Title
Complexity of Finding and Measuring Ground States in Quantum Systems

Author
Meiburg, Alexander Heinz Zhong

Publication Date
2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Santa Barbara

Complexity of Finding and Measuring Ground States in Quantum Systems

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Physics

by

Alexander Heinz Zhong Meiburg

Committee in Charge:

Dr. Bela Bauer, Microsoft, Co-Chair

Professor Leon Balents, Co-Chair

Professor Andrea Young

June 2023

The Dissertation of
Alexander Heinz Zhong Meiburg is approved:

Professor Andrea Young

Dr. Bela Bauer, Microsoft, Co-Chair

Professor Leon Balents, Co-Chair
June 2023

Complexity of Finding and Measuring Ground States in Quantum Systems

by

Alexander Heinz Zhong Meiburg

Acknowledgements

I owe many thanks to my parents, who both worked not only to provide me with wonderful educational opportunities, but even more importantly to give me a curiosity, love, and joy about learning and exploring mathematics. Without their influence, I believe I would be much less content with my life.

I must thank my wife Lindsay for supporting me, and graciously accepting the adventures of moving with me and accepting the geographical constraints of an academic life. Her support and admiration has empowered me more than she knows. It delights me that her name, which she has so graciously let me take as my own, is also printed here.

My advisor, Bela Bauer, has given me guidance and grounding when I didn't even know what to start asking. He has encouraged me when things have gone well, and just as importantly, helped me accept them when they haven't.

Thanks to Eric Vigoda and Scott Aaronson for their encouragement as this physicist has slowly poked his nose into the world of theoretical computer science, and for their guidance when I'm bewildered by how foreign the customs can be. And thank you to a long list of names from Zapata Computing, who gave me a wonderful summer (and then some) seeing the product side of quantum science, and helped me mentally prepare for life after school from many different perspectives.

Thank you to Adolfo Holguin and David Grabovsky for many late and long discussions about math, even if half the time one (or even both) of us was totally lost.

And thank you to my cats, Zenith, Chester, and Karma, for giving me a wholly different kind of support, always happy to lend a very pink pairs of ears when I needed someone to talk to.

Curriculum Vitæ

Alexander Heinz Zhong Meiburg

Education

2023	Doctor of Philosophy in Physics, University of California, Santa Barbara (Expected).
2021	M.A. in Physics, University of California, Santa Barbara.
2018	B.S. in Math and Physics, California Institute of Technology.

Professional Experience

2019 - 2023	Graduate Student Researcher, University of California, Santa Barbara
2022	Instructor, University of California, Santa Barbara
2022	Quantum AI Research Intern, Zapata Computing
2018, 2019	Research Intern, Facebook Reality Labs
2017 - 2018	Research Technician, NASA Jet Propulsion Laboratory
2015	Summer Undergraduate Research Fellow, California Institute of Technology
2014	Intern, Climate Hazards Group, University of California, Santa Barbara

Primary Publications

- Inapproximability of Positive Semidefinite Permanents and Quantum State Tomography. 2022. IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS).
- Linear-time generalized Hartree-Fock algorithm for quasi-one-dimensional systems. Phys. Rev. Research 4, 023128. *Joint work with Bela Bauer.*
- Quantum Constraint Problems can be complete for BQP, QCMA, and more. arXiv:2101.08381.
- Accelerated Green's Function Reconstruction on Quantum Devices. *In preparation. Joint work with Bela Bauer.*
- Generative Learning of Continuous Data by Tensor Networks. *In preparation. Joint work with Jing Chen, Jacob Miller, and Alejandro Perdomo-Ortiz.*

Abstract

Complexity of Finding and Measuring Ground States in Quantum Systems

Alexander Heinz Zhong Meiburg

Given a quantum system that we seek to understand, finding the ground state is often the first and most informative task we can assume, from which we can then examine dynamics, excitations, entanglement geometry, and so on. But there are several different precise senses in which we could ask to "find the ground state", and depending on the exact system and the question, the resulting task could be quite easy or difficult. This thesis examines four distinct problems, drawing from a toolkit of Bayesian statistics, the Density Matrix Renormalization Group, convex optimization, Fourier analysis, and computational complexity theory.

In one setting, we are given a set of local interactions and asked only, is this Hamiltonian frustrated? This question can be very easy or difficult depending on the types of interactions permitted by the symmetry of the system. We show that in fact there are in fact "natural" (in a precise, mathematical sense) interaction types of many different difficulties, including complexity classes BQP and QCMA. In the second setting, we have repeatedly measured an unknown quantum state, and we are tasked with determining the most probable state given the measurements. We show that this task is in fact exponentially difficult (NP-hard) in the

dimension of the Hilbert space. In the third setting, we examine one-dimensional fermionic systems, and show how Gaussian Fermionic Matrix Product States and DMRG can be combined with Hartree-Fock iteration to find approximate groud states very quickly. In the final setting, we use a quantum computer to perform binary measurements of Green's functions and wish to reconstruct the whole function. We show that although classical statistical techniques give an acceptable reconstruction, imposing physicality constraints greatly enhances the sample efficiency and reconstruction quality.

Contents

List of Figures

Chapter 1

Introduction

The four body chapters of this thesis consider four distinct problems, each asking some form of: how can optimization algorithms help us understand quantum systems, and how difficult is this? We Below we give descriptions of and connections between the four parts.

1.1 Quantum Makes Complexity More Complicated

Our physical world has generously provided us with a wide variety of difficult problems to ponder. The inputs to these problems could be physical constants, experimental data, or human design parameters; the outputs could be, for instance, a configuration of particles, or a calculated energy value, or parameters to some new physical model. Today we can simulate 60 trillion particles interacting gravitationally [93], but solving for 256 binary inputs to a circuit is considered implausible with humanity's computing resources, and finding the ground state energy of a 6x6 Hubbard model is beyond the limit of what we compute exactly. The massive gap in difficulty can only be understood in terms of qualitative differences in the nature of these questions. The goal of computational complexity theory is then categorizing these problems by their difficulty.

In many settings, the distinction of "easy" and "hard" problems is now well-understood, often with sharp theorems delineating the transitions. Often, these correspond to phase transitions in statistical physics, for instance satisfiability transitions [80, 129], costs in resource allocation problems [131], or computational questions about partition functions themselves [61, 91].

Living in a quantum world as we do, many of the classical problems we have studied have interesting and natural quantum generalizations. Sometimes, the quantum variants can be solved by appropriately extending existing solutions to the classical problems. Sometimes they require radically new ideas and approaches. And sometimes, unfortunately, the quantum problems turn out to be exponentially harder than their classical versions – or even just unsolvable (*undecidable*), period. Complexity theory is a rich field, and it turns out that introducing quantum mechanics into the mix makes all of it even more complicated. Quantum mechanics makes complexity even more complicated.

Chapters 2 and 3 of this thesis examine two particular instances of this. Chapter 2 examines *constraint satisfaction problems*, a formal way of discussing "discrete variables with local restrictions on configurations". The algebraic theory of constraint problems has filled many books, and precise classification theorems are known, where only two levels of difficulty can occur. Constraint problems naturally quantize to the problem of *frustration-free Hamiltonians*. The work of

Chapter 2 shows that in addition to four known cases of difficulty for frustration-free Hamiltonians, at least four more are possible – including BQP, the precise power of quantum computing itself. Any putative classification theorem would need to account for all of this!

