phi(0,y)|\right)^{1/2} + \sup_{y\in[0,1]}|\phi(0,y)|\right]\right),
\end{aligned}
$$

and

$$
\begin{aligned}
L_{ope,s,2} = L_{ope,g,2} &= \mathcal{O}\left(\frac{T^2}{\epsilon}\left(\sup\left|\frac{\partial^4\phi}{\partial x^4}\right|\right)^{1/2}\right),\\
L_{vec,s,2} = L_{vec,g,2} &= \mathcal{O}\left(\frac{T^{3/2}}{\epsilon^{1/2}}\left((T+1)\sup\left|\frac{\partial^4\phi}{\partial x^4}\right| + \sup\left|\frac{\partial^2\phi}{\partial x^2}\right| + \sup_{y\in[0,1]}|\phi(0,y)|\right)^{1/2}\right).
\end{aligned}
\tag{5.6.35}
$$

Proof. The complexity can be estimated by requiring both error bounds in Lemma 50 and Lemma 51 to be smaller than ϵ. First, by requiring the right hand side of Eq. (5.6.23) to be bounded by ϵ, the scaling of n should be that in Eq. (5.6.31). Plug this back into

Lemma 51 and also let the bounds in Lemma 51 to be bounded by ϵ, we obtain the complexity estimates. □

Theorem 53 clearly illustrates the advantage of vector norm error bounds in terms of the desired level of error ϵ. More precisely, the total number of required Trotter steps estimated from vector norm bounds only scales $\mathcal{O}(1/\epsilon^{1/p})$ for p-th order schemes. This is because the operator norm error bounds depend on the spatial discretization n, where $n = \mathcal{O}(1/\epsilon^{1/2})$ for second order spatial discretization, but the vector norm error bounds do not. We summarize our complexity estimates in terms of the spatial discretization as well as the error level ϵ in Table 5.2, where the simulation time T is $\mathcal{O}(1)$.

The best scaling is achieved by the second order standard and generalized Trotter formulae with the vector norm error bound, which is the result we are referring to as 'This work' in Table 5.1 for comparison with existing estimates. As discussed earlier, in order to demonstrate the behavior of the Trotter formulae for unbounded operators, we require n to grow as poly$(1/\epsilon)$. Therefore we choose $V(x)$ to be a C^4 function so that the commutator scaling of the second order Trotter formulae are valid.

Numerical tests in Section 5.7 demonstrate that the complexity estimates from vector norm error bounds are sharp in terms of ϵ for all the schemes we consider.

Remark 54 (*a priori* estimates of the solution ϕ)**.** *Due to the potential growth of the derivatives of the exact solution with respect to T,* a priori *estimates are required if we would like to obtain the overall scalings in T. Such* a priori *estimates, where the spatial derivatives are bounded by polynomials of T, have been established in the literature for various special cases, such as when $f_1 \equiv 1$, f_2 is smooth in t and V is a real potential, smooth in x and periodic in x as considered in [25], and for strictly positive f_1 in [118]. The corresponding estimates are usually technical, while the common approach to derive them is a combination of various analytical tools and a careful capture on the resonance in the dynamics. Detailed discussions are orthogonal to the topic here and are omitted.*

As we have already observed in Theorem 42, the generalized Trotter formula exhibits commutator type error bounds, while the standard Trotter formula does not. However, the commutator error bound only translates to improved asymptotic complexity for the first order generalized Trotter scheme. For second order schemes, there is no significant difference in the scaling with respect to ϵ between the standard and generalized Trotter formulae. As discussed before, this is due to the fact that $\|H_1\|$ and $\|[H_1, [H_1, H_2]]\|$ have the same asymptotic scaling in n. The generalized Trotter is less restrictive on the control functions, namely, the p-th order generalized Trotter formula ($p = 1, 2$) only requires the boundedness of the derivatives of control functions up to the $(p-1)$-th order while the p-th standard one requires the boundedness up to the p-th order. We expect the same situation for higher order Trotter formulae.

5.7 Numerical Results

To illustrate the difference between the operator norm and vector norm, we consider the following Hamiltonian

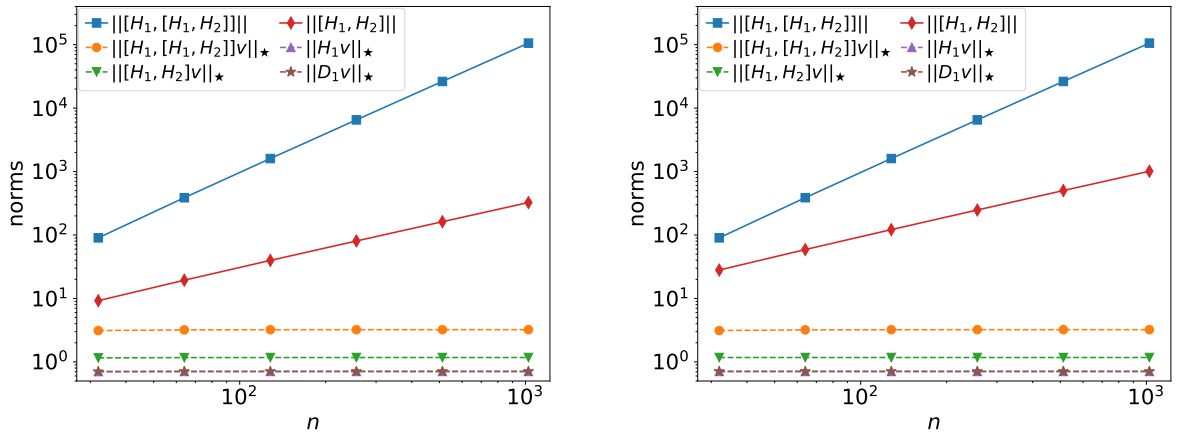
$$H(t) = -\frac{1}{2}(2+\sin(at+0.5))\Delta+(1+\cos(t))V(x), \quad V(x) = 1-\cos(x), \quad x \in [-\pi, \pi] \quad (5.7.1)$$

with periodic boundary conditions. Here $a > 0$ controls the magnitude of the derivatives $\|f_1'\|_\infty$ and $\|f_1''\|_\infty$. These sizes play a role in the preconstants of the errors as shown in Theorem 40. As discussed in Section 5.6, H_1 and H_2 correspond to the discretized matrices of $-\Delta$ and $V(x)$, respectively. Besides the second order finite difference scheme, we also demonstrate that our estimates are equally applicable to the Fourier discretization. Though in this particular example $V(x)$ is smooth, the scaling of n is still chosen according to Eq. (5.6.31), which only requires the regularity of V up to its fourth order derivatives and hence works for more general potentials.

We first demonstrate the scaling of the vector norms and the operator norms, respectively. Consider the vector $\vec{v}$ as the discretization of the smooth function $\cos(x)$. Fig. 5.7.1 plots the operator norms and the vector norms for various number of spatial grids n using the finite difference and Fourier spatial discretization. We find that $\|[H_1, [H_1, H_2]]\|$ grows quadratically with respect the the number of spatial grids while $\|[H_1, H_2]\|$ scales linearly, which agrees with Lemma 48. However, the vector norms $\|[H_1, [H_1, H_2]]\vec{v}\|_\star$, $\|[H_1, H_2]\vec{v}\|_\star$ remain of the same scale. This behavior is not restricted to the specific spatial discretization. Moreover, the scalings of $\|D_1\vec{v}\|_\star$ and $\|H_1\vec{v}\|_\star$ are found to be the same as those of $\|[H_1, [H_1, H_2]]\vec{v}\|_\star$ and $\|[H_1, H_2]\vec{v}\|_\star$. This verifies the in assumptions Eqs. (5.5.1) and (5.5.2), which is also proved for the finite difference scheme in Lemma 47.

We then verify the scaling of the errors with respect to n. The initial wavefunction is $\phi(x, 0) = \cos(x)$. The time step size h is fixed to be 10^{-4}. We run the Trotter formulae for 10 steps, which is sufficient for demonstrating the difference in scalings. The relative errors for both the operator and vector norms are plotted in Fig. 5.7.2 for $a = 1$ and $a = 10$. In terms of the operator norm, the generalized Trotter formula has a smaller error compared to the standard one: the relative error in the operator norms for the first-order standard Trotter scheme scales quadratically with respect to the number of grids while the first-order generalized Trotter schemes admits a linear scaling thanks to the commutator bounds. On the other hand, the relative errors in the vector norm do not grow with respect to n.

For second-order schemes, it can been seen that the errors measured by the operator norm for both methods grow quadratically with respect to n, while the corresponding errors in the vector norm are stable as n increases. These results agree with Lemma 51. Note that though the operator norm errors of the second-order schemes have the same asymptotic scaling in n, their preconstants may differ. When $a = 1$, the sizes of $\|f_1\|_\infty, \|f_1'\|_\infty, \|f_1''\|_\infty$



(a) Finite difference discretization

(b) Fourier discretization

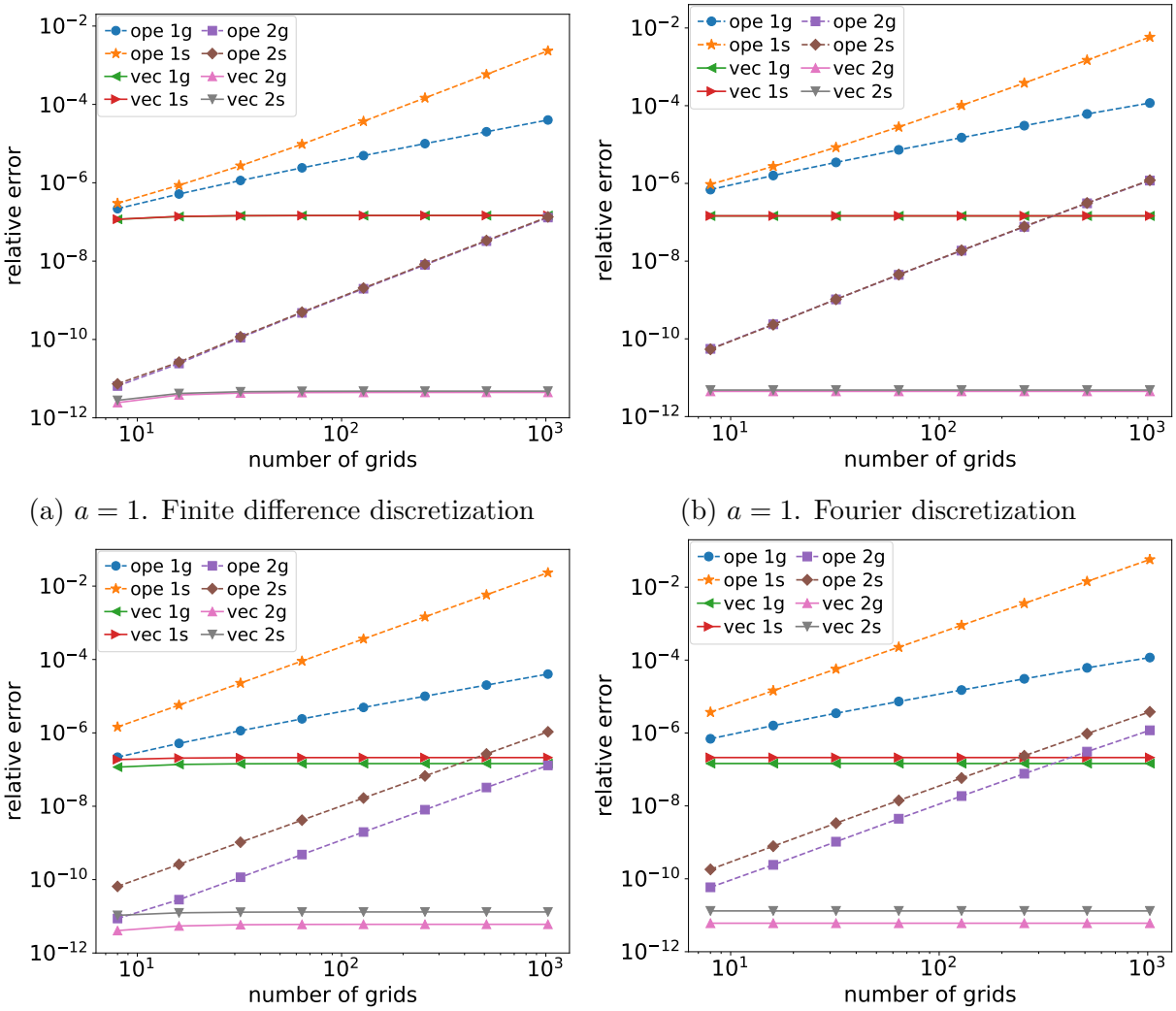
Figure 5.7.1: Operator and vector norms of a smooth vector for various numbers of spatial grids n. $\|[H_1,[H_1,H_2]]\|$ and $\|[H_1,H_2]\|$ scales quadratically and linearly with respect to n, but the vector norms do not grow as n gets larger.

are comparable, and there is no significant difference in the preconstants. However, when $a = 10$, $\|f_1''\|_\infty$ is one order of magnitude larger than $\|f_1'\|_\infty, \|f_1\|_\infty$. In this case, we find from Fig. 5.7.2 that the generalized Trotter formula has a smaller preconstant, which agrees with the preconstant estimates as described in Theorem 40.

Moreover, we compare the scaling of the number of Trotter steps for various precision ϵ, measuring the relative error via the vector norm. We fix $a = 10$, $T = 0.16$, and consider the precision ϵ as 2^{-10}, 2^{-12}, 2^{-14}, 2^{-16}, 2^{-18}, 2^{-20} and take $n \propto \epsilon^{-0.5}$ as 2^5, 2^6, 2^7, 2^8, 2^9 and 2^{10}. As is presented in Fig. 5.7.3, both second-order Trotter formulae requires the number of Trotter steps $L = \mathcal{O}(\epsilon^{-0.5})$ while it requires $L = \mathcal{O}(\epsilon^{-1})$ for both first-order Trotter formulae. These results agree with Theorem 53.

5.8 Conclusion

We have studied in detail the behavior of first and second order standard and generalized Trotter formulae for time-dependent Hamiltonian simulation with unbounded, control type Hamiltonians. We demonstrated that the error of the Hamiltonian simulation for a given initial state can often be overestimated using the standard analysis based on operator norms, which overestimates the computational cost. By taking into account the information of the



(a) $a = 1$. Finite difference discretization

(b) $a = 1$. Fourier discretization

(c) $a = 10$. Finite difference discretization

(d) $a = 10$. Fourier discretization

Figure 5.7.2: Relative Errors in the operator and vector norms. In the legend, "g" stands for the generalized Trotter formula and "s" for the standard Trotter formula. The error in operator norm is labeled as "ope" while the one in vector norm as "vec". First Row: $a = 1$ with slowly varying control functions. Second Row: $a = 10$ with fast varying control functions.

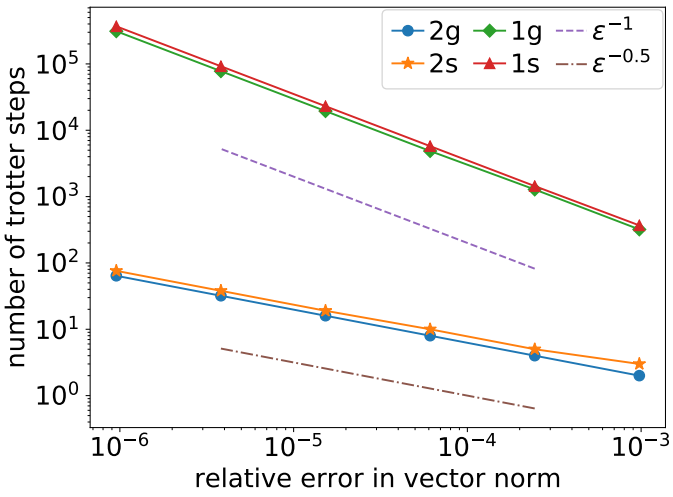


Figure 5.7.3: The number of Trotter steps required to achieve various precision for the relative error in the vector norm. The spatial discretization is finite difference. Both second-order Trotter formulae scales proportionally to $\epsilon^{-0.5}$ while the first-order formulae scales as ϵ^{-1}, which agrees with the theoretical bounds.

initial state in the error analysis, sharper error estimates can be derived via the vector norm scaling. As a side product, we also obtained improved error bounds of the standard and generalized Trotter formulae in operator norm as well in the time-dependent setting.

As an example, we applied our results to the time-dependent Schrödinger equation with a time-dependent effective mass and frequency. While the complexities of existing quantum algorithms for time-dependent Hamiltonian simulation scale at least linearly in the spatial discretization parameter n, we demonstrate that, the error bounds in vector norm do not suffer from such overheads (for both the standard and generalized Trotter formulae). Thus in this setting, our results outperform all existing quantum algorithms, including higher order Trotter and post-Trotter methods.

The bilinear form in Eq. (5.1.2) facilitates the discussion of the error of the Trotter formulae. For the most general Hamiltonian $H(t) = H_1(t) + H_2(t)$, it has been demonstrated that the error bound can be much more complicated even in the second order case [78]. Nevertheless, under suitable modifications, the main conclusion of this chapter can still be applicable to more general time-dependent Hamiltonian under further assumption that $\partial_t^k H_j(s)$ and $\partial_t^{k'} H_j(s')$ commute for any $j = 1, 2$ and k, k', s, s' (thus no essential difference is introduced in taking derivatives of unitaries and deriving error representation). This allows us to simulate e.g. Schrödinger equation governed by $H = -\Delta + V(x, t)$ where the mass is no

longer time-dependent but the potential $V(x,t)$ can have general anisotropy time dependency beyond control form.

A natural extension of this chapter is to consider general high order time-dependent standard and generalized Trotter formulae defined by Suzuki recursion [145, 155]. For the operator norm error bound, our results can be generalized to higher order case with a control Hamiltonian Eq. (5.1.2). More specifically, let $\mathcal{C}_k$ denote the set of the norms of all possible k-th order nested commutators of H_1 and H_2, for example $\mathcal{C}_0 = \{\|H_1\|, \|H_2\|\}$, $\mathcal{C}_1 = \{\|[H_1, H_2]\|\}$, and $\mathcal{C}_2 = \{\|[H_1, [H_1, H_2]]\|, \|[H_2, [H_2, H_1]]\|\}$. For p-th order schemes, we expect that the one-step operator norm error bounds for the standard and generalized Trotter formula scales as $\mathcal{O}(\alpha_{s,p} h^{p+1})$, $\mathcal{O}(\alpha_{g,p} h^{p+1})$, respectively. Here $\alpha_{s,p}$ is a linear combination of terms in the set $\bigcup_{k=0}^{p} \mathcal{C}_k$, while $\alpha_{g,p}$ is expressed as a linear combination of terms in the set $\bigcup_{k=1}^{p} \mathcal{C}_k$. Hence the difference lies in whether $\mathcal{C}_0$ is included, and generalized Trotter formula allows a commutator scaling. Notice that such an error bound will improve the best existing estimate [155], which depends on the norms of the Hamiltonians as well as their high order derivatives, and does not demonstrate possible commutator scalings.