Chapter 3 studies a form of quantum Bayesian inference. The classical problem is simple: given n samples from some unknown probability distribution P over d elements, what P best explains the samples (maximum likelihood)? Besides having a closed form solution, the convexity of the problem permits efficient and precise solutions even in the presence of complicated priors or constraints. The quantum analog has an unknown d-dimensional mixed state ρ with n measurements (or more generally, POVMs) and we want to reconstruct ρ. Perhaps surprisingly, this problem turns out to be exponentially hard in d (harder than NP, in fact) – and so *doubly* exponentially in the number of particles q, as $d \approx 2^q$ in general. This result, as it turns, is relevant to some purely linear algebraic questions as well.

1.2 Efficiently Solvable Ground States

A second theme is that of trying to find *ground states* of systems efficiently. If a quantum system is defined by its Hamiltonian (or equivalently, a unitary), then the

ground state is the natural extremum of the system. If the world is an optimization problem, then the ground state is surely the solution[1]. Unfortunately, ground states can be very difficult to approximate even for simple systems[77], forcing us to constrain our model in some way.

The fact that the Hamiltonians of Chapter 2 must be *frustration-free* relax our problem, by allowing us to focus on only states that are local ground states as well. This corresponds to the classical statement that "satisfy all the rules" is much easier to reason about than "satisfy as many rules as possible". This stricter, easier optimization problem is easier for us to work with.

One-dimensional systems admit[2]particularly good approximate ground states with easy descriptions. For free-fermion systems, the Gaussian state formalism compresses this further, and these can be combined as in Bauer & Schuch[108] as the GFMPS algorithm. Chapter 4 concerns adapting this to systems that are not free-fermion, using Hartree-Fock iteration. By limiting our search space to Gaussian states of bounded entanglement, we can find our (restricted) ground state in linear time.

[1] There is an exception to be made for physics at nonzero temperature, where the optimization objective is replaced by $H \cdot \rho + T\, S(\rho)$, and the extremum is the thermal state $\rho = Z^{-1} \exp(-\beta H)$.

[2] They admit these descriptions, under some assumptions about local interactions and entanglement entropy. Any proofs of convergence go out the window once we go to free-fermion states and Hartree-Fock iteration, unfortunately; it is unlikely that rigorous error bounds exist.

1.3 Bayesian State Estimation

Probability and uncertainty are essential to any operational definition of quantum mechanics. Scientific inquiry in any probabilistic setting must ultimately employ Bayesian reasoning to model the underlying quantities. Chapters 3 and 5 both address variants of this problem.

Chapter 3 has our experimentalist repeatedly preparing some unknown quantum state (by, for instance, doing an experiment, or running a circuit on a quantum computer) and taking measurements to reconstruct the state. Given their observed data, what should we expect the state to be? This natural question is surprisingly difficult, a difficulty attributable to the nonlinearity of $| < \psi_{obs}|\psi_{experiment} > |^2$. We are able to give a $O(n^d)$ algorithm for computing the answer, though, and show that this is essentially optimal.

Dual to reconstructing states is, of course, reconstructing time evolution. In the setting most realistic for realizing on near-term quantum computers, we have access to a real-time Green's function $G(t)$ for different perturbations on some state of interest ρ. Measurements take a simple mathematical form, as a Bernoulli variable with $p = \frac{1+e^{i\theta}G(t)}{2}$ with a t and θ we can choose. The Bayesian scientist wants to estimate the function $G(t)$ optimally from these measurements, a linear but infinite-dimensional problem. Chapter 5 explores how we can reconstruct $G(t)$

efficiently, both in terms of sample efficiency and computational time, to resolve $G(t)$ accurately and quickly.

With these themes in mind, let us move on to the meat.

Chapter 2

BQP-Complete Frustration-Free Hamiltonians[1]

[1] Based on the work in 10.48550/arXiv.2101.08381

Quantum constraint problems ask whether a Hamiltonian is frustration-free: given a collection of local operators, is there a state that is in the ground state of each operator simultaneously? It has previously been shown that these problems can be in P, NP-complete, MA-complete, or QMA_1-complete, but this list has not been shown to be exhaustive. We show that there are also constraint problems that are BQP_1-complete (also known as coRQP), i.e. they capture all efficient quantum computations. Our construction is constructive and provide the first natural complete problem for BQP_1. We also extend our results to QCMA-complete and coRP-complete CSPs, and show that all quantum constraint problems can be realized on qubits, a trait not shared with classical constraint problems. These results suggest a significant diversity of complexity classes present in quantum constraint problems.

2.1 An Introduction to Constraint Problems

In the classical world of constraint problems, it is known that all constraint problems fall either inside the class P of polynomial-time solvable problems, or are NP-complete and expected to be intractable [132]. In this sense, there is a simple binary classification of their complexity. This classification requires a

certain notion of "natural" constraint problem classes, where constraints can be freely attached between any two pair of variables.

In the quantum setting, constraint problems can be naturally translated as a question about local Hamiltonians. To maintain the same corresponding notion of naturality used in complexity theory, the Hamiltonians allow local terms between any small set of particles. This may seem physically implausible (for instance, we have no *spatial* locality), but provides the simplest setting for studying the complexity of the resulting problems. Our new results are showing that quantum constraint problems can be complete for BQP_1, QCMA_1, and coRP.

We first give precise definitions of our problem settings (Section 2.2) and an overview of results and techniques (Sections 2.3 and 2.4). In Section 2.5 we prove the BQP_1 result by exhibiting a particular language. Although the language is complicated to define, it satisfies our goal of showing the complexity class as possible as a constraint problem. In Sections 2.6, 2.7 we show to modify the BQP_1 language to build languages complete for QCMA_1 and coRP. The remainder are technical notes and future directions.

2.2 Background

2.2.1 Classical constraint problems

The classical notion of *constraint problem* or *constraint satisfaction problem* (CSP) take the form of a *domain* of variable values, and *clauses*: relations on a set of variables. Usually these terms are specifically in reference to *finite* constraint problems, where the domain D is finite, and the clauses C are of bounded arity k. A clause of arity $m \leq k$ is a subset of D^k. We will focus our attention on finite CSPs, referring to them simply as CSPs.

Four representative examples could be 2SAT, 3SAT, 3COLOR, and Mod3 (the set of linear equations in variables mod 3, where each equation has at most 3 variables). A CSP *instance* is a finite number of variables n, and a list of clauses applied to certain variables. The instance is satisfiable iff there is an assignment $A : [n] \to D$ such that, for every clause $c \in C$ applied to variables $(v_1, \ldots v_m)$, the assigned values $(A(v_1), \ldots A(v_m)) \in c$. The problem corresponding to a CSP is determining the satisfiabiliy of its instances. 2SAT and Mod3 are in P, while 3SAT and 3COLOR are NP-Complete, that is, in NPC.

A landmark theorem by Dmitry Zhuk [132], the so-called CSP Dichotomy Theorem, showed that *every* constraint problem is either in P or NPC. This does not rule out the possibility of other NP-Intermediate languages (the set

NP\(P∪NPC)), problems that cannot be expressed as **constraint** problems in particular. Problems such as graph isomorphism or integer factorization are believed to be in this class. A single instance of graph isomorphism (GI) can be easily encoded as a single instance of a constraint problem (such as 3SAT), but then there are other instances of 3SAT that are much harder than mere graph isomorphism. Finding a constraint problem that *only* permitted the expression of GI problems would immediately yield either a polynomial time algorithm for GI, or a subexponential algorithm for NPC, by Zhuk's result. The CSP Dichotomy Theorem also provides a systematically checkable condition for whether a problem is in P or NPC, the existence of a *polymorphism*.

A natural question is how this result might translate to the world of quantum problems. We define the quantum version of constraint problems, and emphasize the distinction from quantum *optimization* problems. We review known results about quantum constraint problems. The main contribution is providing three new quantum constraint problems, that are complete for the classes $\mathsf{BQP}_1 = \mathsf{coRQP}$, QCMA, and coRP. These imply that any putative quantum dichotomy theorem would need at least 7 distinct cases – or a proof that some of these 7 complexity classes are actually equal to one another – in stark contrast to the 2 cases in the classical case.

2.2.2 Quantum Constraint Problems

Quantum Constraint Satisfaction Problems, or QCSPs, can be viewed as a quantum version of a CSP or as a question about frustration free Hamiltonians. A QCSP has a domain size d, and a set of clauses or *interactions* $C = \{\mathcal{H}_i\}$. A clause $\mathcal{H}_i$ of arity m is a Hermitian operator on the space of m many d-qudits, $(\mathbb{C}^d)^{\otimes m}$. We require that each clause is a projector, i.e. $\mathcal{H}_i^2 = \mathcal{H}_i$. For a problem instance on n variables, the interactions H_i extend naturally to operators on the whole Hilbert space. If the arity of all interactions is at most k, then the QCSP is *k-local*. An instance of a QCSP is then a collection of the interactions applied at different qudits. We use fonts to distinguish the clause types of the QCSP $\mathcal{H}_i$, from the particular clauses of the instance H_i. An instance is satisfiable if there is a state $|\psi\rangle \neq 0$ such that $\forall_i H_i |\psi\rangle = 0$, equivalently if the total Hamiltonain $H = \sum_i H_i$ satisfies $H |\psi\rangle = 0$.

This can be viewed physically as the question, is H frustration-free? Frustration-free Hamiltonians have applications in one-way computation [35], and are often easier to study in terms of entanglement structure.

If there is a satisfying assignment (a YES instance), we expect we should be able to verify this state exactly, and accept with probability one. If the ground state has positive energy bounded from below by an inverse polynomial $1/p(n)$, then running $p(n)$ rounds should suffice to detect the failure with high probabil-

ity. But if the true ground state has positive but exponentially small energy, we may be unable to observe this small energy, and erroneously accept. Thus this chapter will discuss a promise problem variants of QCSPs, which excludes super-polynomially small gaps. We are given a constant $b > 1/poly(n)$, and promised that the either the total Hamiltonian has a frustration-free ground state, or that the ground state has energy at least b. All complexity classes mentioned herein are formulated as promise problem classes, with perhaps with trivial promises (P and NP, which requires no promise), and all completeness theorems refer to promise problem completeness. Probabilistic classes such as QMA are semantic classes, not syntactic, which largely precludes the possibility of a non-promise problem being complete; and so virtually all discussion of the classes concerns their promise variants. [30] discusses this distinction in more detail.

2.2.3 Previous Results

We review some known statements about quantum constraint problems.

Proposition 2.1 (Folklore.)**.** *Every classical CSP can be efficiently mapped to a corresponding QCSP, preserving satisfiability.*

This occurs by simply writing each classical clause in the classical basis, where they are diagonal.

Definition 2.1 (k-QSAT [27])**.** *The QCSP k-*QSAT *is the QCSP with $d = 2$ qubits, and with clauses C as the set of all k-local interactions.*

Theorem 2.1 ([27])**.** *2-*QSAT *is in* P*.*

Definition 2.2 (QMA_1 [56])**.** *A language L belongs to the class QMA_1 iff there is a uniform family of quantum circuits U of polynomial size, such that for an input x:*

Perfect Completeness*: If $x \in L$, then there exists a state $|y\rangle$, such that measuring the first qubit of $U|x\rangle|y\rangle$ is 1 with probability 1.*

Soundness*: If $x \notin L$, then for any state $|y\rangle$, such that measuring the first qubit of $U|x\rangle|y\rangle$ is 0 with probability at least $2/3$.*

Theorem 2.2 ([56])**.** *3-*QSAT *is QMA_1 complete.*

Definition 2.3 ((r,s)-QSAT)**.** *The QCSP (r,s)-*QSAT *is most naturally described as having a mixture of r-qudits and s-qudits, with 2-local clauses only between r- and s-qudits. This can be defined in our definition of QCSP, using $d = r + s$, where allowed clauses are 2-local projectors that project onto the first r states in the first qudit, and the last s states in the second qudit.*

Note that $(2,2)$-QSAT is the same as 2-QSAT, and that (r',s')-QSAT is at least as hard (r,s)-QSAT if $r' \geq r$, $s' \geq s$.

Theorem 2.3 ([47])**.** *(3,5)-*QSAT *is QMA_1 complete.*

It is worth also noting that the analogous (2,3)-SAT is NP-complete, but it is unknown if (2,3)-QSAT is QMA_1 complete or not. Some 2-local qutrit Hamiltonians such as the AKLT Hamiltonian[4] exhibit interesting entanglement structure in a frustration free system.

Definition 2.4 (k-STOQ-QSAT [30])**.** *The QCSP k-*STOQ-QSAT *is the version of k-*QSAT *restricted so that all clauses have nonpositive off-diagonal elements (they are "stoquastic").*

Theorem 2.4 ([30])**.** *For any $k \geq 6$, the problem k-*STOQ-QSAT *is* MA*-complete. For $k < 6$, it is contained in* MA*.*

2.2.4 Quantum Optimization Problems

As an aside, it is worth pointing out the difference between quantum constraint problems and quantum *optimization* problems. A quantum optimization problem has a similar form: an allowed set of local operators, and we ask if there is a a quantum state with sufficiently low energy. The two key differences are that (1) the local operators are not necessarily projectors, and (2) we ask if the ground state has energy below some a, as opposed to being in the ground state of all clauses simultaneously. The Hamiltonians constructed in satisfiable instances are not necessarily frustration free, then.

The complexity classes of these problems were classified[2] by Cubitt and Montanaro in [39], where it was shown they fall into P, NP-complete, StoqMA-complete, and QMA-complete. Indeed, it is known the optimization problem 2-Local-Hamiltonian is already QMA-complete [77], analogous to MAX-2-SAT already being NP-complete. Cubitt and Montanaro's classification is analogous to the "Min CSP classification theorem" on classical optimization problems, an optimization-oriented analog of the dichotomy theorem. We will not further discuss optimization problems here.

2.3 Statement of results

It is known that some CSPs are not simply in P, but in fact complete for P: they model all efficient classical computation. It seems natural to ask whether there are QCSPs that capture efficient quantum computation, viewed as BQP. The class BQP consists of the problems for which there exists uniform quantum circuits of polynomial size that return the correct answer with probability at least 2/3. However, we are discussing exact satisfying assignments of a Hamiltonian, so allowing a solution that may be wrong 1/3 of the time would prove very difficult, and we must restrict ourselves to one-sided error. This is the same reason that [56] uses QMA_1 over QMA. We define the class as following:

[2] Technically, this result was shown only for problems on qubits; and then, only for when the clauses are all 2-local, or all Pauli matrices are allowed clauses. It seems reasonable to expect they generalize.

Definition 2.5 (BQP_1)**.** *A language L belongs to the class BQP_1 iff there is a uniform family of quantum circuits U of polynomial size, such that for an input x:*

***Perfect Completeness**: If $x \in L$, then measuring the first qubit of $U|x\rangle$ gives 1 with probability 1.*

***Soundness**: If $x \notin L$, then measuring the first qubit of $U|x\rangle$ gives 0 with probability at least $2/3$.*

It is worth noting that BQP_1 could also be called coRQP, the set of complements to RQP: quantumly solvable problems with perfect soundness and bounded-error completeness [20, 19]. Our first main result is

Theorem 2.5. *There is a fixed set $\mathcal{C}$ of 5-local projectors on 13-dimensional qudits, such that the QCSP for $\mathcal{C}$ is complete for BQP_1.*

The majority of the work shall be in proving this theorem. In the process of constructing these projectors, it will become apparent that two more complexity classes could be handled as well, QCMA_1 and coRP.