The extension of our vector norm error bounds to p-th order time-dependent Trotter formula is also possible. The corresponding assumption on the bounds of commutators (*i.e.* counterpart of Assumption 43 in this work) becomes

$$\|[\underbrace{H_1, [H_1, \cdots, [H_1}_{k \text{ repeats}}, H_2]] \cdots]\vec{v}\| \leq \mathcal{O}\left(\|H_1^{k/2}\vec{v}\| + \|\vec{v}\|\right) \tag{5.8.1}$$

for any $1 \leq k \leq p$. Compared with the operator norm error bounds, for the Schrödinger equation with a time-dependent mass and frequency, such vector norm error bounds can still remove the dependence on the spatial discretization thus provide speedup in terms of the accuracy. However, the significance of such improvement might be subtle: in order to satisfy the assumption in Eq. (5.8.1), the potential function $V(x)$ needs to be much smoother with bounded higher order derivatives. Hence, the dependence of n on the error ϵ may become much weaker by employing higher order discretization schemes. In such a scenario, the spectral norms $\|H_1\|$ and $\|H_2\|$ may even become comparable, and the Hamiltonian $H(t)$ may not be regarded as an unbounded operator after all.

In this chapter, we mainly focus on the improvement brought by vector norm error bounds in terms of the accuracy. It is also interesting to study whether vector norm error bounds can improve the scalings of other parameters. For example, if the Schrödinger equation is in d dimension rather than one dimension considered in this chapter, then a vector norm error bound may offer speedup in terms of d, since the degree of freedom for spatial discretization can scale linearly in d [95]. Another related topic is the scaling with respect to the number of the particles in quantum many-body systems. Recently [143] obtained an improved estimate in terms of the number of electrons for electronic structure problem with plane-wave basis

functions in a second quantized formulation, by combining sparsity, commutator scalings and initial-state knowledge and bounding the operator norm on an η-electron sub-manifold. Although much smaller than that on the entire space, the operator norm on the η-electron sub-manifold may still overestimate the error, and a vector norm error bound might offer further improvement by taking the smoothness and low-energy property of the wavefunction into consideration. It is also an interesting question to investigate whether a vector norm error bound can provide any benefit for other applications such as spin systems.

This chapter suggests that it may be of interest to explore the gap between the operator norm and vector norm error bounds in other schemes for Hamiltonian simulations with unbounded operators. Note that such a gap may not exist in all methods. For instance, for time-independent Hamiltonian simulation, the quantum signal processing (QSP) method [109] is based on the polynomial approximation to the function $\cos(xt)$ and $\sin(xt)$, and we do not expect that the error bound can be significantly improved by considering vector norms. However, it may be possible to prove vector norm error bounds for other post-Trotter methods.

Chapter 6

Quantum linear system solver based on time-optimal adiabatic quantum computing and quantum approximate optimization algorithm

6.1 Introduction

Linear system solvers are used ubiquitously in scientific computing. Quantum algorithms for solving large systems of linear equations, also called the quantum linear system problem (QLSP), have received much attention recently [73, 43, 37, 64, 144, 157, 33, 158, 26]. The goal of QLSP is to efficiently compute $|x\rangle = A^{-1}|b\rangle / \|A^{-1}|b\rangle\|_2$ on a quantum computer, where $A \in \mathbb{C}^{N\times N}$, and $|b\rangle \in \mathbb{C}^N$ is a normalized vector (for simplicity we assume $N = 2^n$, and $\|A\|_2 = 1$). The ground-breaking Harrow, Hassidim, and Lloyd (HHL) algorithm obtains $|x\rangle$ with cost $\mathcal{O}(\text{poly}(n)\kappa^2/\epsilon)$, where κ is the condition number of A, and ϵ is the target accuracy. On the other hand, the best classical iterative algorithm is achieved by the conjugate gradient method, where the cost is at least $\mathcal{O}(N\sqrt{\kappa}\log(1/\epsilon))$, with the additional assumptions that A should be Hermitian positive definite and a matrix-vector product can be done with $\mathcal{O}(N)$ cost [135]. The complexity of direct methods based on the Gaussian elimination procedure removes the dependence on κ, but the dependence on N is typically super-linear even for sparse matrices [107]. Therefore the HHL algorithm can be exponentially faster than classical algorithms with respect to N. The undesirable dependence with respect to ϵ is due to the usage of the quantum phase estimation (QPE) algorithm. Recent progresses based on linear combination of unitaries (LCU) [43] and quantum signal processing (QSP) [109, 64] have further improved the scaling to $\mathcal{O}(\kappa^2\text{poly}(\log(\kappa/\epsilon)))$ under different query models, without

using QPE. However, the $\mathcal{O}(\kappa^2)$ scaling can be rather intrinsic to the methods, at least before complex techniques such as variable time amplitude amplification (VTAA) algorithm [3] are applied.

The VTAA algorithm is a generalization of the conventional amplitude amplification algorithm, and allows to quadratically amplify the success probability of quantum algorithms in which different branches stop at different time. In [3], VTAA was first used to successfully improve the complexity of HHL algorithm to $\tilde{\mathcal{O}}(\kappa/\epsilon)$. In [43], the authors further combine VTAA algorithm and a low-precision phase estimate to improve the complexity of LCU to $\tilde{\mathcal{O}}(\kappa\text{poly}(\log(\kappa/\epsilon)))$, which is near-optimal with respect to both κ and ϵ. It is worth noting that the VTAA algorithm is a complicated procedure and is considerably difficult to implement. Thus it remains of great interest to obtain alternative algorithms to solve QLSP with near-optimal complexity scaling without resorting to VTAA.

Some of the alternative routes for solving QLSP are provided by the adiabatic quantum computing (AQC) [82, 2] and a closely related method called the randomization method (RM) [21, 144]. The key idea of both AQC and RM is to solve QLSP as an *eigenvalue* problem with respect to a transformed matrix. Assume that a Hamiltonian simulation can be efficiently performed on a quantum computer, it is shown that the runtime of RM scales as $\mathcal{O}(\kappa \log(\kappa)/\epsilon)$ [144], which achieves near-optimal complexity with respect to κ without using VTAA algorithm as a subroutine. The key idea of the RM is to approximately follow the adiabatic path based on the quantum Zeno effect (QZE) using a Monte Carlo method. Although RM is inspired by AQC, the runtime complexity of the (vanilla) AQC is at least $\mathcal{O}(\kappa^2/\epsilon)$ [144, 20, 2]. Therefore the RM is found to be at least quadratically faster than AQC with respect to κ.

In this chapter, we find that with a simple modification of the scheduling function to traverse the adiabatic path, the gap between AQC and RM can be fully closed, along the following two aspects. 1) We propose a family of rescheduled AQC algorithms called AQC(p). Assuming κ (or its upper bound) is known, we demonstrate that for any matrix A (possibly non-Hermitian or dense), when $1 < p < 2$, the runtime complexity of AQC(p) can be only $\mathcal{O}(\kappa/\epsilon)$. Thus AQC(p) removes a logarithmic factor with respect to κ when compared to RM. 2) We propose another rescheduled algorithm called AQC(exp), of which the runtime is $\mathcal{O}(\kappa\ \text{poly}(\log(\kappa/\epsilon)))$. The main benefit of AQC(exp) is the improved dependence with respect to the accuracy ϵ, and this is the near optimal complexity (up to logarithmic factors) with respect to both κ and ϵ. The scheduling function of AQC(exp) is also universal in the sense that we do not even need the knowledge of an upper bound of κ. Existing works along this line [121, 62] only suggest that runtime complexity is $\mathcal{O}(\kappa^3\ \text{poly}(\log(\kappa/\epsilon)))$, which improves the dependence with respect to ϵ at the expense of a much weaker dependence on κ. Our main technical contribution is to again improve the dependence on κ. Since a generic QLSP solver with cost less than $\mathcal{O}(\kappa)$ does not exist [73], our result achieves the near-optimal complexity up to logarithmic factors.

The quantum approximate optimization algorithm (QAOA) [55], as a quantum variational algorithm (QVA), has received much attention recently thanks to the feasibility of being implemented on near-term quantum devices. Due to the natural connection between AQC and QAOA, our result immediately suggests that the time-complexity for solving QLSP with QAOA is also at most $\mathcal{O}(\kappa \operatorname{poly}(\log(\kappa/\epsilon)))$, which is also confirmed by numerical results. We also remark that both QAOA and AQC schemes prepares an approximate solution to the QLSP in the form of a pure state, while RM prepares a mixed state. All methods above can be efficiently implemented on gate-based computers, and are much simpler compared with those that use VTAA algorithm as a subroutine.

The rest of this chapter is organized as follows. Section 6.2 defines the quantum linear system problem. We first focus on the Hermitian positive definite case, and discuss how AQC with a linear interpolation function can solve this problem as well as its complexity in Section 6.3. Two improved AQC-based algorithms are proposed in Section 6.4 and Section 6.5, followed by a discuss on their gate-based implementation in Section 6.6. In Section 6.7, the approach of using QAOA to solve the quantum linear system problem is discussed. We then generalize all of our new methods to the non-Hermitian matrices case in Section 6.8. Numerical results are given in Section 6.9, which verify the effectiveness and theoretical scalings of our methods. Finally, full proofs of all the theorems in this chapter are provided in Section 6.10 and Section 6.11 with thorough technical details.

6.2 Quantum Linear System Problem

Assume $A \in \mathbb{C}^{N\times N}$ is an invertible matrix with condition number κ and $\|A\|_2 = 1$. Let $|b\rangle \in \mathbb{C}^N$ be a normalized vector. Given a target error ϵ, the goal of QLSP is to prepare a normalized state $|x_{\mathrm{a}}\rangle$, which is an ϵ-approximation of the normalized solution of the linear system $|x\rangle = A^{-1}|b\rangle / \|A^{-1}|b\rangle\|_2$, in the sense that $\| |x_{\mathrm{a}}\rangle\langle x_{\mathrm{a}}| - |x\rangle\langle x| \|_2 \leqslant \epsilon$.

For simplicity, we first assume A is Hermitian and positive definite and will discuss the generalization to non-Hermitian case later.

6.3 Vanilla AQC

Let $Q_b = I_N - |b\rangle\langle b|$. We introduce

$$H_0 = \sigma_x \otimes Q_b = \begin{pmatrix} 0 & Q_b \\ Q_b & 0 \end{pmatrix},$$

then H_0 is a Hermitian matrix and the null space of H_0 is $\mathrm{Null}(H_0) = \mathrm{span}\{|\widetilde{b}\rangle, |\bar{b}\rangle\}$. Here $|\tilde{b}\rangle = |0, b\rangle := (b, 0)^\top, |\bar{b}\rangle = |1, b\rangle := (0, b)^\top$. The dimension of H_0 is $2N$ and one ancilla qubit

is needed to enlarge the matrix block. We also define

$$H_1 = \sigma_+ \otimes (AQ_b) + \sigma_- \otimes (Q_b A) = \begin{pmatrix} 0 & AQ_b \\ Q_b A & 0 \end{pmatrix}.$$

Here $\sigma_\pm = \frac{1}{2}(\sigma_x \pm \mathrm{i}\sigma_y)$. Note that if $|x\rangle$ satisfies $A|x\rangle \propto |b\rangle$, we have $Q_b A|x\rangle = Q_b|b\rangle = 0$. Then $\mathrm{Null}(H_1) = \mathrm{span}\{|\widetilde{x}\rangle, |\bar{b}\rangle\}$ with $|\widetilde{x}\rangle = |0, x\rangle$. Since Q_b is a projection operator, the gap between 0 and the rest of the eigenvalues of H_0 is 1. The gap between 0 and the rest of the eigenvalues of H_1 is bounded from below by $1/\kappa$ (see supplemental materials).

QLSP can be solved if we can prepare the zero-energy state $|\widetilde{x}\rangle$ of H_1, which can be achieved by AQC approach. Let $H(f(s)) = (1 - f(s))H_0 + f(s)H_1, 0 \leqslant s \leqslant 1$. The function $f : [0,1] \to [0,1]$ is called a scheduling function, and is a strictly increasing mapping with $f(0) = 0, f(1) = 1$. The simplest choice is $f(s) = s$, which gives the "vanilla AQC". We sometimes omit the s-dependence as $H(f)$ to emphasize the dependence on f. Note that for any s, $|\bar{b}\rangle$ is always in $\mathrm{Null}(H(f(s)))$, and there exists a state $|\widetilde{x}(s)\rangle = |0, x(s)\rangle$, such that $\mathrm{Null}(H(f(s))) = \{|\widetilde{x}(s)\rangle, |\bar{b}\rangle\}$. In particular, $|\widetilde{x}(0)\rangle = |\widetilde{b}\rangle$ and $|\widetilde{x}(1)\rangle = |\widetilde{x}\rangle$, and therefore $|\widetilde{x}(s)\rangle$ is the desired adiabatic path. Let $P_0(s)$ be the projection to the subspace $\mathrm{Null}(H(f(s)))$, which is a rank-2 projection operator $P_0(s) = |\widetilde{x}(s)\rangle\langle\widetilde{x}(s)| + |\bar{b}\rangle\langle\bar{b}|$. Furthermore, the eigenvalue 0 is separated from the rest of the eigenvalues of $H(f(s))$ by a gap $\Delta(f(s)) \geqslant \Delta_*(f(s)) = 1 - f(s) + f(s)/\kappa$.

Gap of $H(f(s))$

The Hamiltonian $H(f)$ can be written in the block matrix form as

$$H(f) = \begin{pmatrix} 0 & ((1-f)I + fA)Q_b \\ Q_b((1-f)I + fA) & 0 \end{pmatrix}. \tag{6.3.1}$$

Let λ be an eigenvalue of H, then

$$\begin{aligned} 0 &= \det \begin{pmatrix} \lambda I & -((1-f)I + fA)Q_b \\ -Q_b((1-f)I + fA) & \lambda I \end{pmatrix} \\ &= \det\left(\lambda^2 I - ((1-f)I + fA)Q_b^2((1-f)I + fA)\right) \end{aligned}$$

where the second equality holds because the bottom two blocks are commutable. Thus λ^2 is an eigenvalue of $((1-f)I + fA)Q_b^2((1-f)I + fA)$, and $\Delta^2(f)$ equals the smallest non-zero eigenvalue of $((1-f)I + fA)Q_b^2((1-f)I + fA)$. Applying a proposition of matrices that XY and YX have the same non-zero eigenvalues, $\Delta^2(f)$ also equals the smallest non-zero eigenvalue of $Q_b((1-f)I + fA)^2 Q_b$.

Now we focus on the matrix $Q_b((1-f)I+fA)^2Q_b$. Note that $|b\rangle$ is the unique eigenstate corresponding to the eigenvalue 0, all eigenstates corresponding to non-zero eigenvalues must be orthogonal with $|b\rangle$. Therefore

$$\begin{aligned}\Delta^2(f) &= \inf_{\langle b|\varphi\rangle=0,\langle\varphi|\varphi\rangle=1} \left\langle \varphi \left| Q_b((1-f)I+fA)^2Q_b \right| \varphi \right\rangle \\ &= \inf_{\langle b|\varphi\rangle=0,\langle\varphi|\varphi\rangle=1} \left\langle \varphi \left| ((1-f)I+fA)^2 \right| \varphi \right\rangle \\ &\geq \inf_{\langle\varphi|\varphi\rangle=1} \left\langle \varphi \left| ((1-f)I+fA)^2 \right| \varphi \right\rangle \\ &= (1-f+f/\kappa)^2,\end{aligned}$$

and $\Delta(f) \geq \Delta_*(f) = 1-f+f/\kappa$.

Adiabatic theorem

Consider the adiabatic evolution

$$\frac{1}{T}\mathrm{i}\partial_s |\psi_T(s)\rangle = H(f(s)) |\psi_T(s)\rangle, \quad |\psi_T(0)\rangle = |\widetilde{b}\rangle, \tag{6.3.2}$$

where $0 \leq s \leq 1$, and the parameter T is called the runtime of AQC. The quantum adiabatic theorem [82, Theorem 3] states that for any $0 \leqslant s \leqslant 1$,

$$|1 - \langle\psi_T(s)|P_0(s)|\psi_T(s)\rangle| \leqslant \eta^2(s), \tag{6.3.3}$$

where

$$\eta(s) = C\Big\{ \frac{\|H^{(1)}(0)\|_2}{T\Delta^2(0)} + \frac{\|H^{(1)}(s)\|_2}{T\Delta^2(f(s))} + \frac{1}{T}\int_0^s \left(\frac{\|H^{(2)}(s')\|_2}{\Delta^2(f(s'))} + \frac{\|H^{(1)}(s')\|_2^2}{\Delta^3(f(s'))} \right) ds' \Big\}. \tag{6.3.4}$$

The derivatives of H are taken with respect to s, *i.e.* $H^{(k)}(s) := \frac{d^k}{ds^k}H(f(s)), k = 1, 2$. Here and throughout the chapter we shall use a generic symbol C to denote constants independent of s, Δ, T.

Intuitively, the quantum adiabatic theorem in Eq. (6.3.3) says that, if the initial state is an eigenstate corresponding to the eigenvalue 0, then for large enough T the state $|\psi_T(s)\rangle$ will almost stay in the eigenspace of $H(s)$ corresponding to the eigenvalue 0, where there is a double degeneracy and only one of the eigenstate $|\widetilde{x}(s)\rangle$ is on the desired adiabatic path. However, such degeneracy will not break the effectiveness of AQC for the following reason. Note that $\langle\bar{b}|\psi_T(0)\rangle = 0$, and $H(f(s))|\bar{b}\rangle = 0$ for all $0 \leqslant s \leqslant 1$, so the Schrödinger dynamics (6.3.2) implies $\langle\bar{b}|\psi_T(s)\rangle = 0$, which prevents any transition of $|\psi_T(s)\rangle$

to $|\bar{b}\rangle$. Therefore the adiabatic path will stay along $|\widetilde{x}(s)\rangle$. Using $\langle\bar{b}|\psi_T(s)\rangle = 0$, we have $P_0(s)\,|\psi_T(s)\rangle = |\widetilde{x}(s)\rangle\,\langle\widetilde{x}(s)|\psi_T(s)\rangle$. Therefore the estimate (6.3.3) becomes

$$1 - |\,\langle\psi_T(s)|\widetilde{x}(s)\rangle\,|^2 \leqslant \eta^2(s).$$

This also implies that, according to Lemma 55 to be stated and proves at the end of this section,

$$\||\psi_T(s)\rangle\,\langle\psi_T(s)| - |\widetilde{x}(s)\rangle\,\langle\widetilde{x}(s)|\|_2 \leqslant \eta(s).$$

Therefore $\eta(1)$ can be an upper bound of the distance of the density matrix. If we simply assume $\|H^{(1)}\|_2, \|H^{(2)}\|_2$ are constants, and use the worst case bound that $\Delta \geqslant \kappa^{-1}$, we arrive at the conclusion that in order to have $\eta(1) \leqslant \epsilon$, the runtime of vanilla AQC is $T \gtrsim \kappa^3/\epsilon$.