Definition 2.6 (QCMA_1)**.** *A language L belongs to the class QCMA_1 iff there is a uniform family of quantum circuits U of polynomial size, and a polynomial $p(n)$, such that for an input x:*

***Perfect Completeness**: If $x \in L$, there is a classical bitstring $y \in \{0,1\}^{p(|x|)}$, such that measuring the first qubit of $U|x\rangle|y\rangle$ gives 1 with probability 1.*

***Soundness**: If $x \notin L$, then for all $y \in \{0,1\}^{p(|x|)}$, measuring the first qubit of $U|x\rangle|y\rangle$ gives 0 with probability at least $2/3$.*

Definition 2.7 (coRP)**.** *A language L belongs to the class* coRP *iff there is a random Turing machine T, such that T always runs in polynomial time, and for any input x:*

***Perfect Completeness**: If $x \in L$, T accepts x with probability 1.*

***Soundness**: If $x \notin L$, T rejects x with probability $\geq 2/3$.*

The class QCMA_1 is a one-sided error variant of QCMA [6]. In [75] it was shown that $\mathsf{QCMA}_1 = \mathsf{QCMA}$, so our QCMA_1-complete problem is equivalently QCMA-complete. coRP, and its more common complement class RP, are standard [95]. They are the one-sided error versions of BPP. Our other two main results are,

Theorem 2.6. *There is a fixed set $\mathcal{C}$ of 5-local projectors on 15-dimensional qudits, such that the QCSP for $\mathcal{C}$ is complete for* QCMA*.*

Theorem 2.7. *There is a fixed set $\mathcal{C}$ of 5-local projectors on 15-dimensional qudits, such that the QCSP for $\mathcal{C}$ is complete for* coRP*.*

This seems to be the first discussion of BQP_1, and its difficulty is likely similar to that of BQP. In the context of coRP, the distinction between promise problems

and decision problems is more important, and so it is technically a Promise coRP-complete problem.

Our last results establish that all QCSPs can be described entirely on qubits, something which is not expected to hold in the classical setting. This implies that there are also QCSPs for BQP_1, QCMA, and coRP on qubits alone.

Theorem 2.8. *(Informal) Every QCSP $\mathcal{C}$ on d-qudits is equivalent in difficulty to some QCSP $\mathcal{C}'$ on qubits.*

2.4 Construction Techniques

Before diving into the construction of the BQP_1-complete Hamiltonian, we go over a few of the tools.

2.4.1 Universal Gate Set

In their proof that Quantum 3-SAT is QMA_1-Complete, Gosset and Nagaj [56] use a gate set $\mathcal{G}_\prime = \{\hat{H}, T, CNOT\}$. This has the property that every matrix element is of the form $(a + ib + \sqrt{2}c + i\sqrt{2}d)/4$, see their Definition 3. For reasons that will become clear later, we would like a gate set that leaves no classical basis state unchanged, and so we use the modified gate set

$$\mathcal{G} = \{\hat{H}, \hat{H}T, (\hat{H} \otimes \hat{H})CNOT\} \tag{2.1}$$

Through appropriate multiplications by $\hat{H}$, this allows the construction of all of $\mathcal{G}_\prime$, so this is still universal. It is straightforward to compute all matrix elements and check they have the same form. A language L is BQP_1-hard if there is a (classical) polynomial time reduction from any $L' \in \mathsf{BQP}_1$ to L. A language is BQP_1-complete if it is both BQP_1-hard and in BQP_1 itself. It is immediate that $\mathsf{EQP} \subseteq \mathsf{BQP}_1 \subseteq \mathsf{BQP}$. Given that EQP and BQP are commonly understood to capture at least some of the power of quantum computing (e.g. strong oracle separations from P), BQP_1 can be understood to be lower- and upper-bounded between these, and so has some essential quantum nature itself.

Unfortunately, since BQP_1 is a class with a vanishing error probability (on one-side), the notion of "universal gate set" is delicate. The gate that conditionally rotates phase by $\pi/3$, for instance, cannot be built exactly from $\mathcal{G}_\prime$, although it can be approximated exponentially well. This is a problem faced in Gosset and Nagaj [56] as well.

For this reason, BQP_1 is not a well-defined complexity class on its own; it requires an assumption on the allowed gate set that the circuit U should be built from. If we fix the gate set $\mathcal{G}$ above, we get a class $\mathsf{BQP}_{1,\mathcal{G}}$, of problems with one-sided error solvable using $\mathcal{G}$. The QCSP we will construct below is complete for the class $\mathsf{BQP}_{1,\mathcal{G}}$. If we preferred a different gate set $\mathcal{G}'$, say be adding $R(\pi/3)$ gate to $\mathcal{G}$, we would have a new class $\mathsf{BQP}_{1,\mathcal{G}'}$, and the problem we construct is

complete for that class. In this sense, the construction is generic, and we simply write that the QCSP is BQP_1-complete. In section 2.8 we give a definition of *weak* QCSPs that attempts to fix these irritating details.

2.4.2 Ternary Logic

We will use ternary logic (also known as "dual rail logic"), a standard tool in proving P-completeness. While 2-SAT is a problem that can be solved efficiently on a classical computer, it is not believed to capture the full power of classical computing. But another problem, Horn-SAT, is P-complete: it captures the full power of classical computation. The ideas in that proof will be important in our construction, lifted to a quantum setting, so we briefly outline it here.

Horn-SAT is the boolean CSP which allows any OR clauses in at most 3 variables with at most one negative variable: clauses like $(v_i \vee v_j)$, $(\neg v_i \vee v_j \vee v_k)$, (v_i), $(\neg v_i)$, but not $(\neg v_i \vee \neg v_j)$. By setting all variables to true, all clauses are satisifed except those of form $(\neg v_i)$, which imply v_i must be false. This propagates, possibly reducing some $(v_i \vee \neg v_j) \rightarrow (\neg v_j)$, flipping more variables to false until a satisfying instance or conflict is found. Thus, it is in P.

To be P-complete, a CSP must be flexible enough to establish an arbitrary computational graph, usually as a circuit. And yet, it must not permit the construction of problems that require guessing the input, such as a circuit with the

input left blank and the output forced "True", because this would lead to NP-hardness. Informally, Horn-SAT accomplishes P-completness through *dual-rail logic*: for each Boolean variable v_i in the original circuit, we make *two* variables in the Horn-SAT instance, $v_{i,T}$ and $v_{i,F}$, representing the assertion that v_i is true or false, respectively. We add the constraints that $v_{i,T} \implies \neg v_{i,F}$. This implies that at most one of them can be true – but does not rule out the case that both are false. This is the third state, "Undefined", of the ternary logic. All logic gates of the circuit are then constructed so that they are trivially satisfied, if both $v_{i,T}$ and $v_{i,F}$ are false. A Horn-SAT problem can then constrain the inputs to a circuit by requiring one or the other variable to be true, but if the input is left unconstrained (as in the NP proof-checking setup), then the circuit can be satisfied by just leaving both $v_{i,T}$ and $v_{i,F}$ false.

If our variables are 3-state (instead of Boolean), we can construct P-Complete problems more easily. The dual-rail variable is replaced with a single variable, whose states are labelled "true", "false", and "undefined". The output of a gate with "undefined" at any input can be anything, but if a gate has two defined inputs, it must compute its output appropriately. This allows the implication of variable states to only travel forward in the computation graph (conducting computation along the way) and not backwards (which would allow the construction of NP-hard input-guessing problems).

2.4.3 Monogamy

The monogamy of entanglement [37] states that a subsystem A cannot be fully entangled with subsystem B and with subsystem C at the same time. A 2-qubit clause such as $H_{BellPair} = I - (|00\rangle + |11\rangle)(\langle 00| + \langle 11|)$ has a unique ground state, a Bell pair, which is fully entangled between its two qubits. In a QCSP with $H_{BellPair}$ applied once to qubits 1 and 2, and applied a second time to qubits 2 and 3, we could immediately reject: any satisfying assignment would require qubit 1 to be fully entangled with qubit 2 and qubit 3, violating monogamy. This is a trick we can use to force certain clauses to pair up certain variables, without allowing the creation of any more complicated constraint graphs.

2.5 Existence of BQP_1-complete Constraint Satisfaction Problems

We now proceed to show that there exist CSPs that are complete for BQP_1. While our proof proceeds constructively, i.e. considers a specific class of such Hamiltonians, this specific class is rather specialized and implausible of a real physical system. However, the crucial point is that such a CSP can exist at all, rather than its specific form.

The problem will be built with 13-dimensional qudits. We will soon define the problem in terms of the allowed clauses, but first give a more convenient notation for the 13-dimensional space. There is a subspace spanned by three states, labelled as

$$|0_L\rangle, |1_L\rangle, |U_L\rangle$$

another two states labelled as

$$|0_{EC}\rangle, |1_{EC}\rangle$$

and an eight dimensional subspace which is the tensor product of three 2-dimensional spaces:

$$\mathbb{C}(|0_{CL}\rangle, |1_{CL}\rangle) \otimes \mathbb{C}(|0_{CA}\rangle, |1_{CA}\rangle) \otimes \mathbb{C}(|0_{CB}\rangle, |1_{CB}\rangle)$$

We will abuse notation somewhat and write operators such as $|1_{CL}\rangle\langle 0_{CL}|$, which should be understood as shorthand for $\mathbf{0}^5 \oplus (|1_{CL}\rangle\langle 0_{CL}| \otimes I^4)$. Here $\mathbf{0}^5$ is a 5-dimensional zero operator, on the the first 5 of the 13 states, and I^4 is the 4-dimensional identity operator on the CA and CB subspace. This is the most natural extension of $|1_{CL}\rangle\langle 0_{CL}|$, which is an operator on a two-dimensional Hilbert space, to an operator on the 13-dimensional Hilbert space.

We will also write equations such as $X = Y_{12} + Z_{23}$, to mean that X is a 3-local operator, built from the 2-local operators Y and Z, which act on the first two and last two qudits respectively. That is, $X = Y \otimes I + I \otimes Z$.

With this notation, we can define the problem now, after which we'll elaborate on how it was constructed, which will motivate the definition.

Definition 2.8 (Quantum-Clock-Ternary-SAT). *The problem* Quantum-Clock-Ternary-SAT *is the quantum constraint problem with 5 clauses. The allowed clauses are* H_{Start}, H_{End}, *and for each unitary* U *in the gate set* $\mathcal{G}$, *a clause* $H_{prop,U}$. *Since* $|\mathcal{G}| = 3$ *above, this is five clauses total.*

To give expressions for the clauses, first define the 1-local operators,

$$H_L = I - |0_L\rangle\langle 0_L| - |1_L\rangle\langle 1_L| - |U_L\rangle\langle U_L| \tag{2.2}$$

$$H_E = I - |0_{EC}\rangle\langle 0_{EC}| - |1_{EC}\rangle\langle 1_{EC}| \tag{2.3}$$

$$H_C = |0_L\rangle\langle 0_L| + |1_L\rangle\langle 1_L| + |U_L\rangle\langle U_L| + |0_{EC}\rangle\langle 0_{EC}| + |1_{EC}\rangle\langle 1_{EC}| \tag{2.4}$$

$$P_D = |0_L\rangle\langle 0_L| + |1_L\rangle\langle 1_L| \tag{2.5}$$

and the 2-local operator,

$$H_{BP} = I - (|0_{CB}0_{CA}\rangle + |1_{CB}1_{CA}\rangle)(\langle 0_{CB}0_{CA}| + \langle 1_{CB}1_{CA}|) \tag{2.6}$$

and for any 2-qubit unitary U, *define the 2-local* $T(U)$ *by the product*

$$T(U) = U_B(P_D \otimes P_D) \tag{2.7}$$

where U_B *acts on* $|0_L\rangle$ *and* $|1_L\rangle$ *the way that* U *would act on* $|0\rangle$ *and* $|1\rangle$, *and the zero operator if either input is something else. That is, mapping* $|0\rangle$ *and* $|1\rangle$ *to*

$|0_L\rangle$ and $|1_L\rangle$ induces an isometry $\mathbb{C}^2 \to \mathbb{C}^{13}$; this extends naturally to operators $U : (\mathbb{C}^2)^2 \to (\mathbb{C}^2)^2$.

There are three more clauses to define, but they are difficult to write down simply as projectors. In terms of determining frustration-free ground states, we really only care about the fact that our operators are positive semidefinie, and geometry of their kernels, not the energies of any excited states. Any non-projector operator can be "normalized" by adjusting the energies of all excited states to be 1. For example, a 2-qubit operator that applies H_C on the first qubit and H_E on the second qubit could be written $H_{C,1} + H_{E,2}$, and this is shorthand for the normalized projector $I - (I - H_C) \otimes (I - H_E) = H_{C,1} + H_{E,2} - H_{C,1}H_{E,2}$. The following three definitions use this shorthand.

The clauses are H_{Start} and H_{End} are 4-local and defined by

$$\begin{aligned} H_{Start} = &(I - (|0_{EC}0_{CA}\rangle + |1_{EC}1_{CA}\rangle)(\langle 0_{EC}0_{CA}| + \langle 1_{EC}1_{CA}|))_{12} \\ &+ (|0_{CL}\rangle\langle 0_{CL}| \otimes (I - |0_L\rangle\langle 0_L|))_{24} \\ &+ \Big((|0_{CL}\rangle\langle 0_{CL}| + |1_{CL}\rangle\langle 1_{CL}| - |0_{CL}\rangle\langle 1_{CL}| - |1_{CL}\rangle\langle 0_{CL}|) \otimes |0_{CL}\rangle\langle 0_{CL}|\Big)_{23} \\ &+ H_{BP,23} \\ &+ H_{E,1} + H_{C,2} + H_{C,3} + H_{L,4} \end{aligned} \tag{2.8}$$

$$
\begin{aligned}
H_{End} = &(I - (|0_{EC}0_{CB}\rangle + |1_{EC}1_{CB}\rangle)(\langle 0_{EC}0_{CB}| + \langle 1_{EC}1_{CB}|))_{12} \\
&+ (|1_{CL}\rangle\langle 1_{CL}|) \otimes |1_L\rangle\langle 1_L|)_{24} \\
&+ \Big((|0_{CL}\rangle\langle 0_{CL}| + |1_{CL}\rangle\langle 1_{CL}| - |0_{CL}\rangle\langle 1_{CL}| - |1_{CL}\rangle\langle 0_{CL}|) \otimes |1_{CL}\rangle\langle 1_{CL}|\Big)_{23} \\
&+ H_{BP,32} \\
&+ H_{E,1} + H_{C,2} + H_{C,3} + H_{L,4}
\end{aligned}
\tag{2.9}
$$

Each $H_{prop,U}$ is 5-local, defined by

$$
\begin{aligned}
H_{prop,U} = &P_D \otimes P_D \otimes |1_C0_C0_C\rangle\langle 1_C0_C0_C| + I^{\otimes 2} \otimes |1_C1_C0_C\rangle\langle 1_C1_C0_C| \\
&- T(U) \otimes |1_C1_C0_C\rangle\langle 1_C0_C0_C| - T(U)^\dagger \otimes |1_C0_C0_C\rangle\langle 1_C1_C0_C| \\
&+ H_{BP,34} + H_{BP,45} \\
&+ H_{L,1} + H_{L,2} + H_{C,3} + H_{C,4} + H_{C,5}
\end{aligned}
\tag{2.10}
$$

(These are not projectors as written, and they should be interpreted as the normalized versions; really, that $H_{prop,U}$ is the unique projector with the same kernel as the operator on the right-hand side.)