Lemma 55. *(i) The following equation holds,*

$$|1-\langle\psi_T(s)|P_0(s)|\psi_T(s)\rangle\,| = 1-|\langle\psi_T(s)|\widetilde{x}(s)\rangle|^2 = \|\,|\psi_T(s)\rangle\,\langle\psi_T(s)| - |\widetilde{x}(s)\rangle\,\langle\widetilde{x}(s)|\,\|_2^2. \quad (6.3.5)$$

(ii) Assume that

$$|1 - \langle\psi_T(s)|P_0(s)|\psi_T(s)\rangle\,| \leqslant \eta^2(s).$$

Then the fidelity can be bounded from below by $1-\eta^2(s)$, and the 2-norm error of the density matrix can be bounded from above by $\eta(s)$.

Proof. It suffices only to prove part (i). Note that $|\bar{b}\rangle$ is the eigenstate for both H_0 and H_1 corresponding the 0 eigenvalue, we have $H(f(s))\,|\bar{b}\rangle = ((1-f(s))H_0 + f(s)H_1)\,|\bar{b}\rangle = 0$, and thus $\frac{d}{ds}\,\langle\bar{b}|\psi_T(s)\rangle = 0$. Together with the initial condition $\langle\bar{b}|\psi_T(0)\rangle = 0$, the overlap of $|\bar{b}\rangle$ and $|\psi_T(s)\rangle$ remains to be 0 for the whole time period, *i.e.* $\langle\bar{b}|\psi_T(s)\rangle = 0$. Since $P_0(s) = |\widetilde{x}(s)\rangle\,\langle\widetilde{x}(s)| + |\bar{b}\rangle\,\langle\bar{b}|$, we have $P_0(s)\,|\psi_T(s)\rangle = |\widetilde{x}(s)\rangle\,\langle\widetilde{x}(s)|\psi_T(s))\rangle$. Therefore

$$|1 - \langle\psi_T(s)|P_0(s)|\psi_T(s)\rangle\,| = |1 - \langle\psi_T(s)|\widetilde{x}(s)\rangle\,\langle\widetilde{x}(s)|\psi_T(s)\rangle\,| = 1 - |\langle\psi_T(s)|\widetilde{x}(s)\rangle|^2\,.$$

To prove the second equation, let $M = |\psi_T(s)\rangle\,\langle\psi_T(s)| - |\widetilde{x}(s)\rangle\,\langle\widetilde{x}(s)|$. Note that $\|M\|_2^2 = \lambda_{\max}(M^\dagger M)$, we study the eigenvalues of $M^\dagger M$ by first computing that

$$\begin{aligned} M^\dagger M = {} & |\psi_T(s)\rangle\,\langle\psi_T(s)| + |\widetilde{x}(s)\rangle\,\langle\widetilde{x}(s)| \\ & - \langle\psi_T(s)|\widetilde{x}(s)\rangle\,|\psi_T(s)\rangle\,\langle\widetilde{x}(s)| - \langle\widetilde{x}(s)|\psi_T(s)\rangle\,|\widetilde{x}(s)\rangle\,\langle\psi_T(s)|\,. \end{aligned}$$

Since for any $|y\rangle \in \mathrm{span}\{|\psi_T(s)\rangle\,, |\widetilde{x}(s)\rangle\}^\perp$, $M^\dagger M\,|y\rangle = 0$, and

$$\begin{aligned} M^\dagger M\,|\psi_T(s)\rangle &= (1 - |\langle\psi_T(s)|\widetilde{x}(s)\rangle|^2)\,|\psi_T(s)\rangle\,, \\ M^\dagger M\,|\widetilde{x}(s)\rangle &= (1 - |\langle\psi_T(s)|\widetilde{x}(s)\rangle|^2)\,|\widetilde{x}(s)\rangle\,, \end{aligned}$$

we have $\|M\|_2^2 = \lambda_{\max}(M^\dagger M) = 1 - |\langle\psi_T(s)|\widetilde{x}(s)\rangle|^2$.

□

6.4 AQC(p) method

Our goal is to reduce the runtime by choosing a proper scheduling function. The key observation is that the accuracy of AQC depends not only on the gap $\Delta(f(s))$ but also on the derivatives of $H(f(s))$, as revealed in the estimate (6.3.4). Therefore it is possible to improve the accuracy if a proper time schedule allows the Hamiltonian $H(f(s))$ to slow down when the gap is close to 0. We consider the following schedule [82, 2]

$$\dot{f}(s) = c_p \Delta_*^p(f(s)), \quad f(0) = 0, \quad p > 0. \tag{6.4.1}$$

Here $c_p = \int_0^1 \Delta_*^{-p}(u)du$ is a normalization constant chosen so that $f(1) = 1$. When $1 < p \leq 2$, Eq. (6.4.1) can be explicitly solved as

$$f(s) = \frac{\kappa}{\kappa - 1}\left[1 - \left(1 + s(\kappa^{p-1} - 1)\right)^{\frac{1}{1-p}}\right]. \tag{6.4.2}$$

Note that as $s \to 1$, $\Delta_*(f(s)) \to \kappa^{-1}$, and therefore the dynamics of $f(s)$ slows down as $f \to 1$ and the gap decreases towards κ^{-1}. We refer to the adiabatic dynamics (6.3.2) with the schedule (6.4.1) as the AQC(p) scheme. Our main result is given in Theorem 56 (See Section 6.10 for the proof).

Theorem 56. *Let $A \in \mathbb{C}^{N\times N}$ be a Hermitian positive definite matrix with condition number κ. For any choice of $1 < p < 2$, the error of the AQC(p) scheme satisfies*

$$\| \, |\psi_T(1)\rangle \langle \psi_T(1)| - |\widetilde{x}\rangle \langle \widetilde{x}| \, \|_2 \leq C\kappa/T. \tag{6.4.3}$$

Therefore in order to prepare an $\epsilon-$approximation of the solution of QLSP it suffices to choose the runtime $T = \mathcal{O}(\kappa/\epsilon)$. Furthermore, when $p = 1, 2$, the bound for the runtime becomes $T = \mathcal{O}(\kappa \log(\kappa)/\epsilon)$.

The runtime complexity of the AQC(p) method with respect to κ is only $\mathcal{O}(\kappa)$. Compared to Ref. [144], AQC(p) further removed the $\log(\kappa)$ dependence when $1 < p < 2$, and hence reaches the optimal complexity with respect to κ. Interestingly, though not explicitly mentioned in [144], the success of RM for solving QLSP relies on a proper choice of the scheduling function, which approximately corresponds to AQC(p=1). It is this scheduling function, rather than the QZE or its Monte Carlo approximation *per se* that achieves the desired $\mathcal{O}(\kappa \log \kappa)$ scaling with respect to κ. Furthermore, the scheduling function in RM is similar to the choice of the schedule in the AQC(p=1) scheme. The speedup of AQC(p) versus the vanilla AQC is closely related to the quadratic speedup of the optimal time complexity of AQC for Grover's search [132, 82, 131, 2], in which the optimal time scheduling reduces the runtime from $T \sim \mathcal{O}(N)$ (i.e. no speedup compared to classical algorithms) to $T \sim \mathcal{O}(\sqrt{N})$ (i.e. Grover speedup). In fact, the choice of the scheduling function in Ref. [132] corresponds to AQC($p = 2$) and that in Ref. [82] corresponds to AQC($1 < p < 2$).

6.5 AQC(exp) method

Although AQC(p) achieves the optimal runtime complexity with respect to κ, the dependence on ϵ is still $\mathcal{O}(\epsilon^{-1})$, which limits the method from achieving high accuracy. It turns out that the when T is sufficiently large, the dependence on ϵ could be improved to $\mathcal{O}(\mathrm{poly}\log(1/\epsilon))$, by choosing an alternative scheduling function.

The basic observation is as follows. In AQC(p) method, the adiabatic error bound we consider, *i.e.* Eq. (6.3.4), is the so-called instantaneous adiabatic error bound, which holds true for all $s \in [0,1]$. However, when using AQC for solving QLSP, it suffices only to focus on the error bound at the final time $s = 1$. It turns out that this allows us to obtain a tighter error bound. In fact, such an error bound can be exponentially small with respect to the runtime [121, 154, 62, 2]. Roughly speaking, with an additional assumption for the Hamiltonian $H(f(s))$ that the derivatives of any order vanish at $s = 0, 1$, the adiabatic error can be bounded by $c_1 \exp(-c_2 T^{\alpha})$ for some positive constants c_1, c_2, α. Furthermore, it is proved in [62] that if the target eigenvalue is simple, then $c_1 = \mathcal{O}(\Delta_*^{-1})$ and $c_2 = \mathcal{O}(\Delta_*^3)$. Note that $\Delta_* \geq \kappa^{-1}$ for QLSP, thus, according to this bound, to obtain an ϵ-approximation, it suffices to choose $T = \mathcal{O}(\kappa^3 \ \mathrm{poly}(\log(\kappa/\epsilon)))$. This is an exponential speedup with respect to ϵ, but the dependence on the condition number becomes cubic again.

However, it is possible to reduce the runtime if the change of the Hamiltonian is slow when the gap is small, as we have already seen in the AQC(p) method. For QLSP the gap monotonically decreases and the smallest gap occurs uniquely at the final time, where the Hamiltonian $H(s)$ happens to be very slow if satisfying the assumption of vanishing derivatives at the boundary.

We consider the following schedule

$$f(s) = c_e^{-1} \int_0^s \exp\left(-\frac{1}{s'(1-s')}\right) \mathrm{d}s' \tag{6.5.1}$$

where $c_e = \int_0^1 \exp\left(-1/(s'(1-s'))\right) \mathrm{d}s'$ is a normalization constant such that $f(1) = 1$. This schedule can assure that $H^{(k)}(0) = H^{(k)}(1) = 0$ for all $k \geq 1$. We refer to the adiabatic dynamics (6.3.2) with the schedule (6.5.1) as the AQC(exp) scheme. Our main result is given in Theorem 57 (see Section 6.11 for the proof).

Theorem 57. *Let $A \in \mathbb{C}^{N\times N}$ be a Hermitian positive definite matrix with condition number κ. Then for large enough $T > 0$, the final time error $\| \, |\psi_T(1)\rangle \langle \psi_T(1)| - |\widetilde{x}\rangle \langle \widetilde{x}| \, \|_2$ of the AQC(exp) scheme is bounded by*

$$C \log(\kappa) \exp\left(-C \left(\frac{\kappa \log^2 \kappa}{T}\right)^{-\frac{1}{4}}\right). \tag{6.5.2}$$

Therefore for any $\kappa > e$, $0 < \epsilon < 1$, in order to prepare an $\epsilon-$approximation of the solution of QLSP it suffices to choose the runtime $T = \mathcal{O}\left(\kappa \log^2(\kappa) \log^4\left(\frac{\log \kappa}{\epsilon}\right)\right)$.

Compared with RM and AQC(p), although the $\log(\kappa)$ dependence reoccurs, AQC(exp) achieves an exponential speedup over RM and AQC(p) with respect to ϵ (and hence giving its name), thus is more suitable for preparing the solution of QLSP with high fidelity. Furthermore, the time scheduling of AQC(exp) is universal and AQC(exp) does not require knowledge on the bound of κ.

6.6 Gate-based implementation of AQC

We briefly discuss how to implement AQC(p) and AQC(exp) on a gate-based quantum computer. Since $|\psi_T(s)\rangle = \mathcal{T} \exp(-\mathrm{i}T \int_0^s H(f(s'))ds') |\psi_T(0)\rangle$, where $\mathcal{T}$ is a time-ordering operator, it is sufficient to implement an efficient time-dependent Hamiltonian simulation of $H(f(s))$.

One straightforward approach is to use a Trotter splitting method. The lowest order approximation takes the form

$$\begin{aligned} \mathcal{T} \exp\left(-\mathrm{i}T \int_0^s H(f(s'))\, \mathrm{d}s'\right) &\approx \prod_{m=1}^{M} \exp\left(-\mathrm{i}ThH(f(s_m))\right) \\ &\approx \prod_{m=1}^{M} \exp\left(-\mathrm{i}Th(1-f(s_m))H_0\right) \exp\left(-\mathrm{i}Thf(s_m)H_1\right) \end{aligned} \tag{6.6.1}$$

where $h = 1/M, s_m = mh$. It is proved in [48] that the error of such an approximation is $\mathcal{O}(\mathrm{poly}(\log(N))T^2/M)$, which indicates that to achieve an ϵ approximation, it suffices to choose $M = \mathcal{O}(\mathrm{poly}(\log(N))T^2/\epsilon)$. On a quantum computer, the operation $e^{-\mathrm{i}\tau H_0}, e^{-\mathrm{i}\tau H_1}$ requires a time-independent Hamiltonian simulation process, which can be implemented via techniques such as LCU and QSP [16, 109]. For a d-sparse matrix A, according to [17], the query complexity is $\widetilde{\mathcal{O}}(d\tau \log(d\tau/\epsilon))$ for a single step. Here the $\widetilde{\mathcal{O}}$ means that we neglect the log log factors. Note that the total sum of the simulation time of single steps is exactly T regardless of the choice of M, the total query complexity is $\widetilde{\mathcal{O}}(dT \log(dT/\epsilon))$. Using Theorem 56 and 57, the query complexity of AQC(p) and AQC(exp) is $\widetilde{\mathcal{O}}(d\kappa/\epsilon \log(d\kappa/\epsilon))$ and $\widetilde{\mathcal{O}}(d\kappa\, \mathrm{poly}(\log(d\kappa/\epsilon)))$, respectively. Nevertheless, M scales as $\mathcal{O}(T^2)$ with respect to the runtime T, which implies that the number of time slices should be at least $\mathcal{O}(\kappa^2)$. This breaks the linear dependence on κ if we consider the number of qubits and the gate complexity. The scaling of the Trotter expansion can be improved using high order Trotter-Suzuki formula as well as the recently developed commutator based error analysis [44], but we will not pursue this direction here.

	AQC(p)	AQC(exp)
Queries	$\widetilde{\mathcal{O}}(d\kappa/\epsilon \log(d\kappa/\epsilon))$	$\widetilde{\mathcal{O}}(d\kappa \text{ poly}(\log(d\kappa/\epsilon)))$
Qubits	$\widetilde{\mathcal{O}}(n + \log(d\kappa/\epsilon))$	$\widetilde{\mathcal{O}}(n + \log(d\kappa/\epsilon))$
Primitive gates	$\widetilde{\mathcal{O}}(nd\kappa/\epsilon \log(d\kappa/\epsilon))$	$\widetilde{\mathcal{O}}(nd\kappa \text{ poly}(\log(d\kappa/\epsilon)))$

Table 6.1: Computational costs of AQC(p) and AQC(exp) via a time-dependent Hamiltonian simulation using truncated Dyson expansion [108].

There is an efficient way is to directly perform time evolution of $H(f(s))$ without using the splitting strategy, following the algorithm proposed by Low and Wiebe in [108], where the time-dependent Hamiltonian simulation is performed based on a truncated Dyson expansion. We refer to [106] for more details on the implementation in a query model in the context of AQC. The costs of AQC(p) and AQC(exp) are summarized in Table 6.1, where for both AQC(p) and AQC(exp) almost linear dependence with respect to κ is achieved. The almost linear κ dependence cannot be expected to be improvable to $\mathcal{O}(\kappa^{1-\delta})$ with any $\delta > 0$ [73], thus both AQC(p) and AQC(exp) are almost optimal with respect to κ, and AQC(exp) further achieves an exponential speedup with respect to ϵ.

6.7 QAOA for solving QLSP

The quantum approximate optimization algorithm (QAOA) [55] considers the following parameterized wavefunction

$$|\psi_\theta\rangle := e^{-\mathrm{i}\gamma_P H_1} e^{-\mathrm{i}\beta_P H_0} \cdots e^{-\mathrm{i}\gamma_1 H_1} e^{-\mathrm{i}\beta_1 H_0} |\psi_i\rangle . \tag{6.7.1}$$

Here θ denotes the set of $2P$ adjustable real parameters $\{\beta_i, \gamma_i\}_{i=1}^{2P}$, and $|\psi_i\rangle$ is an initial wavefunction. The goal of QAOA is to choose $|\psi_i\rangle$ and to tune θ, so that $|\psi_\theta\rangle$ approximates a target state. In the context of QLSP, we may choose $|\psi_i\rangle = |\widetilde{b}\rangle$. Then with a sufficiently large P, the optimal Trotter splitting method becomes a special form of Eq. (6.7.1). Hence Theorem 57 implies that the runtime complexity of QAOA, defined to be $T := \sum_{i=1}^{P}(|\beta_i| + |\gamma_i|)$, is at most $\mathcal{O}(\kappa \text{ poly}(\log(\kappa/\epsilon)))$. We remark that the validity of such an upper bound requires a sufficiently large P and optimal choice of θ. On the other hand, our numerical results suggests that the same scaling can be achieved with a much smaller P.

For a given P, the optimal θ maximizes the fidelity as

$$\max_\theta F_\theta := |\langle \psi_\theta | \widetilde{x} \rangle|^2.$$

However, the maximization of the fidelity requires the knowledge of the exact solution $|\widetilde{x}\rangle$ which is not practical. We may instead solve the following minimization problem

$$\min_{\theta} \langle \psi_\theta | H_1^2 | \psi_\theta \rangle \,. \tag{6.7.2}$$

Since the null space of H_1 is of dimension 2, the unconstrained minimizer $|\psi_\theta\rangle$ seems possible to only have a small overlap with $|\widetilde{x}\rangle$. However, this is not a problem due to the choice of the initial state $|\psi_i\rangle = |\widetilde{b}\rangle$. Notice that by the variational principle the minimizer $|\psi_\theta\rangle$ maximizes $\langle \psi_\theta | P_0(1) | \psi_\theta \rangle$. Using the fact that $e^{-\mathrm{i}\beta H_0} |\bar{b}\rangle = e^{-\mathrm{i}\gamma H_1} |\bar{b}\rangle = |\bar{b}\rangle$ for any β, γ, we obtain $\langle \bar{b} | \psi_\theta \rangle = \langle \bar{b} | \widetilde{b} \rangle = 0$, which means the QAOA ansatz prevents the transition to $|\bar{b}\rangle$, similar to AQC. Then $\langle \psi_\theta | P_0(1) | \psi_\theta \rangle = \langle \psi_\theta | \widetilde{x} \rangle \langle \widetilde{x} | \psi_\theta \rangle = F_\theta$, so the minimizer of Eq. (6.7.2) indeed maximizes the fidelity.