End definition.

Now it is time to give some meaning to the parts of the problem. We can give better names to the 13 basis states. The first three basis vector span the "logical subspace", on which we will do ternary logic: 0_L and 1_L represent logical 0 and

1 qubit states, and U_L represents an "undefined" qubit. The operator H_L just requires that a particular 13-qudit is, in fact, a logical qubit.

Since our construction is based on Kitaev's circuit-to-Hamiltonian mapping, we need clock qudits, but we will assign them separate states than the logical qudits. The latter 10 vectors of our 13-dimensional space form the "clock subspace", which will perform this role. This allows to avoid worrying whether a $|0\rangle$ is a "clock zero" or a "logic zero". Compare this with Kitaev's Hamiltonian, where the same $|0\rangle$ state is used for both, but many constraint problems can be built that look nothing like a circuit.

Eight of the clock states are a tensor product of three 2-dimensional subspaces: $\{|0_{CL}\rangle, |1_{CL}\rangle\}$, $\{|0_{CA}\rangle, |1_{CA}\rangle\}$, and $\{|0_{CB}\rangle, |1_{CB}\rangle\}$. The $_{CL}$ component is "logical clock" states, corresponding to the 0 and 1 clock states in Kitaev's clock construction, and these actually carry the information of the current time is in the circuit's evaluation. The $_{CA}$ and $_{CB}$ components do not carry information on timing, and will be used for something else. The operator H_C requires that a 13-qudit is a clock qudit.

Since we will be able to do the same clock-to-Hamiltonian mapping, we will have BQP_1-hardness. But in order to keep the difficulty within BQP_1, we want to avoid building any problem that looks like something *other* than a clock-to-Hamiltonian mapping. In particular, things would be very complicated if the

chain of clock states branched, instead of forming a linear path of time. We will use monogamy to uniquely pair each clock qudit – and therefore, each moment in time – with a *unique* predecessor and successor.

The CA and CB components are the auxiliary parts of the clock state, used in the monogamy construction. By establishing a Bell pair between the CA component of a 13-qudit x the CB component of another 13-qudit y, we are stating that y is the clock qudit immediately following x in time. Since the CA component of x cannot form another Bell pair by monogamy, x now has a uniquely following moment in time. H_{BP} expresses this constraint between two clock qudits.

2.5.1 Initializing and terminating

Unlike QMA problems, we do not want to leave the input up to guessing: the circuit should start in the $|0^n\rangle$ state. For that purpose H_{Start} is designed to force a particular qubit to start in the $|0\rangle$ state. Since the start of time is indicated by the first clock qudit being zero, H_{Start} really only needs to say that: either the clock qudit is one, or the logical qudit is zero. This is the second line of (2.8).

But it would be a headache if it H_{Start} was applied somewhere other than the start of time, effectively forcing logical qubits to be zero in the middle of execution. To avoid this mess, we've added two more "clock endpoint" states, $|0_{EC}\rangle$ and $|1_{EC}\rangle$. These can be entangled with a $_{CA}$ or $_{CB}$ to form a Bell pair,

terminating a chain of clock qudits. For the same reasons of monogamy, a single $_{CA}$ can't be entangled with anything else if it's entangled with a $_{EC}$ qubit.

So, H_{Start} can be attached a clock qudit x, and require that x's $_{CA}$ subspace is entangled with an $_{EC}$ subspace; this means that x must be the first clock qudit in the chain. Then H_{Start} says that, if x is zero, the logical qubit y must be zero as well. This accomplishes initialization of the input.

Breaking down (2.8), it is an operator on 4 qudits: an endpoint qudit, a clock qudit, a second clock qudit, and a logical qudit. The fifth line forces them to be of these these types. The first line requires that the first two qudits form a maximal Bell pair with their $_{EC}$ and $_{CA}$ states. The second line requires that, when the second (clock) qudit is zero, the fourth (logical) qudit is also zero.

To understand lines three and four of (2.8), we need to note a particular detail with how time is encoded, and the propagator terms. In Feynman's description, each propagator has the form

$$I \otimes |t\rangle \langle t| - I \otimes |t+1\rangle \langle t+1| - U_t \otimes |t\rangle \langle t+1| - U_t^\dagger \otimes |t+1\rangle \langle t|$$

and Kitaev noted that checking the time $|t\rangle$ can be done in a 3-local way with a unary encoding, where $|t\rangle$ is indicated by the pattern $|110\rangle_{t-1,t,t}$. This form doesn't apply when $t = 0$ or $t = L$, though, and they instead need terms like $|00\rangle_{0,1}$ and $|11\rangle_{t_{Max}-1,t_{Max}}$. This means that we can't use the same form of propagator clause for all times. One solution would be to have additional clause types, specifically

for the start and end, but this could make the reasoning about correctness considerably more complicated. The resolution is to have H_{Start} and H_{End} include a propagator of their own, applying an identity gate in the process that doesn't modify the state. Then our main propagator term can begin at $t = 2$ with the same form for each application.

H_{End} does in (2.9) almost the exact opposite of H_{Start}. Instead of pairing with $_{CA}$, the endpoint qudit forms a Bell pair with the $_{CB}$ subspace, because we want H_{End} to come at the end of the chain. The second line of (2.9) requires that, when the second (clock) qudit is one, the third (logical) qudit is either 0_L or U_L – zero or undefined. This way, H_{End} states that, at the end of execution, we should not get a "1" as a result. If all the input bits are defined, we will get a well-defined ouptut, and H_{End} performs the same role as H_{out} in [81] and [78].

2.5.2 Propagating ternary logic

We are not guaranteed that all of our qudits have an H_{Start} clause on them, and we want to avoid having to guess the input as in QMA problems. In a classical circuit, ternary logic would address this problem as follows: should any of the input bits be uninitialized, all downstream bits can be undefined, leaving the output undefined as well, which is always a satisfying assignment.

In our quantum world, where the satisfying state is a superposition of the computing history across all time, there's a simpler solution: just end time itself, and "destroy the universe" if we try to compute on an undefined input.

In [81], the propagating clause for a two-qubit unitary U was:

$$H_{Kitaev} = I \otimes (|t\rangle\langle t| + |t-1\rangle\langle t-1|) - U \otimes |t\rangle\langle t-1| - U^\dagger \otimes |t-1\rangle\langle t|$$

$$\text{where} \quad |t-1\rangle = |100\rangle\,,\ |t\rangle = |110\rangle\,.$$

For a candidate solution $|\psi\rangle$, it straightforward to check that $|\psi\rangle$ can only be in the nullspace of H if:

$$U\,(I \otimes \langle t-1|)\,|\psi\rangle = (I \otimes \langle t|)\,|\psi\rangle$$

that is, the logical state encoded in $|\psi\rangle$ for time t must be equal to U applied to the state at time $t-1$. They must also have the same amplitude. The fact that the amplitudes are equal at each step implies $|\psi\rangle$ must be a uniform superposition across all times.

For our $H_{prop,U}$ defined in (2.10), we need these ingredients, that for *defined* inputs we propagate the state with the same amplitude. If the input is in the undefined state $|U_L\rangle$, we will drop the requirement that the amplitude remains the same. The computation can terminate early, because our solution $|\psi\rangle$ is no longer required to have any component on times after $t-1$.

The first line of (2.10) replaces an I in H_{Kitaev} with a $P_D \otimes P_D$, which projects away $|U_L\rangle$ states. $H_{prop,U}$ has all the same nullspace as H_{Kitaev}:

$$|0_L 0_L\rangle \otimes |1_C 0_C 0_C\rangle + (U |0_L 0_L\rangle) \otimes |1_C 1_C 0_C\rangle ,$$

$$|0_L 1_L\rangle \otimes |1_C 0_C 0_C\rangle + (U |0_L 1_L\rangle) \otimes |1_C 1_C 0_C\rangle ,$$

$$|1_L 0_L\rangle \otimes |1_C 0_C 0_C\rangle + (U |1_L 0_L\rangle) \otimes |1_C 1_C 0_C\rangle ,$$

$$|1_L 1_L\rangle \otimes |1_C 0_C 0_C\rangle + (U |1_L 1_L\rangle) \otimes |1_C 1_C 0_C\rangle$$

with the additional options of:

$$|U_L 0_L\rangle \otimes |1_C 0_C 0_C\rangle ,$$

$$|U_L 1_L\rangle \otimes |1_C 0_C 0_C\rangle ,$$

$$|0_L U_L\rangle \otimes |1_C 0_C 0_C\rangle ,$$

$$|1_L U_L\rangle \otimes |1_C 0_C 0_C\rangle ,$$

$$|U_L U_L\rangle \otimes |1_C 0_C 0_C\rangle$$

The third line of (2.10) requires that $H_{prop,U}$ be placed on the correct sequence of clock bits – that if it is placed on, say, the 3rd, 5th, and 12th sites of the clock chain, that will be unsatisfiable because of the contradictory Bell pairs. The fourth line of (2.10) simply requires the right types of particles at each site.

Now that we have motivated the definition, we can proceed to the main result.

NameIgnored 1 (Theorem 2.5 (restated))**.** Quantum-Clock-Ternary-SAT *is* BQP_1*-complete.*

Proof. We need to show provide a BQP_1 algorithm for deciding instances of Quantum-Clock-Ternary-SAT , and show its completeness and soundness, and then that Quantum-Clock-Ternary-SAT is BQP_1-hard.

Part 1: Quantum-Clock-Ternary-SAT *is in* BQP_1.

First, a brief note about the form of the input. Providing every term of the input Hamiltonian in the standard basis would take up exponential space. We assume that the input is provided as a list of clauses (and the qudits they operate on), or a list of 5-local interactions (which might not be manifestly of the form allowed). It possible to find the set of clauses corresponding to a list of 5-local interactions by solving a system of linear equations in $O(n^5)$ time; that this preprocessing can be performed in polynomial time on a classical machine means we don't have to care about the input form. Henceforth we assume that the input is a list of $(clause, sites)$ data, indicating that clause number $clause$ is acting on the sites $sites$.

The following BQP_1 algorithm, we claim, solves this problem. The algorithm proceeds by identifying the structure of (possibly several) BQP_1 circuits in the problem, which is a classical operation that can be completed in polynomial time. Then it executes each circuit and verifies that the result passes.

Algorithm 1

1. For each 13-qudit, check all clauses it occurs in. Each clause will apply one of H_L, H_C, or H_E to it. A given qudit should only ever have one of these applied to it, otherwise we immediately reject. Label each qudit as a logical, clock, or endpoint qudit, depending on which is applied. Any qudits that have none of these applied, have no clauses applied at all, and so can be ignored for the rest of the problem.

2. For each clock qudit and endpoint qudit, inspect all H_{BP}, $H_{Start,BP}$, and $H_{End,BP}$ terms applied to it. These should only ever create bell pairs between the same pairs of underlying qubits, in the CA, CB, and EC subspaces. If any underlying qubit is paired with multiple others, reject.

3. Every qudit labelled as an endpoint qudit has at least one of H_{Start} or H_{End} applied to it. If any endpoint qudit has *both* H_{Start} and H_{End}, reject. Otherwise, label it as a "start qudit" or "end qudit" accordingly, and proceed.

4. The pairs from step 2 induce a linear structure where endpoints are connected to at most one clock qudit, and each clock qudit is connected to at most two qudits on either side. Following these connections, all clock qudits and endpoint qudits can be linked into some collection of paths and cycles.

5. For any cycle from step 4 (which necessarily consists entirely of clock qudits), or any path without any endpoint qudits, assign all clock qudits the $|0_C\rangle$ state, and they can be ignored for the rest of the problem.

6. If any remaining paths have no start qudit, assign all clock qudits in that chain $|1_C\rangle$ and ignore them for the rest of the problem.

7. If any remaining paths have no end qudit, assign all clock qudits in that chain $|0_C\rangle$ and ignore them for the rest of the problem.

8. All remaining paths have at least one start qudit and end qudit. Since a single start or end can't be entangled with multiple others, it must be exactly one start qudit and end qudit – otherwise it would have been rejected in step 2.

9. At this point we are left with a collection of clock paths, with associated unitaries from each $H_{prop,U}$ acting on logical qubits. In the case of a single clock path, this is a (ternary-logic) quantum circuit, and we need to evaluate it. In the case of multiple clock paths operating on the same qubits, we need to ensure compatibility of the two circuits, which is more complicated. We first describe the case of a single clock path, and then generalize to multiple clock paths.

Case: Single clock path Every logical qubit with a H_{Start} on it must be in the $|0_L\rangle$ state at $t = 0$. Every other logical qubit could be in any state, including being entangled with each other; but we will see that we can assume that they all begin in the $|U_L\rangle$ state, representing an "undefined" logical state, without losing completeness.

Any time a qubit in the $|U_L\rangle$ state reaches a unitary gate, we can find a satisfying assignment for that circuit by terminating history there. In a fully functional circuit with unitaries $U_1, \ldots U_T$ and initial state $|0^n\rangle$, the full solution to the constraint problem would be the state

$$|\Psi\rangle = \frac{1}{\sqrt{T}} \sum_{i=1}^{T} (U_i U_{i-1} \ldots U_1 |0^n\rangle) \otimes |Clock_i\rangle \tag{2.11}$$

and the H_{End} checks the $|Clock_T\rangle$ subspace. But if an input qudit $|U_L\rangle$ is operated on at time t, then a solution to the constraint problem is the state

$$|\Psi\rangle = \frac{1}{\sqrt{t}} \sum_{i=1}^{t} (U_i U_{i-1} \ldots U_1 |0^n\rangle) \otimes |Clock_i\rangle \tag{2.12}$$

This is in the ground state of each H_{prop}, and when H_{End} checks the $|Clock_T\rangle$ subspace, we are trivially in its ground state, because the projection of $|\Psi\rangle$ in that subspace is the zero vector.

So, if any undefined qubit is acted on by a unitary, we know we can accept, without any further checking. If no such case arises, then any solution to the QCSP must be a valid computational history. A quantum computer can execute

the circuit from the known starting state $|0^n\rangle$, and measure all the check qudits, the qudits with H_{End} applied. If these are all in the $|0\rangle$ state, we accept. If any are in the $|1\rangle$ state, we reject.

Case: Multiple clock paths

If we have multiple clocks that operate on disjoint sets of qubits, then our constraint problem is completely separable, and we need only to verify each part independently. It becomes more difficult if both clock paths are operating on the same set of qubits, as there is no meaningful ordering of time.