For every choice of θ, we evaluate the expectation value $\langle \psi_\theta | H_1^2 | \psi_\theta \rangle$. Then the next θ is adjusted on a classical computer towards minimizing the objective function. The process is repeated till convergence. Efficient classical algorithms for the optimization of parameters in QAOA are currently an active topic of research, including methods using gradient optimization [164, 114], Pontryagin's maximum principle (PMP) [160, 11], reinforcement learning [29, 125], to name a few. Algorithm 1 describes the procedure using QAOA to solve QLSP.

Algorithm 1 QAOA for solving QLSP

1: Initial parameters $\theta^{(0)} = \{\beta_i, \gamma_i\}_{i=1}^{2P}$.
2: **for** $k = 0, 1, \ldots$ **do**
3: Perform Hamiltonian simulation to obtain $\psi_\theta^{(k)}$.
4: Measure $O(\theta^{(k)}) = \langle \psi_\theta^{(k)} | H_1^2 | \psi_\theta^{(k)} \rangle$.
5: If $O(\theta^{(k)}) < \epsilon/\kappa^2$, exit the loop.
6: Choose $\theta^{(k+1)}$ using a classical optimization method.
7: **end for**

Compared to AQC(p) and AQC(exp), QAOA have the following three potential advantages. The first advantage is that the optimization should automatically (at least in principle) achieve or even exceed the result obtained by AQC with the best scheduling function. Second, as discussed before, one way of the implementation of AQC(p) and AQC(exp) using an operator splitting method requires the time interval to be explicitly split into a large number of intervals, while numerical results indicate that the number of intervals P in QAOA be much smaller, thus resulting in a lower depth quantum circuit. Compared to AQC, QAOA has the additional advantage that it only consists of $2P$ *time-independent* Hamiltonian simulation problem, once θ is known.

Despite the potential advantages, several severe caveats of using QAOA for QLSP arise when we consider beyond the time complexity. The first is the cost for the classical optimization is hard to known *a priori*. The optimization may require many iterations, which overwhelms the gain of QAOA's reduced time complexity. The second is related to the accurate computation of the objective function $O(\theta^{(k)})$. Note that the minimal spectrum gap of H_1 is $\mathcal{O}(\kappa^{-1})$. In order to obtain an ϵ-approximation, the precision of measuring $O(\theta) = \langle\psi_\theta|H_1|\psi_\theta\rangle$ should be at least $\mathcal{O}(\epsilon/\kappa^2)$. Hence $\mathcal{O}(\kappa^4/\epsilon^2)$ repeated measurements can be needed to achieve the desired accuracy.

6.8 Generalization to non-Hermitian matrices

Now we discuss the case when A is not Hermitian positive definite. First we still assume that A is Hermitian (but not necessarily positive definite). In this case we adopt the family of Hamiltonians introduced in [144], which overcomes the difficulty brought by the indefiniteness of A at the expense of enlarging the Hilbert space to dimension $4N$ (so two ancilla qubits are needed to enlarge the matrix block). Here we define

$$H_0 = \sigma_+ \otimes [(\sigma_z \otimes I_N)Q_{+,b}] + \sigma_- \otimes [Q_{+,b}(\sigma_z \otimes I_N)]$$

where $Q_{+,b} = I_{2N} - |+,b\rangle\langle+,b|$, and $|\pm\rangle = \frac{1}{\sqrt{2}}(|0\rangle \pm |1\rangle)$. The null space of H_0 is $\text{Null}(H_0) = \text{span}\{|0,-,b\rangle, |1,+,b\rangle\}$. We also define

$$H_1 = \sigma_+ \otimes [(\sigma_x \otimes A)Q_{+,b}] + \sigma_- \otimes [Q_{+,b}(\sigma_x \otimes A)]$$

Note that $\text{Null}(H_1) = \text{span}\{|0,+,x\rangle, |1,+,b\rangle\}$. Therefore the solution of the QLSP can be obtained if we can prepare the zero-energy state $|0,+,x\rangle$ of H_1.

The family of Hamiltonians for AQC(p) is still given by $H(f(s)) = (1-f(s))H_0 + f(s)H_1, 0 \leqslant s \leqslant 1$. Similar to the case of Hermitian positive definite matrices, there is a double degeneracy of the eigenvalue 0, and we aim at preparing one of the eigenstate via time-optimal adiabatic evolution. More precisely, for any s, $|1,+,b\rangle$ is always in $\text{Null}(H(f(s)))$, and there exists a state $|\widetilde{x}(s)\rangle$ with $|\widetilde{x}(0)\rangle = |0,-,b\rangle, |\widetilde{x}(1)\rangle = |0,+,x\rangle$, such that $\text{Null}(H(f(s))) = \{|\widetilde{x}(s)\rangle, |1,+,b\rangle\}$. Such degeneracy will not influence the adiabatic computation starting with $|0,-,b\rangle$ for the same reason we discussed for Hermitian positive definite case (also discussed in [144]), and the error of AQC(p) is still bounded by $\eta(s)$ given in Eq. (6.3.4).

Furthermore, the eigenvalue 0 is separated from the rest of the eigenvalues of $H(f(s))$ by a gap $\Delta(f(s)) \geqslant \sqrt{(1-f(s))^2 + (f(s)/\kappa)^2}$ [144]. For technical simplicity, note that $\sqrt{(1-f)^2 + (f/\kappa)^2} \geq (1 - f + f/\kappa)/\sqrt{2}$ for all $0 \leq f \leq 1$, we define the lower bound of the gap to be $\Delta_*(f) = (1 - f + f/\kappa)/\sqrt{2}$, which is exactly proportional to that for the

Hermitian positive definite case. Therefore, we can use exactly the same time schedules as the Hermitian positive definite case to perform AQC(p) and AQC(exp) schemes, and properties of AQC(p) and AQC(exp) are stated in the following theorems (see Section 6.10 and Section 6.11 for the proof).

Theorem 58. *Let $A \in \mathbb{C}^{N\times N}$ be a Hermitian matrix (not necessarily positive definite) with condition number κ. For any choice of $1 < p < 2$, the AQC(p) scheme gives*

$$\| \left|\psi_T(s)\right\rangle \left\langle\psi_T(s)\right| - \left|0,+,x\right\rangle \left\langle 0,+,x\right| \|_2 \leq C\kappa/T. \tag{6.8.1}$$

Therefore in order to prepare an $\epsilon-$approximation of the solution of QLSP it suffices to choose the runtime $T = \mathcal{O}(\kappa/\epsilon)$. Furthermore, when $p = 1, 2$, the bound of the runtime becomes $T = \mathcal{O}(\kappa\log(\kappa)/\epsilon)$.

Theorem 59. *Let $A \in \mathbb{C}^{N\times N}$ be a Hermitian matrix (not necessarily positive definite) with condition number κ. Then for large enough $T > 0$, the final time error $\| \left|\psi_T(1)\right\rangle \left\langle\psi_T(1)\right| - \left|0,+,x\right\rangle \left\langle 0,+,x\right| \|_2$ of the AQC(exp) scheme is bounded by*

$$C\log(\kappa)\exp\left(-C\left(\frac{\kappa\log^2\kappa}{T}\right)^{-\frac{1}{4}}\right). \tag{6.8.2}$$

Therefore for any $\kappa > e$, $0 < \epsilon < 1$, in order to prepare an $\epsilon-$approximation of the solution of QLSP it suffices to choose the runtime $T = \mathcal{O}\left(\kappa\log^2(\kappa)\log^4\left(\frac{\log\kappa}{\epsilon}\right)\right)$.

For a most general square matrix $A \in \mathbb{C}^{N\times N}$, we may transform it into the Hermitian case at the expense of further doubling the dimension of the Hilbert space. Introduce the solution of the adjoint QLSP $\left|y\right\rangle = (A^\dagger)^{-1}\left|b\right\rangle / \|(A^\dagger)^{-1}\left|b\right\rangle\|_2$, and consider an extended QLSP $\mathfrak{A}\left|\mathfrak{x}\right\rangle = \left|\mathfrak{b}\right\rangle$ in dimension $2N$ where

$$\mathfrak{A} = \sigma_+\otimes A + \sigma_-\otimes A^\dagger = \begin{pmatrix} 0 & A \\ A^\dagger & 0 \end{pmatrix}, \quad \left|\mathfrak{b}\right\rangle = \left|+,b\right\rangle.$$

Note that $\mathfrak{A}$ is a Hermitian matrix of dimension $2N$, with condition number κ and $\|\mathfrak{A}\|_2 = 1$, and $\left|\mathfrak{x}\right\rangle := \frac{1}{\sqrt{2}}(\left|1,x\right\rangle + \left|0,y\right\rangle)$ solves the extended QLSP. Therefore we can directly apply AQC(p) and AQC(exp) for Hermitian matrix $\mathfrak{A}$ to prepare an ϵ-approximation of x and y simultaneously. The total dimension of the Hilbert space becomes $8N$ for non-Hermitian matrix A (therefore three ancilla qubits are needed).

6.9 Numerical results

We first report the performance of AQC(p), AQC(exp) and QAOA for a series of Hermitian positive definite dense matrices with varying condition numbers, together with the performance of RM and vanilla AQC.

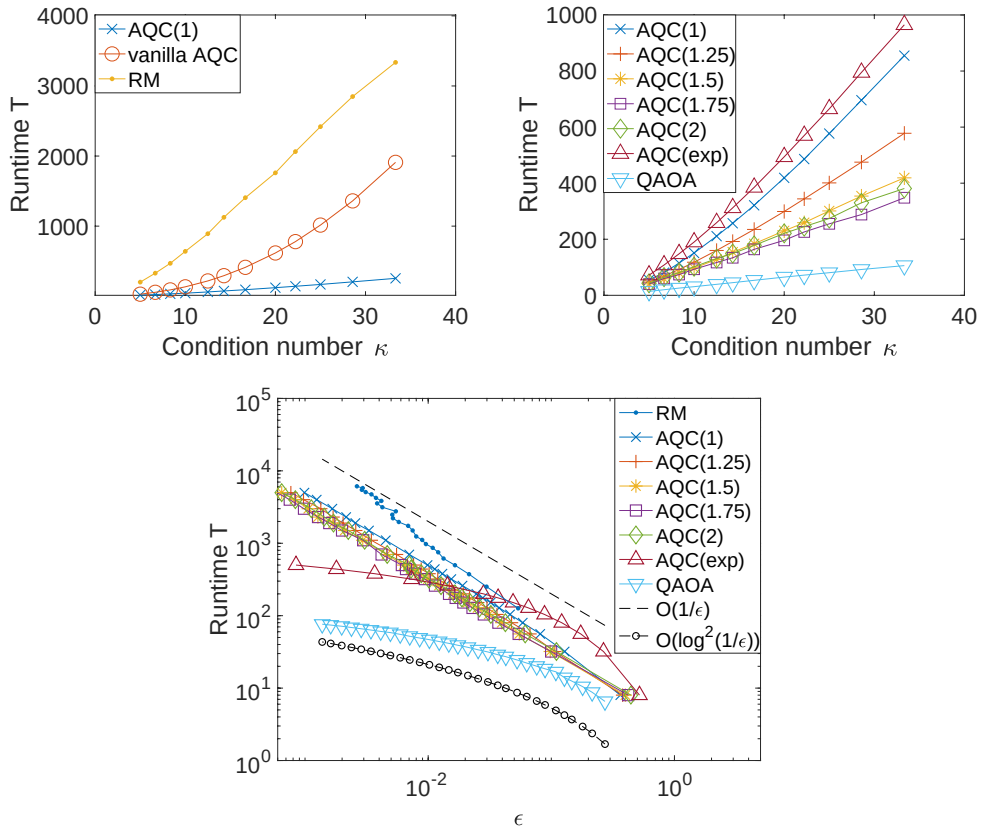


Figure 6.8.1: Simulation results for the Hermitian positive definite example. Top/Middle: the runtime to reach desired fidelity 0.99/0.999 as a function of the condition number. Bottom: a log-log plot of the runtime as a function of the accuracy with $\kappa = 10$.

Setup

For simulation purpose, the AQC schemes are carried out using a Trotter splitting method with a time step size 0.2. We use the gradient descent method to optimize QAOA and record the running time corresponding to the lowest error in each case. In QAOA we also use the true fidelity to measure the error. RM is a Monte Carlo method, and each RM calculation involves performing 200 independent runs to obtain the density matrix $\rho^{(i)}$ for i'th repetition, then we use the averaged density $\bar{\rho} = 1/n_{\text{rep}} \sum \rho^{(i)}$ to compute the error. We report the averaged runtime of each single RM calculation. We perform calculations for a series of 64-dimensional Hermitian positive definite dense matrices A_1, and 32-dimensional non-Hermitian dense matrices A_2 with varying condition number κ.

methods	scaling w.r.t. κ	scaling w.r.t. $1/\epsilon$
vanilla AQC	2.2022	/
RM	1.4912	1.3479
AQC(1)	1.4619	1.0482
AQC(1.25)	1.3289	1.0248
AQC(1.5)	1.2262	1.0008
AQC(1.75)	1.1197	0.9899
AQC(2)	1.1319	0.9904
AQC(exp)	1.3718	0.5377
AQC(exp)	/	1.7326 (w.r.t. $\log(1/\epsilon)$)
QAOA	1.0635	0.4188
QAOA	/	1.4927 (w.r.t. $\log(1/\epsilon)$)

Table 6.2: Numerical scaling of the runtime as a function of the condition number and the accuracy, respectively, for the Hermitian positive definite example.

For concreteness, for the Hermitian positive definite example, we choose $A = U\Lambda U^\dagger$. Here U is an orthogonal matrix obtained by Gram-Schmidt orthogonalization (implemented via a QR factorization) of the discretized periodic Laplacian operator given by

$$L = \begin{pmatrix} 1 & -0.5 & & & & -0.5 \\ -0.5 & 1 & -0.5 & & & \\ & -0.5 & 1 & -0.5 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -0.5 & 1 & -0.5 \\ -0.5 & & & & -0.5 & 1 \end{pmatrix}. \tag{6.9.1}$$

Λ is chosen to be a diagonal matrix with diagonals uniformly distributed in $[1/\kappa, 1]$. More precisely, $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_N)$ with $\lambda_k = 1/\kappa + (k-1)h, h = (1-1/\kappa)/(N-1)$. Such construction ensures A to be a Hermitian positive definite matrix which satisfies $\|A\|_2 = 1$ and the condition number of A is κ. We choose $|b\rangle = \sum_{k=1}^N u_k / \| \sum_{k=1}^N u_k \|_2$ where $\{u_k\}$ is the set of the column vectors of U. Here $N = 64$.

For the non-Hermitian positive definite example, we choose $A = U\Lambda V^\dagger$. Here U is the same as those in the Hermitian positive definite case, except that the dimension is reduced to $N = 32$. $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_N)$ with $\lambda_k = (-1)^k(1/\kappa + (k-1)h), h = (1-1/\kappa)/(N-1)$.

V is an orthogonal matrix obtained by Gram-Schmidt orthogonalization of the matrix

$$K = \begin{pmatrix} 2 & -0.5 & & & & -0.5 \\ -0.5 & 2 & -0.5 & & & \\ & -0.5 & 2 & -0.5 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -0.5 & 2 & -0.5 \\ -0.5 & & & & -0.5 & 2 \end{pmatrix}. \tag{6.9.2}$$

Such construction ensures A to be non-Hermitian, satisfying $\|A\|_2 = 1$ and the condition number of A is κ. We choose the same $|b\rangle$ as that in the Hermitian positive definite example.

Results

Fig 6.8.1 shows how the total runtime T depends on the condition number κ and the accuracy ϵ for the Hermitian positive definite case. The numerical scaling is reported in Table 6.2. For the κ dependence, despite that RM and AQC(1) share the same asymptotic linear complexity with respect to κ, we observe that the preconstant of RM is larger due to its Monte Carlo strategy and the mixed state nature resulting in the same scaling of errors in fidelity and density. The asymptotic scaling of the vanilla AQC is at least $\mathcal{O}(\kappa^2)$. When higher fidelity (0.999) is desired, the cost of vanilla AQC becomes too expensive, and we only report the timing of RM, AQC(p), AQC(exp) and QAOA. For the κ dependence tests, the depth of QAOA ranges from 8 to 60. For the ϵ dependence test, the depth of QAOA is fixed to be 20. We find that the runtime for AQC(p), AQC(exp) and QAOA depends approximately linearly on κ, while QAOA has the smallest runtime overall. It is also interesting to observe that although the asymptotic scalings of AQC(1) and AQC(2) are both bounded by $\mathcal{O}(\kappa \log \kappa)$ instead of $\mathcal{O}(\kappa)$, the numerical performance of AQC(2) is much better than AQC(1); in fact, the scaling is very close to that with the optimal value of p. For the ϵ dependence, the scaling of RM and AQC(p) is $\mathcal{O}(1/\epsilon)$, which agrees with the error bound. Again the preconstant of RM is slightly larger. Our results also confirm that AQC(exp) only depends poly logarithmically on ϵ. Note that when ϵ is relatively large, AQC(exp) requires a longer runtime than that of AQC(p), and it eventually outperforms AQC(p) when ϵ is small enough. The numerical scaling of QAOA with respect to ϵ is found to be only $\mathcal{O}(\log^{1.5}(1/\epsilon))$ together with the smallest preconstant.

Fig 6.9.1 and Table 6.3 demonstrate the simulation results for non-Hermitian matrices. We find that numerical performances of RM, AQC(p), AQC(exp) and QAOA are similar with that of the Hermitian positive definite case. Again QAOA obtains the optimal performance in terms of the runtime. The numerical scaling of the optimal AQC(p) is found to be $\mathcal{O}(\kappa/\epsilon)$, while the time complexity of QAOA and AQC(exp) is only $\mathcal{O}(\kappa \text{ poly}(\log(1/\epsilon)))$.

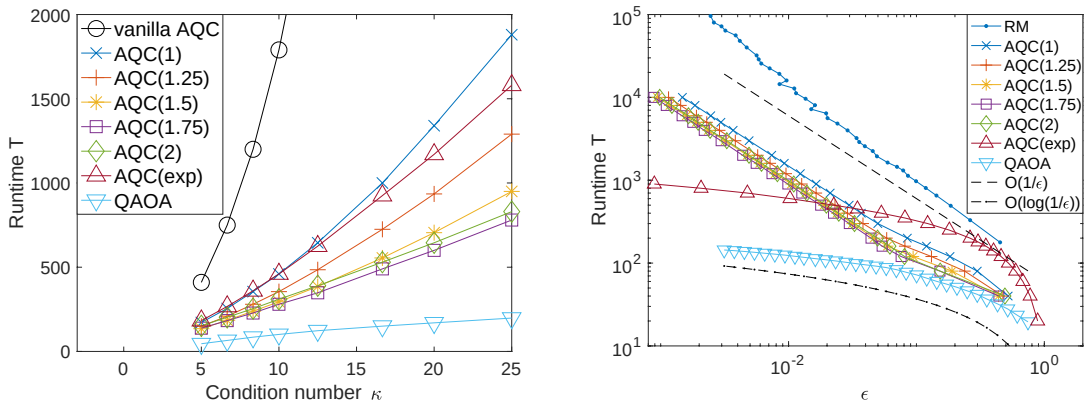


Figure 6.9.1: Simulation results for the non-Hermitian example. Top: the runtime to reach 0.999 fidelity as a function of the condition number. Bottom: a log-log plot of the runtime as a function of the accuracy with $\kappa = 10$.

methods	scaling w.r.t. κ	scaling w.r.t. $1/\epsilon$
vanilla AQC	2.1980	/
RM	/	1.2259
AQC(1)	1.4937	0.9281
AQC(1.25)	1.3485	0.9274
AQC(1.5)	1.2135	0.9309
AQC(1.75)	1.0790	0.9378
AQC(2)	1.0541	0.9425
AQC(exp)	1.3438	0.4415
AQC(exp)		0.9316 (w.r.t. $\log(1/\epsilon)$)
QAOA	0.8907	0.3283
QAOA	/	0.7410 (w.r.t. $\log(1/\epsilon)$)

Table 6.3: Numerical scaling of the runtime as a function of the condition number and the accuracy, respectively, for the non-Hermitian example.

6.10 Proof of Theorem 56 and Theorem 58

The proof of Theorem 56 and Theorem 58 rests on some delicate cancellation of the time derivatives $\|H^{(1)}\|_2, \|H^{(2)}\|_2$ and the gap $\Delta(f(s))$ in the error bound, and can be completed by carefully analyzing the κ-dependence of each term in $\eta(s)$ given in Eq. (6.3.4). Please note that the proof strategy largely follows the proof of Theorem 7, but here we show how the generic quadratic gap dependence can be improved to linear dependence in specific examples when more information on the spectrum is available.

Note that in both cases $H(f) = (1-f)H_0 + fH_1$, and we let $\Delta_*(f) = (1 - f + f/\kappa)/\sqrt{2}$ since such Δ_* can serve as a lower bound of the spectrum gap for both the case of Theorem 56 and 57. We first compute the derivatives of $H(f(s))$ by chain rule as

$$H^{(1)}(s) = \frac{d}{ds}H(f(s)) = \frac{dH(f(s))}{df}\frac{df(s)}{ds} = (H_1 - H_0)c_p\Delta_*^p(f(s)),$$

and

$$\begin{aligned} H^{(2)}(s) = \frac{d}{ds}H^{(1)}(s) &= \frac{d}{ds}\left((H_1 - H_0)c_p\Delta_*^p(f(s))\right) \\ &= (H_1 - H_0)c_p p\Delta_*^{p-1}(f(s))\frac{d\Delta_*(f(s))}{df}\frac{df(s)}{ds} \\ &= \frac{1}{\sqrt{2}}(-1 + 1/\kappa)(H_1 - H_0)c_p^2 p\Delta_*^{2p-1}(f(s)). \end{aligned}$$

Then the first two terms of $\eta(s)$ can be rewritten as

$$\begin{aligned} &\frac{\|H^{(1)}(0)\|_2}{T\Delta^2(0)} + \frac{\|H^{(1)}(s)\|_2}{T\Delta^2(f(s))} \le \frac{\|H^{(1)}(0)\|_2}{T\Delta_*^2(0)} + \frac{\|H^{(1)}(s)\|_2}{T\Delta_*^2(f(s))} \\ =&\frac{\|(H_1 - H_0)c_p\Delta_*^p(f(0))\|_2}{T\Delta_*^2(0)} + \frac{\|(H_1 - H_0)c_p\Delta_*^p(f(s))\|_2}{T\Delta_*^2(f(s))} \\ \le&\frac{C}{T}\left(c_p\Delta_*^{p-2}(0) + c_p\Delta_*^{p-2}(f(s))\right) \\ \le&\frac{C}{T}\left(c_p\Delta_*^{p-2}(0) + c_p\Delta_*^{p-2}(1)\right) \end{aligned}$$

Here C stands for a general positive constant independent of s, Δ, T. To compute the remaining two terms of $\eta(s)$, we use the following change of variable

$$u = f(s'), \quad du = \frac{d}{ds'}f(s')ds' = c_p\Delta_*^p(f(s'))ds',$$

and the last two terms of $\eta(s)$ become

$$\begin{aligned}
&\frac{1}{T}\int_0^s \frac{\|H^{(2)}\|_2}{\Delta^2}ds' \le \frac{1}{T}\int_0^s \frac{\|H^{(2)}\|_2}{\Delta_*^2}ds' \\
=&\frac{1}{T}\int_0^s \frac{\|\frac{1}{\sqrt{2}}(-1+1/\kappa)(H_1-H_0)c_p^2 p\Delta_*^{2p-1}(f(s'))\|_2}{\Delta_*^2(f(s'))}ds' \\
=&\frac{1}{T}\int_0^{f(s)} \frac{\|\frac{1}{\sqrt{2}}(-1+1/\kappa)(H_1-H_0)c_p^2 p\Delta_*^{2p-1}(u)\|_2}{\Delta_*^2(u)}\frac{du}{c_p\Delta_*^p(u)} \\
\le&\frac{C}{T}\left((1-1/\kappa)c_p\int_0^{f(s)}\Delta_*^{p-3}(u)du\right) \\
\le&\frac{C}{T}\left((1-1/\kappa)c_p\int_0^{1}\Delta_*^{p-3}(u)du\right),
\end{aligned}$$

and similarly

$$\begin{aligned}
&\frac{1}{T}\int_0^s \frac{\|H^{(1)}\|_2^2}{\Delta^3}ds' \le \frac{1}{T}\int_0^s \frac{\|H^{(1)}\|_2^2}{\Delta_*^3}ds' \\
=&\frac{1}{T}\int_0^s \frac{\|(H_1-H_0)c_p\Delta_*^p(f(s'))\|_2^2}{\Delta_*^3(f(s'))}ds' \\
=&\frac{1}{T}\int_0^{f(s)} \frac{\|(H_1-H_0)c_p\Delta_*^p(u)\|_2^2}{\Delta_*^3(u)}\frac{du}{c_p\Delta_*^p(u)} \\
\le&\frac{C}{T}\left(c_p\int_0^{f(s)}\Delta_*^{p-3}(u)du\right) \\
\le&\frac{C}{T}\left(c_p\int_0^{1}\Delta_*^{p-3}(u)du\right).
\end{aligned}$$

Summarize all terms above, an upper bound of $\eta(s)$ is

$$\begin{aligned}
\eta(s) \le \frac{C}{T}\Big\{ &\left(c_p\Delta_*^{p-2}(0)+c_p\Delta_*^{p-2}(1)\right) \\
&+\left((1-1/\kappa)c_p\int_0^1\Delta_*^{p-3}(u)du\right)+\left(c_p\int_0^1\Delta_*^{p-3}(u)du\right)\Big\} \\
= \frac{C}{T}\Big\{ &2^{-(p-2)/2}\left(c_p+c_p\kappa^{2-p}\right)+\left((1-1/\kappa)c_p\int_0^1\Delta_*^{p-3}(u)du\right)+\left(c_p\int_0^1\Delta_*^{p-3}(u)du\right)\Big\}.
\end{aligned}$$

Finally, since for $1<p<2$

$$c_p=\int_0^1\Delta_*^{-p}(u)du=\frac{2^{p/2}}{p-1}\frac{\kappa}{\kappa-1}(\kappa^{p-1}-1),$$

and

$$\int_0^1 \Delta_*^{p-3}(u)du = \frac{2^{-(p-3)/2}}{2-p}\frac{\kappa}{\kappa-1}(\kappa^{2-p}-1),$$

we have

$$\begin{aligned}\eta(s) \leq & \frac{C}{T}\Big\{\frac{\kappa}{\kappa-1}(\kappa^{p-1}-1)+\frac{\kappa}{\kappa-1}(\kappa-\kappa^{2-p}) \\ & + \frac{\kappa}{\kappa-1}(\kappa^{p-1}-1)(\kappa^{2-p}-1)+\left(\frac{\kappa}{\kappa-1}\right)^2(\kappa^{p-1}-1)(\kappa^{2-p}-1)\Big\}.\end{aligned}$$

The leading term of the bound is $\mathcal{O}(\kappa/T)$ when $1 < p < 2$.

Now we consider the limiting case when $p = 1, 2$. Note that the bound for $\eta(s)$ can still be written as

$$\begin{aligned}\eta(s) \leq & \frac{C}{T}\Big\{ \left(c_p\Delta_*^{p-2}(0)+c_p\Delta_*^{p-2}(1)\right) \\ & + \left((1-1/\kappa)c_p\int_0^1 \Delta_*^{p-3}(u)du\right)+\left(c_p\int_0^1\Delta_*^{p-3}(u)du\right)\Big\} \\ = & \frac{C}{T}\Big\{2^{-(p-2)/2}\left(c_p+c_p\kappa^{2-p}\right)+(1-1/\kappa)c_pc_{3-p}+c_pc_{3-p}\Big\}.\end{aligned}$$

Straightforward computation shows that

$$c_1 = \int_0^1 \Delta_*^{-1}(u)du = \sqrt{2}\frac{\kappa}{\kappa-1}\log(\kappa)$$

and

$$c_2 = \int_0^1 \Delta_*^{-2}(u)du = 2\frac{\kappa}{\kappa-1}(\kappa-1).$$

Hence when $p = 1, 2$,

$$\eta(s) \leq \frac{C}{T}\Big\{2^{-(p-2)/2}\left(c_p+c_p\kappa^{2-p}\right)+(1-1/\kappa)c_1c_2+c_1c_2\Big\} \leq C\frac{\kappa\log(\kappa)}{T}.$$

This completes the proof of Theorem 56 and Theorem 58.

6.11 Proof of Theorem 57 and Theorem 59

We provide a rigorous proof of the error bound for AQC(exp) scheme. We mainly follow the methodology of [121] and a part of technical treatments of [62]. Our main contribution

is carefully revealing an explicit constant dependence in the adiabatic theorem, which is the key to obtain the $\tilde{\mathcal{O}}(\kappa)$ scaling. In the AQC(exp) scheme, the Hamiltonian $H(s) = (1-f(s))H_0 + f(s)H_1$ with $\|H_0\|, \|H_1\| \leq 1$ and

$$f(s) = \frac{1}{c_e}\int_0^s \exp\left(-\frac{1}{s'(1-s')}\right) \mathrm{d}s'. \tag{6.11.1}$$

The normalization constant $c_e = \int_0^1 \exp(-\frac{1}{t(1-t)})dt \approx 0.0070$. Let $U_T(s)$ denote the corresponding unitary evolution operator, and $P_0(s)$ denote the projector onto the eigenspace corresponding to 0. We use $\Delta_*(f) = (1-f+f/\kappa)/\sqrt{2}$ since this can serve as a lower bound of the spectrum gap for both the case of Theorem 57 and Theorem 59.

We first restate the theorems universally with more technical details as following.

Theorem 60. *Assume the condition number $\kappa > e$. Then the final time adiabatic error $|1 - \langle\psi_T(1)|P_0(1)|\psi_T(1)\rangle|$ of AQC(exp) can be bounded by η_1^2 where*

(a) for arbitrary N,

$$\eta_1^2 = A_1 D \log^2\kappa \left(C_2 \frac{\kappa \log^2\kappa}{T} N^4\right)^N$$

where A_1, D, C_2 are positive constants which are independent of T, κ and N.

(b) if T is large enough such that

$$16eA_1^{-1}D\left(\frac{4\pi^2}{3}\right)^3 \frac{\kappa\log^2\kappa}{T} \leq 1,$$

then

$$\eta_1^2 = C_1 \log^2\kappa \exp\left(-\left(C_2\frac{\kappa\log^2\kappa}{T}\right)^{-\frac{1}{4}}\right)$$

where A_1, D, C_1, C_2 are positive constants which are independent of T and κ.

Corollary 61. *For any $\kappa > e, 0 < \epsilon < 1$, to prepare an ϵ-approximation of the solution of QLSP using AQC(exp), it is sufficient to choose the runtime $T = \mathcal{O}\left(\kappa \log^2\kappa \log^4\left(\frac{\log\kappa}{\epsilon}\right)\right)$.*

Proof. We start the proof by considering the projector $P(s)$ onto an invariant space of H, then $P(s)$ satisfies

$$\mathrm{i}\frac{1}{T}\partial_s P(s) = [H(s), P(s)], \quad P^2(s) = P(s). \tag{6.11.2}$$

We try the ansatz (only formally)

$$P(s) = \sum_{j=0}^{\infty} E_j(s) T^{-j}. \tag{6.11.3}$$

Substitute it into the Heisenberg equation and match terms with the same orders, we get

$$[H(s), E_0(s)] = 0, \quad \mathrm{i}\partial_s E_j(s) = [H(s), E_{j+1}(s)], \quad E_j(s) = \sum_{m=0}^{j} E_m(s) E_{j-m}(s). \tag{6.11.4}$$

It has been proved in [121] that the solution of (6.11.4) with initial condition $E_0 = P_0$ is given by

$$E_0(s) = P_0(s) = -(2\pi\mathrm{i})^{-1} \oint_{\Gamma(s)} (H(s) - z)^{-1} dz, \tag{6.11.5}$$

$$E_j(s) = (2\pi)^{-1} \oint_{\Gamma(s)} (H(s) - z)^{-1} [E_{j-1}^{(1)}(s), P_0(s)](H(s) - z)^{-1} dz + S_j(s) - 2P_0(s)S_j(s)P_0(s) \tag{6.11.6}$$

where $\Gamma(s) = \{z \in \mathbb{C} : |z| = \Delta(s)/2\}$ and

$$S_j(s) = \sum_{m=1}^{j-1} E_m(s) E_{j-m}(s). \tag{6.11.7}$$

Furthermore given $E_0 = P_0$, such solution is unique.

In general, Eq. (6.11.3) does not converge, so for arbitrary positive integer N we define a truncated series as

$$P_N(s) = \sum_{j=0}^{N} E_j(s) T^{-j}. \tag{6.11.8}$$

Then

$$\mathrm{i}\frac{1}{T} P_N^{(1)} - [H, P_N] = \mathrm{i}\frac{1}{T} \sum_{j=0}^{N} E_j^{(1)} T^{-j} - \sum_{j=0}^{N} [H, E_j] T^{-j} = \mathrm{i} T^{-(N+1)} E_N^{(1)}.$$

In Lemma 64, we prove that $P_N(0) = P_0(0)$ and $P_N(1) = P_0(1)$, then the adiabatic error becomes

$$\begin{aligned} |1 - \langle \psi_T(1) | P_0(1) | \psi_T(1) \rangle| &= |\langle \psi_T(0) | P_0(0) | \psi_T(0) \rangle - \langle \psi_T(0) | U_T(1)^{-1} P_0(1) U_T(1) | \psi_T(0) \rangle| \\ &\leq \| P_0(1) - U_T(1)^{-1} P_0(0) U_T(1) \| \\ &= \| P_N(1) - U_T(1)^{-1} P_N(0) U_T(1) \| \\ &= \left\| \int_0^1 ds \frac{d}{ds} \left(U_T^{-1} P_N U_T \right) \right\|. \end{aligned}$$

Straightforward computations show that

$$\frac{d}{ds}(U_T^{-1}) = -U_T^{-1}\frac{d}{ds}(U_T)U_T^{-1} = -U_T^{-1}\frac{T}{\mathrm{i}}HU_TU_T^{-1} = -\frac{T}{\mathrm{i}}U_T^{-1}H,$$

$$\begin{aligned}\frac{d}{ds}\left(U_T^{-1}P_NU_T\right) &= \frac{d}{ds}(U_T^{-1})P_NU_T + U_T^{-1}\frac{d}{ds}(P_N)U_T + U_T^{-1}P_N\frac{d}{ds}(U_T)\\ &= -\frac{T}{\mathrm{i}}U_T^{-1}HP_NU_T + U_T^{-1}\frac{T}{\mathrm{i}}[H,P_N]U_T + U_T^{-1}T^{-N}E_N^{(1)}U_T + \frac{T}{\mathrm{i}}U_T^{-1}P_NHU_T\\ &= T^{-N}U_T^{-1}E_N^{(1)}U_T,\end{aligned}$$

therefore

$$|1 - \langle\psi_T(1)|P_0(1)|\psi_T(1)\rangle\,| \le \left\|\int_0^1 T^{-N}U_T^{-1}E_N^{(1)}U_T ds\right\| \le T^{-N}\max_{s\in[0,1]}\|E_N^{(1)}\|.$$

In Lemma 69, we prove that (the constant $c_f = 4\pi^2/3$)

$$\begin{aligned}\|E_N^{(1)}\| &\le A_1A_2^NA_3\frac{[(N+1)!]^4}{(1+1)^2(N+1)^2}\\ &= \frac{A_1}{4}D\log^2\kappa\left[A_1^{-1}c_f^3\frac{16}{\Delta}D\log^2\kappa\right]^N\frac{[(N+1)!]^4}{(N+1)^2}\\ &\le \frac{A_1}{4}D\log^2\kappa\left[16A_1^{-1}Dc_f^3\kappa\log^2\kappa\right]^N\frac{[(N+1)!]^4}{(N+1)^2}\\ &\le A_1D\log^2\kappa\left[16A_1^{-1}Dc_f^3\kappa\log^2\kappa N^4\right]^N\end{aligned}$$

where the last inequality comes from the fact that $[(N+1)!]^4/(N+1)^2 \le 4N^{4N}$. This completes the proof of part (a).

When T is large enough, we now choose

$$N = \left\lfloor\left(16eA_1^{-1}Dc_f^3\frac{\kappa\log^2\kappa}{T}\right)^{-\frac{1}{4}}\right\rfloor \ge 1,$$

then

$$\begin{aligned}|1 - \langle\psi_T(1)|P_0(1)|\psi_T(1)\rangle\,| &\le A_1D\log^2\kappa\left[16A_1^{-1}Dc_f^3\frac{\kappa\log^2\kappa}{T}N^4\right]^N\\ &\le A_1D\log^2\kappa\exp\left(-\left(16eA_1^{-1}Dc_f^3\frac{\kappa\log^2\kappa}{T}\right)^{-\frac{1}{4}}\right).\end{aligned}$$

This completes the proof of part (b). □

The remaining part is devoted to some preliminary results regarding H, E and the technical estimates for the growth of E_j. It is worth mentioning in advance that in the proof we will encounter many derivatives taken on a contour integral. In fact all such derivatives taken on a contour integral will not involve derivatives on the contour. Specifically, since $(H(s)-z)^{-1}$ is analytic for any $0<|z|<\Delta(s)$, then for any $s_0\in(0,1)$, there exists a small enough neighborhood $B_\delta(s_0)$ such that $\forall s\in B_\delta(s_0)$, $\oint_{\Gamma(s)} G(s,(H(s)-z)^{-1})dz=\oint_{\Gamma(s_0)} G(s,(H(s)-z)^{-1})dz$ for any smooth mapping G. This means locally the contour integral does not depend on the smooth change of the contour, thus the derivatives will not involve derivatives on the contour. In the spirit of this trick, we write the resolvent $R(z,s,s_0)=(H(s)-z)^{-1}$ for $0\le s\le 1, 0\le s_0\le 1, z\in\mathbb{C}$ and $|z|=\Delta(s_0)/2$ and let $R^{(k)}$ denote the partial derivative with respect to s, *i.e.* $\frac{\partial}{\partial s}R(z,s,s_0)$, which means by writing $R^{(k)}$ we only consider the explicit time derivatives brought by H.

Lemma 62. *(a) $H(s)\in C^\infty$ with $H^{(k)}(0)=H^{(k)}(1)=0$ for all $k\ge 1$.*

(b) There's a gap $\Delta(s)\ge\Delta_*(s)=((1-f(s))+f(s)/\kappa)/\sqrt{2}$ which separates 0 from the rest of the spectrum.

The following lemma gives the bound for the derivatives of H.

Lemma 63. *For every $k\ge 1, 0<s<1$,*

$$\|H^{(k)}(s)\|\le b(s)a(s)^k\frac{(k!)^2}{(k+1)^2}, \tag{6.11.9}$$

where

$$b(s)=\frac{2e}{c_e}\exp\left(-\frac{1}{s(1-s)}\right)[s(1-s)]^2,\quad a(s)=\left(\frac{2}{s(1-s)}\right)^2.$$

Proof. We first compute the derivatives of f. Let $g(s)=-s(1-s)$ and $h(y)=\exp(1/y)$, then $f'(s)=c_e^{-1}h(g(s))$. By the chain rule of high order derivatives (also known as Faà di Bruno's formula),

$$f^{(k+1)}(s)=c_e^{-1}\sum\frac{k!}{m_1!1!^{m_1}m_2!2!^{m_2}\cdots m_k!k!^{m_k}}h^{(m_1+m_2+\cdots+m_k)}(g(s))\prod_{j=1}^{k}\left(g^{(j)}(s)\right)^{m_j}$$

where the sum is taken over all k-tuples of non-negative integers $(m_1,\cdots,m_k)$ satisfying

$\sum_{j=1}^{k} jm_j = k$. Note that $g^{(j)}(s) = 0$ for $j \geq 3$, the sum becomes

$$\begin{aligned} f^{(k+1)}(s) &= c_e^{-1} \sum_{m_1+2m_2=k} \frac{k!}{m_1!1!^{m_1} m_2!2!^{m_2}} h^{(m_1+m_2)}(g(s)) \left(g^{(1)}(s)\right)^{m_1} \left(g^{(2)}(s)\right)^{m_2} \\ &= c_e^{-1} \sum_{m_1+2m_2=k} \frac{k!}{m_1!m_2!2^{m_2}} h^{(m_1+m_2)}(g(s)) \left(2s-1\right)^{m_1} 2^{m_2} \\ &= c_e^{-1} \sum_{m_1+2m_2=k} \frac{k!}{m_1!m_2!} h^{(m_1+m_2)}(g(s)) \left(2s-1\right)^{m_1}. \end{aligned}$$

To compute the derivatives of h, we use the chain rule again to get (the sum is over $\sum_{j=1}^{m} jn_j = m$)

$$\begin{aligned} h^{(m)}(y) &= \sum \frac{m!}{n_1!1!^{n_1} n_2!2!^{n_2} \cdots n_m!m!^{n_m}} \exp(1/y) \prod_{j=1}^{m} \left(\frac{d^j(1/y)}{dy^j}\right)^{n_j} \\ &= \sum \frac{m!}{n_1!1!^{n_1} n_2!2!^{n_2} \cdots n_m!m!^{n_m}} \exp(1/y) \prod_{j=1}^{m} \left((-1)^j j! y^{-j-1}\right)^{n_j} \\ &= \sum \frac{(-1)^m m!}{n_1!n_2! \cdots n_m!} \exp(1/y) y^{-m-\sum n_j} \end{aligned}$$

Since $0 \leq n_j \leq m/j$, the number of tuples $(m_1, \cdots, m_n)$ is less than $(m+1)(m/2+1)(m/3+1)\cdots(m/m+1) = \binom{2m}{m} < 2^{2m}$, so for $0 < y < 1$ and $m \leq k$ we have

$$|h^{(m)}(y)| \leq 2^{2k} k! \exp(1/y) y^{-2k}.$$

Therefore $f^{(k+1)}$ can be bounded as

$$\begin{aligned} |f^{(k+1)}(s)| &\leq c_e^{-1} \sum_{m_1+2m_2=k} \frac{k!}{m_1!m_2!} 2^{2k} k! \exp(-\frac{1}{s(1-s)}) \left(\frac{1}{s(1-s)}\right)^{2k} |2s-1|^{m_1} \\ &\leq c_e^{-1} \exp(-\frac{1}{s(1-s)}) \left(\frac{2}{s(1-s)}\right)^{2k} (k!)^2 \sum_{m_1 \leq k} \frac{1}{m_1!} \\ &\leq e c_e^{-1} \exp(-\frac{1}{s(1-s)}) \left(\frac{2}{s(1-s)}\right)^{2k} (k!)^2. \end{aligned}$$

Substitute $k+1$ by k and for every $k \geq 1$

$$\begin{aligned} |f^{(k)}(s)| &\leq e c_e^{-1} \exp\left(-\frac{1}{s(1-s)}\right) \left(\frac{2}{s(1-s)}\right)^{2(k-1)} ((k-1)!)^2 \\ &\leq 4e c_e^{-1} \exp\left(-\frac{1}{s(1-s)}\right) \left(\frac{2}{s(1-s)}\right)^{2(k-1)} \frac{(k!)^2}{(k+1)^2}. \end{aligned}$$

Note that $\|H_0\| \leq 1, \|H_1\| \leq 1$ and $H^{(k)} = (H_1 - H_0)f^{(k)}$, we complete the proof of bounds for $H^{(k)}$. □

The following result demonstrate that E_j $(j \geq 1)$ vanish on the boundary.

Lemma 64. *(a) For all $k \geq 1$, $E_0^{(k)}(0) = P_0^{(k)}(0) = 0, E_0^{(k)}(1) = P_0^{(k)}(1) = 0$.*
(b) For all $j \geq 1, k \geq 0$, $E_j^{(k)}(0) = E_j^{(k)}(1) = 0$.

Proof. We will repeatedly use the fact that $R^{(k)}(0) = R^{(k)}(1) = 0$. This can be proved by taking the kth order derivative of the equation $(H - z)R = I$ and

$$R^{(k)} = -R\sum_{l=1}^{k}\binom{k}{l}(H-z)^{(l)}R^{(k-l)} = -R\sum_{l=1}^{k}\binom{k}{l}H^{(l)}R^{(k-l)}.$$

(a) This is a straightforward result by the definition of E_0 and the fact that $R^{(k)}$ vanish on the boundary.
(b) We prove by induction with respect to j. For $j = 1$, Eq. (6.11.6) tells that

$$E_1 = (2\pi)^{-1}\oint_\Gamma R[P_0^{(1)}, P_0]Rdz.$$

Therefore each term in the derivatives of E_1 must involve at least one of the derivative of R and the derivative of P_0, which means the derivatives of E_1 much vanish on the boundary.

Assume the conclusion holds for $< j$, then for j, first each term of the derivatives of S_j much involve the derivative of some E_m with $m < j$, which means the derivatives of S_j much vanish on the boundary. Furthermore, for the similar reason, Eq. (6.11.6) tells that the derivatives of E_j must vanish on the boundary. □

Before we process, we recall three technical lemmas introduced in [121, 62]. Throughout let $c_f = 4\pi^2/3$ denote an absolute constant.

Lemma 65. *Let $\alpha > 0$ be a positive real number, p, q be non-negative integers and $r = p+q$. Then*

$$\sum_{l=0}^{k}\binom{k}{l}\frac{[(l+p)!(k-l+q)!]^{1+\alpha}}{(l+p+1)^2(k-l+q+1)^2} \leq c_f\frac{[(k+r)!]^{1+\alpha}}{(k+r+1)^2}.$$

Lemma 66. *Let k be a non-negative integer, then*

$$\sum_{l=0}^{k}\frac{1}{(l+1)^2(k+1-l)^2} \leq c_f\frac{1}{(k+1)^2}.$$

Lemma 67. *Let $A(s), B(s)$ be two smooth matrix-valued function defined on $[0,1]$ satisfying*

$$\|A^{(k)}(s)\| \leq a_1(s)a_2(s)^k \frac{[(k+p)!]^{1+\alpha}}{(k+1)^2}, \quad \|B^{(k)}(s)\| \leq b_1(s)b_2(s)^k \frac{[(k+q)!]^{1+\alpha}}{(k+1)^2}$$

for some non-negative functions a_1, a_2, b_1, b_2, non-negative integers p, q and for all $k \geq 0$. Then for every $k \geq 0, 0 \leq s \leq 1$,

$$\|(A(s)B(s))^{(k)}\| \leq c_f a_1(s)b_1(s) \max\{a_2(s), b_2(s)\}^k \frac{[(k+r)!]^{1+\alpha}}{(k+1)^2}$$

where $r = p + q$.

Next we bound the derivatives of the resolvent. This bound provides the most important improvement of the general adiabatic bound.

Lemma 68. *For all $k \geq 0$,*

$$\|R^{(k)}(z, s_0, s_0)\| \leq \frac{2}{\Delta(s_0)} \left(D \log^2 \kappa\right)^k \frac{(k!)^4}{(k+1)^2}$$

where

$$D = c_f \frac{2048\sqrt{2}e^2}{c_e}$$

.

Proof. We prove by induction, and for simplicity we will omit explicit dependence on arguments z, s, s_0. The estimate obviously holds for $k = 0$. Assume the estimate holds for $< k$. Take the kth order derivative of the equation $(H - z)R = I$ and we get

$$R^{(k)} = -R \sum_{l=1}^{k} \binom{k}{l} (H - z)^{(l)} R^{(k-l)} = -R \sum_{l=1}^{k} \binom{k}{l} H^{(l)} R^{(k-l)}.$$

Using Lemma 63 and induction hypothesis, we have

$$\|R^{(k)}\|_2 \leq \frac{2}{\Delta} \sum_{l=1}^{k} \binom{k}{l} b a^l \frac{(l!)^2}{(l+1)^2} \frac{2}{\Delta} \left(D \log^2 \kappa\right)^{k-l} \frac{[(k-l)!]^4}{(k-l+1)^2}$$

To proceed we need to bound the term $\Delta^{-1} b a^l$ for $l \geq 1$. Let us define

$$F(s) = \frac{c_e}{2^{2l} 2\sqrt{2} e} \Delta_*^{-1}(s) b(s) a(s)^l = \frac{\exp(-\frac{1}{s(1-s)})}{(1 - f(s) + f(s)/\kappa)[s(1-s)]^{2l-2}}.$$

Note that $F(0) = F(1) = 0, F(s) > 0$ for $s \in (0,1)$ and $F(1/2+t) > F(1/2-t)$ for $t \in (0,1/2)$, then there exists a maximizer $s_* \in [1/2,1)$ such that $F(s) \le F(s_*), \forall s \in [0,1]$. Furthermore, $F'(s_*) = 0$. Now we compute the F' as

$$\begin{aligned}
&[(1-f+f/\kappa)[s(1-s)]^{2l-2}]^2 F'(s) \\
=&\exp\left(-\frac{1}{s(1-s)}\right)\frac{1-2s}{s^2(1-s)^2}(1-f+f/\kappa)[s(1-s)]^{2l-2} \\
&-\exp\left(-\frac{1}{s(1-s)}\right)\Big[(-f'+f'/\kappa)[s(1-s)]^{2l-2} \\
&\qquad +(1-f+f/\kappa)(2l-2)[s(1-s)]^{2l-3}(1-2s)\Big] \\
=&\exp\left(-\frac{1}{s(1-s)}\right)[s(1-s)]^{2l-4} \\
&\times\Big[(1-f+f/\kappa)(1-2s)[1-(2l-2)s(1-s)] \\
&\qquad -\exp\left(-\frac{1}{s(1-s)}\right)c_e^{-1}(-1+1/\kappa)s^2(1-s)^2\Big] \\
=&\exp\left(-\frac{1}{s(1-s)}\right)[s(1-s)]^{2l-4}G(s)
\end{aligned}$$

where

$$G(s) = (1-f+f/\kappa)(1-2s)[1-(2l-2)s(1-s)] + \exp\left(-\frac{1}{s(1-s)}\right)c_e^{-1}(1-1/\kappa)s^2(1-s)^2.$$

The sign of $F'(s)$ for $s \in (0,1)$ is the same as the sign of $G(s)$.

We now show that s_* cannot be very close to 1. Precisely, we will prove that for all $s \in [1-\frac{c}{l\log\kappa}, 1)$ with $c = \sqrt{c_e}/4 \approx 0.021$, $G(s) < 0$. For such s, we have

$$1-f+f/\kappa \ge f(1/2)/\kappa > 0,$$

$$1-2s < -1/2,$$

and

$$1-(2l-2)s(1-s) \ge 1-(2l-2)(1-s) \ge 1-\frac{2c}{\log\kappa} \ge 1/2,$$

then

$$(1-f+f/\kappa)(1-2s)[1-(2l-2)s(1-s)] \le -\frac{f(1/2)}{4\kappa} = -\frac{1}{8\kappa}.$$

On the other hand,

$$\begin{aligned}\exp\left(-\frac{1}{s(1-s)}\right) &\le \exp\left(-(1-\frac{c}{l\log\kappa})^{-1}\frac{l\log\kappa}{c}\right)\\ &= \kappa^{-(1-\frac{c}{l\log\kappa})^{-1}\frac{l}{c}}\\ &\le \kappa^{-l/c}\\ &\le \kappa^{-1},\end{aligned}$$

then

$$\begin{aligned}&\exp\left(-\frac{1}{s(1-s)}\right)c_e^{-1}(1-1/\kappa)s^2(1-s)^2\\ \le&\frac{1}{\kappa}\frac{1}{c_e}\left(\frac{c}{l\log\kappa}\right)^2\\ \le&\frac{1}{16\kappa}.\end{aligned}$$

Therefore for all $s \in [1-\frac{c}{l\log\kappa}, 1]$ we have $G(s) \le -1/(16\kappa) < 0$, which indicates $s_* \le 1-\frac{c}{l\log\kappa}$.

We are now ready to bound $F(s)$. From the equation $G(s_*) = 0$, we get

$$\frac{\exp\left(-\frac{1}{s_*(1-s_*)}\right)}{1-f+f/\kappa} = \frac{(1-2s_*)[1-(2l-2)s_*(1-s_*)]}{c_e^{-1}(-1+1/\kappa)s_*^2(1-s_*)^2},$$

which gives

$$\begin{aligned}F(s) &\le F(s_*)\\ &= \frac{(1-2s_*)[1-(2l-2)s_*(1-s_*)]}{c_e^{-1}(-1+1/\kappa)[s_*(1-s_*)]^{2l}}\\ &\le \frac{2s_*-1}{c_e^{-1}(1-1/\kappa)[s_*(1-s_*)]^{2l}}\\ &\le 2c_e\cdot 2^{2l}(1-s_*)^{-2l}\\ &\le 2c_e\cdot 2^{2l}\left(\frac{l\log\kappa}{c}\right)^{2l}\\ &= 2c_e\left(\frac{64}{c_e}\right)^l(\log\kappa)^{2l}l^{2l}\\ &\le \frac{2c_e}{e^2}\left(\frac{64e^2}{c_e}\right)^l(\log\kappa)^{2l}(l!)^2.\end{aligned}$$

The last inequality comes from the fact $l^l \le e^{l-1} l!$, which can be derived from the fact that

$$\sum_{i=1}^{n} \log i \geqslant \int_1^n \log x \, \mathrm{d}x = n \log n - (n-1).$$

By definition of $F(s)$ we immediately get

$$\Delta^{-1} b a^l \le \frac{2\sqrt{2}e}{c_e} 4^l F \le \frac{4\sqrt{2}}{e} \left(\frac{256e^2}{c_e}\right)^l (\log \kappa)^{2l} (l!)^2.$$

Now we go back to the estimate of $R^{(k)}$. By Lemma 65,

$$\begin{aligned}
\|R^{(k)}\|_2 &\le \frac{2}{\Delta} \sum_{l=1}^{k} \binom{k}{l} b a^l \frac{(l!)^2}{(l+1)^2} \frac{2}{\Delta} \left(D \log^2 \kappa\right)^{k-l} \frac{[(k-l)!]^4}{(k-l+1)^2} \\
&\le \frac{2}{\Delta} \sum_{l=1}^{k} \binom{k}{l} \frac{8\sqrt{2}}{e} \left(\frac{256e^2}{c_e}\right)^l (\log \kappa)^{2l} (l!)^2 \frac{(l!)^2}{(l+1)^2} \left(D \log^2 \kappa\right)^{k-l} \frac{[(k-l)!]^4}{(k-l+1)^2} \\
&\le \frac{2}{\Delta} (D \log^2 \kappa)^k c_f^{-1} \sum_{l=1}^{k} \binom{k}{l} \frac{(l!)^4 [(k-l)!]^4}{(l+1)^2 (k-l+1)^2} \\
&\le \frac{2}{\Delta} (D \log^2 \kappa)^k \frac{(k!)^4}{(k+1)^2}.
\end{aligned}$$

This completes the proof. □

The next lemma is the main technical result, which gives the bound of derivatives of E_j defined in (6.11.4).

Lemma 69. *(a) For all $k \ge 0$,*

$$\|E_0^k\| = \|P_0^k\| \le (D \log^2 \kappa)^k \frac{(k!)^4}{(k+1)^2}. \tag{6.11.10}$$

(b) For all $k \ge 0, j \ge 1$,

$$\|E_j^{(k)}\| \le A_1 A_2^j A_3^k \frac{[(k+j)!]^4}{(k+1)^2 (j+1)^2} \tag{6.11.11}$$

with

$$\begin{aligned}
A_1 &= \frac{1}{2} \left[c_f^2 \left(1 + 2c_f^2\right)\right]^{-1}, \\
A_2 &= A_1^{-1} c_f^3 \frac{16}{\Delta} D \log^2 \kappa, \\
A_3 &= D \log^2 \kappa.
\end{aligned}$$

Remark 70. *The choice of A_1, A_2 can be rewritten as*

$$c_f^3 \frac{16}{\Delta} D \log^2 \kappa = A_1 A_2,$$
$$c_f^2 \left(1 + 2c_f^2\right) A_1 = \frac{1}{2}.$$

Furthermore, using $c_f > 1$, we have

$$c_f^3 \frac{16}{\Delta} \frac{A_3}{A_2} = A_1 \leq \frac{1}{2}.$$

These relations will be used in the proof later.

Proof. (a) By Lemma 68,

$$\|P_0^k(s_0)\| = \|(2\pi \mathrm{i})^{-1} \oint_{\Gamma(s_0)} R^{(k)}(z, s_0, s_0) dz\| \leq (D \log^2 \kappa)^k \frac{(k!)^4}{(k+1)^2}$$

(b) We prove by induction with respect to j. For $j = 1$, Eq. (6.11.6) tells

$$\|E_1^{(k)}\| = \|(2\pi)^{-1} \oint_{\Gamma} \frac{d^k}{ds^k}(R[P_0^{(1)}, P_0]R) dz\| \leq \frac{\Delta}{2} \|\frac{d^k}{ds^k}(R[P_0^{(1)}, P_0]R)\|.$$

By Lemma 67 and Lemma 68,

$$\begin{aligned}\|E_1^{(k)}\| &\leq \Delta c_f^3 \left(\frac{2}{\Delta}\right)^2 D \log^2 \kappa (D \log^2 \kappa)^k \frac{[(k+1)!]^4}{(k+1)^2} \\ &\leq A_1 A_2 A_3^k \frac{[(k+1)!]^4}{(k+1)^2 (1+1)^2}.\end{aligned}$$

Now assume $< j$ the estimate holds, for j, by Lemma 66, Lemma 67 and induction hypothesis,

$$\begin{aligned}\|S_j^{(k)}\| &\leq \sum_{m=1}^{j-1} c_f A_1 A_2^m A_1 A_2^{j-m} A_3^k \frac{[(k+j)!]^4}{(k+1)^2 (m+1)^2 (j-m+1)^2} \\ &= A_1^2 A_2^j A_3^k \frac{[(k+j)!]^4}{(k+1)^2} c_f \sum_{m=1}^{j-1} \frac{1}{(m+1)^2 (j-m+1)^2} \\ &\leq c_f^2 A_1^2 A_2^j A_3^k \frac{[(k+j)!]^4}{(k+1)^2 (j+1)^2}.\end{aligned}$$

Again by Lemma 67, Lemma 68 and induction hypothesis,

$$\begin{aligned}
\|E_j^{(k)}\| &\leq \|\frac{d^k}{ds^k}\left((2\pi)^{-1}\oint_\Gamma R[E_{j-1}^{(1)},P_0]Rdz\right)\| + \|\frac{d^k}{ds^k}S_j\| + \|\frac{d^k}{ds^k}\left(2P_0S_jP_0\right)\| \\
&\leq \Delta c_f^3\left(\frac{2}{\Delta}\right)^2 A_1A_2^{j-1}A_3\frac{1}{j^2}A_3^k\frac{[(k+j)!]^4}{(k+1)^2} + c_f^2A_1^2A_2^jA_3^k\frac{[(k+j)!]^4}{(k+1)^2(j+1)^2} \\
&\quad + 2c_f^2c_f^2A_1^2A_2^j\frac{1}{(j+1)^2}A_3^k\frac{[(k+j)!]^4}{(k+1)^2} \\
&\leq c_f^3\frac{16}{\Delta}A_1A_2^{j-1}A_3^{k+1}\frac{[(k+j)!]^4}{(k+1)^2(j+1)^2} + c_f^2A_1^2A_2^jA_3^k\frac{[(k+j)!]^4}{(k+1)^2(j+1)^2} \\
&\quad + 2c_f^4A_1^2A_2^jA_3^k\frac{[(k+j)!]^4}{(k+1)^2(j+1)^2} \\
&= \left[c_f^3\frac{16}{\Delta}\frac{A_3}{A_2} + c_f^2\left(1+2c_f^2\right)A_1\right]\times\left[A_1A_2^jA_3^k\frac{[(k+j)!]^4}{(k+1)^2(j+1)^2}\right] \\
&\leq A_1A_2^jA_3^k\frac{[(k+j)!]^4}{(k+1)^2(j+1)^2}.
\end{aligned}$$

□

6.12 Discussion

By reformulating QLSP into an eigenvalue problem, AQC provides an alternative route to solve QLSP other than those based on phase estimation (such as HHL) and those based on approximation of matrix functions (such as LCU and QSP). However, the scaling of the vanilla AQC is at least $\mathcal{O}(\kappa^2/\epsilon)$, which is unfavorable with respect to both κ and ϵ. Thanks to the explicit information of the energy gap along the adiabatic path, we demonstrate that we may reschedule the AQC and dramatically improve the performance of AQC for solving QLSP. When the target accuracy ϵ is relatively large, the runtime complexity of the AQC(p) method ($1 < p < 2$) is reduced to $\mathcal{O}(\kappa/\epsilon)$; for highly accurate calculations with a small ϵ, the AQC(exp) method is more advantageous, and its runtime complexity is $\mathcal{O}(\kappa\ \mathrm{poly}(\log(\kappa/\epsilon)))$. Due to the close connection between AQC and QAOA, the runtime complexity of QAOA for solving QLSP is also bounded by $\mathcal{O}(\kappa\ \mathrm{poly}(\log(\kappa/\epsilon)))$. Both AQC and QAOA can be implemented on gate-based quantum computers.

Our numerical results can be summarized using the following relation:

$$\text{QAOA} \lesssim \text{AQC(exp)} \lesssim \text{AQC}(p) < \text{RM} < \text{vanilla AQC}.$$

Here $A < B$ means that the runtime of A is smaller than that of B. The two exceptions are: QAOA $\lesssim$ AQC(exp) means that the runtime of QAOA is smaller only when the optimizer

θ is found, while AQC(exp) $\lesssim$ AQC(p) holds only when ϵ is sufficiently small. While the runtime complexity of AQC(exp) readily provides an upper bound of the runtime complexity of QAOA, numerical results indicate that the optimizer of QAOA often involves a much smaller depth and hence the dynamics of QAOA does not necessarily follow the adiabatic path. It is therefore of interest to find alternative routes to directly prove the scaling of the QAOA runtime without relying on AQC. Based on the wide range of applications of linear systems and the simplicity of the AQC scheme, we think that our AQC based QLSP solver can be a useful subroutine for future design of quantum algorithms.

Bibliography

[1] G. R. Ahokas. *Improved algorithms for approximate quantum Fourier transforms and sparse Hmailtonian simulations*. University of Calgary, 2004.

[2] T. Albash and D. A. Lidar. "Adiabatic quantum computation". *Rev. Mod. Phys.* 90 (2018), p. 015002.

[3] A. Ambainis. "Variable time amplitude amplification and quantum algorithms for linear algebra problems". *STACS'12 (29th Symposium on Theoretical Aspects of Computer Science)*. Vol. 14. 2012, pp. 636–647.

[4] D. An, D. Fang, and L. Lin. "Time-dependent unbounded Hamiltonian simulation with vector norm scaling". *Quantum* 5 (2021), p. 459.

[5] D. An and L. Lin. "Quantum dynamics with the parallel transport gauge". *Multiscale Model. Simul.* 18 (2020), p. 612.

[6] D. An, D. Fang, and L. Lin. "Parallel transport dynamics for mixed quantum states with applications to time-dependent density functional theory" (2021). arXiv: 2105.14755.

[7] D. An and L. Lin. "Quantum linear system solver based on time-optimal adiabatic quantum computing and quantum approximate optimization algorithm" (2020). arXiv: 1909.05500.

[8] D. G. Anderson. "Iterative procedures for nonlinear integral equations". *J. Assoc. Comput. Mach.* 12 (1965), pp. 547–560.

[9] J. E. Avron and A. Elgart. "Adiabatic Theorem without a Gap Condition". *Commun. Math. Phys.* 203.2 (1999), pp. 445–463.

[10] S. Bachmann, W. D. Roeck, and M. Fraas. "The adiabatic theorem and linear response theory for extended quantum systems". *Commun. Math. Phys.* 361 (2018), p. 997.

[11] S. Bao, S. Kleer, R. Wang, and A. Rahmani. "Optimal control of superconducting gmon qubits using Pontryagin's minimum principle: Preparing a maximally entangled state with singular bang-bang protocols". *Phys. Rev. A* 97 (2018), p. 062343.

[12] W. Bao, S. Jin, and P. A. Markowich. "On time-splitting spectral approximations for the Schrödinger equation in the semiclassical regime". *J. Comput. Phys.* 175.2 (2002), pp. 487–524.

[13] D. W. Berry. "High-order quantum algorithm for solving linear differential equations". *J. Phys. A: Math. Theor.* 47 (2014), p. 105301.

[14] D. W. Berry, G. Ahokas, R. Cleve, and B. C. Sanders. "Efficient quantum algorithms for simulating sparse Hamiltonians". *Commun. Math. Phys.* 270.2 (2007), pp. 359–371.

[15] D. W. Berry and A. M. Childs. "Black-box Hamiltonian simulation and unitary implementation". *Quantum Information & Computation* 12.1-2 (2012), pp. 29–62.

[16] D. W. Berry, A. M. Childs, R. Cleve, R. Kothari, and R. D. Somma. "Simulating Hamiltonian dynamics with a truncated Taylor series". *Phys. Rev. Lett.* 114 (2015), p. 090502.

[17] D. W. Berry, A. M. Childs, and R. Kothari. "Hamiltonian simulation with nearly optimal dependence on all parameters". *Proceedings of the 56th IEEE Symposium on Foundations of Computer Science* (2015), pp. 792–809.

[18] D. W. Berry, A. M. Childs, Y. Su, X. Wang, and N. Wiebe. "Time-dependent Hamiltonian simulation with L^1-norm scaling". *Quantum* 4 (2020), p. 254.

[19] D. W. Berry, A. M. Childs, R. Cleve, R. Kothari, and R. D. Somma. "Exponential improvement in precision for simulating sparse Hamiltonians". *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*. 2014, pp. 283–292.

[20] S. Boixo and R. D. Somma. "Necessary condition for the quantum adiabatic approximation". *Phys. Rev. A* 81 (2010), p. 032308.

[21] S. Boixo, E. Knill, and R. D. Somma. "Eigenpath traversal by phase randomization." *Quantum Info. Comput.* 9 (2009), pp. 833–855.

[22] M. Born and V. A. Fock. "Beweis des Adiabatensatzes". *Zeitschrift für Physik A.* 3-4 (1928), p. 51.

[23] M. Born and R. Oppenheimer. "Zur quantentheorie der molekeln". *Ann. Phys. (Leipzig)* 389 (1927), pp. 457–484.

[24] F. A. Bornemann and C. Schütte. "On the singular limit of the quantum-classical molecular dynamics model". *J. Appl. Math.* 59 (1999), pp. 1208–1224.

[25] J. Bourgain. "On growth of sobolev norms in linear schrödinger equations with smooth time dependent potential". *Journal d'Analyse Mathématique* 77.1 (1999), pp. 315–348.

[26] C. Bravo-Prieto, R. LaRose, M. Cerezo, Y. Subasi, L. Cincio, and P. J. Coles. "Variational Quantum Linear Solver" (2020). arXiv: `1909.05820`.

[27] C. L. Bris. *Handbook of Numerical Analysis, Special Volume Computational Chemistry*. First. Elsevier, 2003.

[28] A. Bruner, S. Hernandez, F. Mauger, P. M. Abanador, D. J. LaMaster, M. B. Gaarde, K. J. Schafer, and K. Lopata. "Attosecond charge migration with TDDFT: Accurate dynamics from a well-defined initial state". *J. Phys. Chem. Let.*. 8.17 (2017), pp. 3991–3996.

[29] M. Bukov, A. G. Day, D. Sels, P. Weinberg, A. Polkovnikov, and P. Mehta. "Reinforcement Learning in Different Phases of Quantum Control". *Phys. Rev. X* 8.3 (2018), p. 031086.

[30] E. Campbell. "Random compiler for fast Hamiltonian simulation". *Phys. Rev. Lett.* 123.7 (2019), p. 070503.

[31] H. Candy and W. Rozmus. "A symplectic integration algorithm for seperate Hamiltonian functions". *J. Comput. Phys.* 92 (1991), pp. 230–256.

[32] Y. Cao and J. Lu. "Stochastic dynamical low-rank approximation method". *J. Comput. Phys.* 372 (2018), pp. 564–586.

[33] Y. Cao, A. Papageorgiou, I. Petras, J. Traub, and S. Kais. "Quantum algorithm and circuit design solving the Poisson equation". *New J. Phys.* 15.1 (2013), p. 013021.

[34] Z. Cao and A. Elgart. "On the efficiency of Hamiltonian-based quantum computation for low-rank matrices". *J. Math. Phys* 53 (2012), p. 032201.

[35] R. Carles and C. Fermanian-Kammerer. "A nonlinear adiabatic theorem for coherent states". *Nonlinearity* 1 (2011), p. 24.

[36] A. Castro, M. Marques, and A. Rubio. "Propagators for the time-dependent Kohn-Sham equations". *J. Chem. Phys.* 121 (2004), pp. 3425–33.

[37] S. Chakraborty, A. Gilyén, and S. Jeffery. "The Power of Block-Encoded Matrix Powers: Improved Regression Techniques via Faster Hamiltonian Simulation". *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)* 132 (2019), 33:1–33:14.

[38] C.-F. Chen, H.-Y. Huang, R. Kueng, and J. A. Tropp. *Concentration for random product formulas*. 2021. arXiv: `2008.11751`.

[39] Z. Chen and E. Polizzi. "Spectral-based propagation schemes for time-dependent quantum systems with application to carbon nanotubes". *Phys. Rev. B* 82.20 (2010), p. 205410.

[40] A. M. Childs, R. Cleve, E. Deotto, E. Farhi, S. Gutmann, and D. A. Spielman. "Exponential algorithmic speedup by a quantum walk". *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*. 2003, pp. 59–68.

[41] A. M. Childs, A. Ostrander, and Y. Su. "Faster quantum simulation by randomization". *Quantum* 3 (2019), p. 182.

[42] A. M. Childs and Y. Su. "Nearly optimal lattice simulation by product formulas". *Phys. Rev. Lett.* 123.5 (2019), p. 050503.

[43] A. M. Childs, R. Kothari, and R. D. Somma. "Quantum Algorithm for Systems of Linear Equations with Exponentially Improved Dependence on Precision". *SIAM J. Comput.* 46 (2017), pp. 1920–1950.

[44] A. M. Childs, Y. Su, M. C. Tran, N. Wiebe, and S. Zhu. "Theory of Trotter Error with Commutator Scaling". *Phys. Rev. X* 11 (2021), p. 011020.

[45] A. M. Childs and J. Young. "Optimal state discrimination and unstructured search in nonlinear quantum mechanics". *Phys. Rev. A* 93.2 (2016).

[46] D. Cohen, T. Jahnke, K. Lorenz, and C. Lubich. "Numerical integrators for highly oscillatory Hamiltonian systems: a review". *Analysis, modeling and simulation of multiscale problems*. Springer, 2006, pp. 553–576.

[47] H. D. Cornean, D. Monaco, and S. Teufel. "Wannier functions and Z2 invariants in time-reversal symmetric topological insulators". *Rev. Math. Phys.* 29 (2017), p. 1730001.

[48] W. van Dam, M. Mosca, and U. Vazirani. "How Powerful is Adiabatic Quantum Computation?" *Proceedings of the 42nd Annual Symposium on Foundations of Computer Science* (2001), pp. 279–287.

[49] C. M. A. Dantas, I. A. Pedrosa, and B. Baseia. "Harmonic oscillator with time-dependent mass and frequency and a perturbative potential". *Physical Review A* 45.3 (1992), pp. 1320–1324.

[50] S. Descombes and M. Thalhammer. "An exact local error representation of exponential operator splitting methods for evolutionary problems and applications to linear Schrödinger equations in the semi-classical regime". *BIT Numer. Math.* 50.4 (2010), pp. 729–749.

[51] P. A. M. Dirac. "Note on Exchange Phenomena in the Thomas Atom". *Proc. Cambridge Philos. Soc.* 26 (1930), p. 376.

[52] A. Dranov, j. Kellendonk, and R. Seiler. "Discrete time adiabatic theorems for quantum mechanical systems". *J. Math. Phys.* 39 (1998), p. 1340.

[53] A. Elgart and G. A. Hagedorn. "A note on the switching adiabatic theorem". *J. Math. Phys* 53 (2012), p. 102202.

[54] P. Elliott and N. T. Maitra. "Propagation of initially excited states in time-dependent density-functional theory". *Phys. Rev. A* 85.5 (2012), p. 052510.

[55] E. Farhi, J. Goldstone, and S. Gutmann. "A Quantum Approximate Optimization Algorithm" (2014). arXiv: 1411.4028.

[56] M. Feng. "Complete solution of the Schrödinger equation for the time-dependent linear potential". *Phys. Rev. A* 64.3 (2001), p. 034101.

[57] C. Fermanian-Kammerer and A. Joye. "A nonlinear quantum adiabatic approximation". *Nonlinearity* 33.9 (2020), pp. 4715–4751.

[58] A. L. Fetter and J. D. Walecka. *Quantum theory of many-particle systems*. Courier Corp., 2003.

[59] R. P. Feynman. "Simulating physics with computers". *Int. J. Theor. Phys* 21 (1982), p. 467.

[60] S. A. Fischer, C. J. Cramer, and N. Govind. "Excited state absorption from real-time time-dependent density functional theory". *J. Chem. Theory Comput.* 11.9 (2015), pp. 4294–4303.

[61] Z. Gang and P. Grech. "Adiabatic theorem for the Gross–Pitaevskii equation". *Commun. Partial. Differ. Equ.* 42.5 (2017), pp. 731–756.

[62] Y. Ge, A. Molnár, and J. I. Cirac. "Rapid Adiabatic Preparation of Injective Projected Entangled Pair States and Gibbs States". *Phys. Rev. Lett.* 116 (2016), p. 080503.

[63] P. Giannozzi et al. "QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials". *J. Phys.: Condens. Matter* 21 (2009), pp. 395502–395520.

[64] A. Gilyén, Y. Su, G. H. Low, and N. Wiebe. "Quantum singular value transformation and beyond: exponential improvements for quantum matrix arithmetics". *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. 2019, pp. 193–204.

[65] A. Gómez Pueyo, M. A. Marques, A. Rubio, and A. Castro. "Propagators for the time-dependent Kohn-Sham equations: multistep, Runge-Kutta, exponential Runge-Kutta, and commutator free Magnus methods". *J. Chem. Theory. Comput.* (2018).

[66] L. K. Grover. "A fast quantum mechanical algorithm for database search". *Proceedings, 28th Annual ACM Symposium on the Theory of Computing (STOC)* (1996), pp. 212–219.

[67] G. A. Hagedorn and A. Joye. "Elementary exponential error estimates for the adiabatic approximation". *J. Math. Anal. Appl.* 267.1 (2002), pp. 235–246.

[68] E. Hairer, C. Lubich, and G. Wanner. *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations*. Second. Springer-Verlag Berlin Heidelberg, 2006.

[69] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving ordinary differential equation I: nonstiff problems*. Vol. 8. Springer, 1987.

[70] E. Hairer and G. Wanner. *Solving ordinary differential equation II: stiff and differential-algebraic problems*. Vol. 8. Springer, 1991.

[71] D. R. Hamann. "Optimized norm-conserving Vanderbilt pseudopotentials". *Phys. Rev. B* 88 (2013), p. 085117.

[72] L. Hari. "Coherent States for Systems of L^2-supercritical Nonlinear Schrödinger Equations". *Commun. Partial. Differ. Equ.* 529 (2013), p. 38.

[73] A. W. Harrow, A. Hassidim, and S. Lloyd. "Quantum algorithm for linear systems of equations". *Phys. Rev. Lett.* 103 (2009), p. 150502.

[74] J. Heyd, G. E. Scuseria, and M. Ernzerhof. "Hybrid functionals based on a screened Coulomb potential". *J. Chem. Phys.* 118.18 (2003), pp. 8207–8215.

[75] M. Hochbruck and C. Lubich. "On Magnus integrators for time-dependent Schrödinger equations". *SIAM J. Numer. Anal.* 41.3 (2003), pp. 945–963.

[76] W. Hu, L. Lin, and C. Yang. "DGDFT: A massively parallel method for large scale density functional theory calculations". *J. Chem. Phys.* 143 (2015), p. 124110.

[77] W. Humphrey, A. Dalke, and K. Schulten. "VMD – Visual Molecular Dynamics". *J. Molec. Graphics* 14 (1996), pp. 33–38.

[78] J. Huyghebaert and H. De Raedt. "Product formula methods for time-dependent Schrödinger problems". *J. Phys. A* 23.24 (1990), pp. 5777–5793.

[79] A. Iserles. *A first course in the numerical analysis of differential equations*. 44. Cambridge Univ. Pr., 2009.

[80] T. Jahnke and C. Lubich. "Numerical integrators for quantum dynamics close to the adiabatic limit". *Numer. Math.* 94 (2003), pp. 289–314.

[81] T. Jahnke and C. Lubich. "Error bounds for exponential operator splittings". *BIT* 40.4 (2000), pp. 735–744.

[82] S. Jansen, M.-B. Ruskai, and R. Seiler. "Bounds for the adiabatic approximation with applications to quantum computation". *J. Math. Phys.* 48.10 (2007), p. 102111.

[83] J.-Y. Ji, J. K. Kim, S. P. Kim, and K.-S. Soh. "Exact wave functions and nonadiabatic Berry phases of a time-dependent harmonic oscillator". *Physical Review A* 52.4 (1995), pp. 3352–3355.

[84] J. Jia and J. Huang. "Krylov deferred correction accelerated method of lines transpose for parabolic problems". *J. Comput. Phys.* 227.3 (2008), pp. 1739–1753.

[85] W. Jia, D. An, L.-W. Wang, and L. Lin. "Fast real-time time-dependent density functional theory calculations with the parallel transport gauge". *J. Chem. Theory Comput.* 14 (2018), p. 5645.

[86] W. Jia and L. Lin. "Fast real-time time-dependent hybrid functional calculations with the parallel transport gauge and the adaptively compressed exchange formulation". *Comput. Phys. Commun.* 240 (2019), p. 21.

[87] W. Jia, L.-W. Wang, and L. Lin. "Parallel transport time-dependent density functional theory calculations with hybrid functional on summit". *SC '19 Proceedings of the International Conference for High Performance Computing*. 79. 2019, pp. 1–23.

[88] A. Joye. "Proof of the Landau–Zener formula". *Asymptotic Analysis* 9 (1994), p. 209.

[89] A. Joye, H. Kunz, and C.-E. Pfister. "Exponential Decay and Geometric Aspect of Transition Probabilities in the Adiabatic Limit". *Ann. Phys.* 208 (1991), p. 299.

[90] A. Joye. "An adiabatic theorem for singularly perturbed Hamiltonians". *Ann. Inst. H. Poincaré Phys. Théor.* 63.2 (1995), pp. 231–250.

[91] A. Joye. "General adiabatic evolution with a gap condition". *Comm. Math. Phys.* 275.1 (2007), pp. 139–162.

[92] A.-K. Kassam and L. N. Trefethen. "Fourth-order time-stepping for stiff PDEs". *SIAM J. Sci. Comput.* 26 (2005), pp. 1214–1233.

[93] T. Kato. "On the adiabatic theorem of quantum mechanics". *J. Phys. Soc. J. Jpn.* 5 (1950), pp. 435–439.

[94] C. T. Kelley. *Iterative methods for optimization*. Vol. 18. SIAM, 1999.

[95] I. D. Kivlichan, N. Wiebe, R. Babbush, and A. Aspuru-Guzik. "Bounding the costs of quantum simulation of many-body physics in real space". *J. Phys. A Math. Theor.* 50 (2017), p. 305301.

[96] A. W. Knapp. *Basic Real Analysis*. Springer Science & Business Media, 2005.

[97] O. Koch and C. Lubich. "Dynamical low-rank approximation". *SIAM J. Matrix Anal. Appl.* 29 (2007), pp. 434–454.

[98] A. V. Krasheninnikov, Y. Miyamoto, and D. Tománek. "Role of electronic excitations in ion collisions with carbon nanostructures". *Phys. Rev. Lett.* 99 (2007), p. 016104.

[99] L. D. Landau. "Zur theorie der energieubertragung ii". *Phys. Z. Sowjetunion* 2 (1932), p. 46.

[100] C. Le Bris and P. Rouchon. "Low-rank numerical approximations for high-dimensional Lindblad equations". *Phys. Rev. A* 87.2 (2013), pp. 1–4.

[101] C. Le Bris, P. Rouchon, and J. Roussel. "Adaptive low-rank approximation and denoised Monte Carlo approach for high-dimensional Lindblad equations". *Phys. Rev. A* 92.6 (2015), pp. 1–5.

[102] R. J. LeVeque. *Finite difference methods for ordinary and partial differential equations: Steady-state and time-dependent problems*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007, pp. xvi+341.

[103] X. Li, J. C. Tully, H. B. Schlegel, and M. J. Frisch. "*Ab initio* Ehrenfest dynamics". *J. Chem. Phys.* 123 (2005), p. 084106.

[104] L. Lin, J. Lu, L. Ying, and W. E. "Adaptive local basis set for Kohn-Sham density functional theory in a discontinuous Galerkin framework I: Total energy calculation". *J. Comput. Phys.* 231 (2012), pp. 2140–2154.

[105] L. Lin and C. Yang. "Elliptic preconditioner for accelerating self consistent field iteration in Kohn-Sham density functional theory". *SIAM J. Sci. Comp.* 35 (2013), S277–S298.

[106] L. Lin and Y. Tong. "Optimal polynomial based quantum eigenstate filtering with application to solving quantum linear systems". *Quantum* 4 (2020), p. 361.

[107] J. Liu. "The multifrontal method for sparse matrix solution: Theory and practice". *SIAM Rev.* 34 (1992), pp. 82–109.

[108] G. H. Low and N. Wiebe. "Hamiltonian Simulation in the Interaction Picture" (2019). arXiv: 1805.00675.

[109] G. H. Low and I. L. Chuang. "Optimal Hamiltonian Simulation by Quantum Signal Processing". *Phys. Rev. Lett.* 118 (2017), p. 010501.

[110] G. H. Low, V. Kliuchnikov, and N. Wiebe. *Well-conditioned multiproduct Hamiltonian simulation*. 2019. arXiv: 1907.11679.

[111] C. Lubich. *From quantum to classical molecular dynamics: reduced models and numerical analysis*. European Mathematical Society, 2008.

[112] C. Lubich. "On splitting methods for Schrodinger-Poisson and cubic nonlinear Schrodinger equations". *Math. Comp.* 77 (2008), pp. 2141–2153.

[113] C. Lubich and I. V. Oseledets. "A projector-splitting integrator for dynamical low-rank approximation". *BIT Numer. Math.* 54.1 (2014), pp. 171–188.

[114] Y. Maday and G. Turinici. "New formulations of monotonically convergent quantum control algorithms". *J. Chem. Phys.* 118.18 (2003), pp. 8191–8196.

[115] O. B. Malcıoğlu, R. Gebauer, D. Rocca, and S. Baroni. "turboTDDFT–A code for the simulation of molecular spectra using the Liouville–Lanczos approach to time-dependent density-functional perturbation theory". *Comput. Phys. Commun.* 182 (2011), pp. 1744–1754.

[116] R. Martin. *Electronic Structure: Basic Theory and Practical Methods*. Cambridge Univ. Pr., 2008.

[117] H.-D. Meyer, U. Manthe, and L. Cederbaum. "The multi-configurational time-dependent Hartree approach". *Chem. Phys. Lett* 165.1 (1990), pp. 73–78.

[118] R. Montalto. "On the growth of Sobolev norms for a class of linear Schrödinger equations on the torus with superlinear dispersion". *Asymptotic Analysis* 108 (2018), pp. 85–114.

[119] M. Nakahara. *Geometry, topology and physics*. CRC Press, 2003.

[120] G. Nenciu. "Existence of the exponentially localised Wannier functions". *Comm. Math. Phys.* 91 (1983), pp. 81–85.

[121] G. Nenciu. "Linear adiabatic theory. Exponential estimates". *Comm. Math. Phys.* 152 (1993), pp. 479–496.

[122] P. Nettesheim, F. A.Bornemann, B. Schmidt, and C. Schütte. "An explicit and symplectic integrator for quantum-classical molecular dynamics". *Chem. Phys. Lett.* 256 (1996), pp. 581–588.

[123] P. Nettesheim and C. Schütte. "Numerical Integrators for Quantum-Classical Molecular Dynamics" (1999). Ed. by P. Deuflhard, J. Hermans, B. Leimkuhler, A. E. Mark, S. Reich, and R. D. Skeel, pp. 396–411.

[124] M. A. Nielsen and I. Chuang. *Quantum computation and quantum information*. 2000.

[125] M. Y. Niu, S. Boixo, V. N. Smelyanskiy, and H. Neven. "Universal quantum control through deep reinforcement learning". *npj Quantum Info.* 5 (2019), p. 33.

[126] G. Onida, L. Reining, and A. Rubio. "Electronic excitations: density-functional versus many-body Green's-function approaches". *Rev. Mod. Phys.* 74 (2002), p. 601.

[127] I. A. Pedrosa. "Exact wave functions of a harmonic oscillator with time-dependent mass and frequency". *Phys. Rev. A* 55.4 (1997), pp. 3219–3221.

[128] I. A. Pedrosa, G. P. Serra, and I. Guedes. "Wave functions of a time-dependent harmonic oscillator with and without a singular perturbation". *Phys. Rev. A* 56.5 (1997), pp. 4300–4303.

[129] J. P. Perdew, K. Burke, and M. Ernzerhof. "Generalized gradient approximation made simple". *Phys. Rev. Lett.* 77 (1996), pp. 3865–3868.

[130] D. A. Rehn, Y. Shen, M. E. Buchholz, M. Dubey, R. Namburu, and E. J. Reed. "ODE integration schemes for plane-wave real-time time-dependent density functional theory". *J. Chem. Phys.* 150.1 (2019), p. 014101.

[131] A. T. Rezakhani, W.-J. Kuo, A. Hamma, D. A. Lidar, and P. Zanardi. "Quantum Adiabatic Brachistochrone". *Phys. Rev. Lett.* 103 (2009), p. 080502.

[132] J. Roland and N. J. Cerf. "Quantum search by local adiabatic evolution". *Phys. Rev. A* 65.4 (2002), p. 042308.

[133] E. Runge and E. K. U. Gross. "Density-functional theory for time-dependent systems". *Phys. Rev. Lett.* 52 (1984), p. 997.

[134] A. Russakoff, Y. Li, S. He, and K. Varga. "Accuracy and computational efficiency of real-time subspace propagation schemes for the time-dependent density functional theory". *J. Chem. Phys.* 144 (2016), p. 204125.

[135] Y. Saad. *Iterative methods for sparse linear systems*. Vol. 82. SIAM, 2003.

[136] B. Şahinoğlu and R. D. Somma. "Hamiltonian simulation in the low energy subspace" (2020). arXiv: 2006.02660.

[137] A. Schleife, E. W. Draeger, Y. Kanai, and A. A. Correa. "Plane-wave pseudopotential implementation of explicit integrators for time-dependent Kohn-Sham equations in large-scale simulations". *J. Chem. Phys.* 137 (2012), 22A546.

[138] M. Schlipf and F. Gygi. "Optimization algorithm for the generation of ONCV pseudopotentials". *Comput. Phys. Commun.* 196 (2015), pp. 36–44.

[139] A. Schulze-Halberg. "Form-Preserving Transformations of Time-Dependent Schrödinger Equation with Time- and Position-Dependent Mass". *Commun. Theor. Phys* 43.4 (2005), pp. 657–665.

[140] P. W. Shor. "Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer". *SIAM J. Comput.* 26.5 (1997), pp. 1484–1509.

[141] C. Sparber. "Weakly nonlinear time-adiabatic theory". *Ann. Henri Poincáre* 17 (2016), pp. 913–936.

[142] R. B. Stinchcombe. "Ising model in a transverse field. I. Basic theory". *J. Phys. C: Solid State Phys.* 6 (1973), p. 2459.

[143] Y. Su, H.-Y. Huang, and E. T. Campbell. "Nearly tight Trotterization of interacting electrons". *Quantum* 5 (2021), p. 495.

[144] Y. Subaşı, R. D. Somma, and D. Orsucci. "Quantum Algorithms for Systems of Linear Equations Inspired by Adiabatic Quantum Computing". *Phys. Rev. Lett.* 122 (2019), p. 060504.

[145] M. Suzuki. "General decomposition theory of ordered exponentials". *Proc. Japan Acad.* 69 (1993), pp. 161–166.

[146] Y. Takimoto, F. D. Vila, and J. J. Rehr. "Real-time time-dependent density functional theory approach for frequency-dependent nonlinear optical response in photonic molecules". *J. Chem. Phys.* 127 (2007), p. 154114.

[147] S. Teufel. *Adiabatic perturbation theory in quantum dynamics*. First. Springer-Verlag Berlin Heidelberg, 2003.

[148] S. Teufel. "A note on the adiabatic theorem without gap condition". *Lett. Math. Phys.* 58.3 (2001), 261–266 (2002).

[149] M. Thalhammer. "High-order exponential operator splitting methods for time-dependent Schrödinger equations". *SIAM J. Numer. Anal.* 46.4 (2008), pp. 2022–2038.

[150] J. W. Thomas. *Numerical partial differential equations: finite difference methods*. Vol. 22. Texts in Applied Mathematics. Springer-Verlag, New York, 1995, pp. xx+437.

[151] C. A. Ullrich. *Time-dependent density-functional theory: concepts and applications*. Oxford Univ. Pr., 2011.

[152] Z. Wang, S.-S. Li, and L.-W. Wang. "Efficient Real-Time Time-Dependent Density Functional Theory Method and its Application to a Collision of an Ion with a 2D Material". *Phys. Rev. Lett.* 114 (2015), pp. 1–5.

[153] D. Wecker, M. B. Hastings, N. Wiebe, B. K. Clark, C. Nayak, and M. Troyer. "Solving strongly correlated electron models on a quantum computer". *Phys. Rev. A* 92 (2015), p. 062318.

[154] N. Wiebe and N. S. Babcock. "Improved error-scaling for adiabatic quantum evolutions". *New J. Phys.* 14 (2012), pp. 1–10.

[155] N. Wiebe, D. Berry, P. Høyer, and B. C. Sanders. "Higher order decompositions of ordered operator exponentials". *J. Phys. A* 43.6 (2010), p. 065203.

[156] W. Wootters and W. Zurek. "A Single Quantum Cannot be Cloned". *Nature* 299 (1982), p. 802.

[157] L. Wossnig, Z. Zhao, and A. Prakash. "Quantum Linear System Algorithm for Dense Matrices". *Phys. Rev. Lett.* 120.5 (2018), p. 050502.

[158] X. Xu, J. Sun, S. Endo, Y. Li, S. C. Benjamin, and X. Yuan. "Variational algorithms for linear algebra" (2019). arXiv: 1909.03898.

[159] K. Yabana and G. F. Bertsch. "Time-dependent local-density approximation in real time". *Phys. Rev. B* 54 (1996), pp. 4484–4487.

[160] Z.-C. Yang, A. Rahmani, A. Shabani, H. Neven, and C. Chamon. "Optimizing Variational Quantum Algorithms Using Pontryagin's Minimum Principle". *Phys. Rev. X* 7 (2017), p. 021027.

[161] S.-H. Yeh, A. Manjanath, Y.-C. Cheng, J.-D. Chai, and C.-P. Hsu. "Excitation energies from thermally assisted-occupation density functional theory: Theory and computational implementation". *J. Chem. Phys.* 153.8 (2020), p. 084120.

[162] D. C. Yost, Y. Yao, and Y. Kanai. "Propagation of maximally localized Wannier functions in real-time TDDFT". *J. Chem. Phys.* 150.19 (2019), p. 194113.

[163] C. Zener. "Non-adiabatic crossing of energy levels". *Proc. R. Soc. London Ser. A* 137 (1932), p. 696.

[164] W. Zhu and H. Rabitz. "A rapid monotonically convergent iteration algorithm for quantum optimal control over the expectation value of a positive definite operator". *J. Chem. Phys.* 109.2 (1998), pp. 385–391.