To handle that case, will show that either the clock paths can be reduced in a way such that they no longer share any logical qubits; or, if this is impossible, then it must be a frustrated instance. Picking up from Step 9 above, we have processed our problem to have only *complete* clock paths – those that start with H_{Start}, have a path of $H_{Prop,U}$, and terminate with H_{End}. We would like to know that each qubit is either defined or not, for any times:

Theorem 1. *If a qubit Q is initialized by H_{Start} in any complete clock path, then its support must be entirely in the $|0_L\rangle$ and $|1_L\rangle$ subspace; that is, any frustration-free ground state would never produce a measurement $|U_L\rangle$.*

Proof. We can decompose a candidate ground state as,

$$|\Psi\rangle = \sum_{t=0}^{t=t_{Max}} \alpha_t |\psi_t\rangle \otimes |t\rangle \tag{2.13}$$

where the factor states represent everything other than the clock path, and the time. The H_{Start} implies that when $t = 0$, Q is in the pure $ket0_L$ state. The propagators (which exist for each time, because the clock path is complete) imply that

$$\alpha_{t+1} |\psi_{t+1}\rangle = \alpha_t T(U) |\psi_t\rangle \tag{2.14}$$

The operator $T(U)$ will either map a defined state to another defined state, or an undefined state to a zero vector. But, it will never map a defined state to an undefined state. Since Q starts out defined at $t = 0$, it cannot become undefined at any t. □ □

This means we can unambiguously separate qubits into "defined" logical qubits (those with an H_{Start} from a complete clock path) and "undefined" qubits (the others). Now we look at what happens if two clock paths share defined logical qubits:

Theorem 2. *If two complete clock paths operate only on defined qubits, one clock path has an H_{Start} term on a qubit, and another clock path acts with a gate on that qubit, then the resulting Hamiltonian is frustrated.*

Proof. Suppose there is a state $|\Psi\rangle$ with zero energy. Any state on three subsystems A, B, C can be expanded in terms of a basis on each of those subsystems. In this case, we will expand $|\Psi\rangle$ in terms of the (logical qubits)×(first clock

path)×(second cloth path) subsystems:

$$|\Psi\rangle = \sum_{\substack{a=0,\, b=0,\, c=0}}^{\substack{a=2^{N_{logical}} \\ b=2^{1+tMax_1} \\ c=2^{1+tMax_2}}} \gamma_{a,b,c} \, |a\rangle \otimes |b\rangle \otimes |c\rangle \tag{2.15}$$

so that the coefficients $\gamma_{a,b,c}$ determine the state in the standard basis. As before, our support must be entirely on clock strings like $1 \ldots 10 \ldots 0$, as any "01" configuration would incur an energy penalty. Thus our $|\Psi\rangle$ must exist entirely on the valid clock states:

$$|\Psi\rangle = \sum_{\substack{a=0 \\ \tilde{i} \in \text{ clock states} \\ \tilde{j} \in \text{ clock states}}}^{a=2^{N_{logical}}} \gamma_{a,\tilde{i},\tilde{j}} \, |a\rangle \otimes |\tilde{i}\rangle \otimes |\tilde{j}\rangle \tag{2.16}$$

Here $\tilde{i}$ is running over only $1 + tMax_1$ many values, in place or i running over all 2^{1+tMax_1}, and likewise for $\tilde{j}$. We know that the other entries of γ are all zero, those corresponding to invalid clock states. For this reason we adapt our indexing and write $\tilde{i}$ as running from 0 to $tMax_1$. The time coordinates (the clock states) are then given by a pair $(\tilde{i}, \tilde{j})$. To each pair of times $(\tilde{i}, \tilde{j})$, there is corresponding logical state $\sum_{a=0} \gamma_{a,\tilde{i},\tilde{j}} \, |a\rangle$ – although that expression is not normalized. So let us define $|\psi_{i,j}\rangle$ by

$$\alpha_{\tilde{i},\tilde{j}} \, |\psi_{\tilde{i},\tilde{j}}\rangle = \sum_{a=0} \gamma_{a,\tilde{i},\tilde{j}} \, |a\rangle \, , \quad \langle \psi_{\tilde{i},\tilde{j}} | \psi_{\tilde{i},\tilde{j}} \rangle = 1, \quad 0 \leq \alpha_{\tilde{i},\tilde{j}} \leq 1. \tag{2.17}$$

In this form, $|\psi_{\tilde{i},\tilde{j}}\rangle$ indicates the logical state at a given pair of times, and $\alpha_{\tilde{i},\tilde{j}}$ tells us its relative magnitude. Explicitly,

$$\alpha_{\tilde{i},\tilde{j}} = \sqrt{\sum_{a=0}^{a=2^{N_logical}} |\gamma_{a,\tilde{i},\tilde{j}}|^2} \tag{2.18}$$

(At this point we drop the tildes on $\tilde{i}$ and $\tilde{j}$ as there is no confusion with the earlier indices.) These $|\psi_{i,j}\rangle$ are the correctly normalized states. Then the whole state, the one that by hypothesis has zero energy, is

$$|\Psi\rangle = \sum_{\substack{i=0,\, j=0 \\ i=tMax_1 \\ j=tMax_2}} \alpha_{i,j} |\psi_{i,j}\rangle \otimes |i\rangle \otimes |j\rangle \tag{2.19}$$

At this point, we have only slightly narrowed the space of possible $|\Psi\rangle$s, which are those with illegal clock states (clock states with a "01" configuration). The form above is superficially similar to a history state, but the different scalars $\alpha_{i,j}$ mean it is not necessarily a history state. A history state would have the additional requirement that the mangitudes of all α's are equal.

Each propagator $H_{Prop,U}$ on circuit 1 is only satisfied when

$$\alpha_{i+1,j} |\psi_{i+1,j}\rangle = U_i \alpha_{i,j} |\psi_{i,j}\rangle , \tag{2.20}$$

and likewise for circuit 2,

$$\alpha_{i,j+1} |\psi_{i,j+1}\rangle = U_j \alpha_{i,j} |\psi_{i,j}\rangle . \tag{2.21}$$

As the Us are unitary, we can conclude that all $\alpha_{i,j}$ are the same magnitude, and in particular can all be taken to be equal to $1/\sqrt{t_{Max,1} t_{Max,2}}$. At this point we *do*

know that $|\Psi\rangle$ must be a history state, albeit an unusual one with multiple times coordinates.

The hypothesis of this lemma is that there is a qubit Q acted on with a gate U in circuit 1, and acted on with H_{Start} in circuit 2. The H_{Start} in circuit 2 requires that whenever $j = 0$, Q must be in the pure 0 state. The gate in circuit 1 is associated to some time t_G, and it requires that

$$|\psi_{t_G+1,j}\rangle = U\,|\psi_{t_G,j}\rangle \tag{2.22}$$

for all j. When $j = 0$, these together imply

$$|0\rangle\,|\psi'_{t_G+1,j}\rangle = U\,|0\rangle\,|\psi'_{t_G,j}\rangle \tag{2.23}$$

where U is acting on $|0\rangle$, and $|\psi'\rangle$ contains the other logical qubits. But we chose our gates in 2.1 so that a pure $|0\rangle$ state will not remain a pure $|0\rangle$ state after the gate. This form a condtradiction, so there can be no frustration-free state in this configuration. □ □

This leaves only the case of what happens when logical qubits are shared by two clock paths, but left undefined (no H_{Start}). We see that we can always put them in a pure $|U_L\rangle$ state, and this cannot make the problem unsatisfiable. For when we do so, the $H_{Prop,U}$ terminates the clock path early. This in turn removes gates from the path, which can only remove the problems that Lemma 2 describes. Terminating the clock path early also disconnects the H_{End}, which means that we

cannot fail the final check there. With these lemmas we can now write down the algorithm for multiple clock paths:

Continuing from the other pre-processing, we are left with a collection of complete linear clock paths, and logical qubits.

10. For any logical qubit with no H_{Start}, mark it as an undefined qubit. Mark the remaining logical qubits as defined. (Lemma 1)

11. For any gate that operates on an undefined qubit, reduce that clock path: it now only extends up to the time of that gate. Because all qubits without H_{Start} are undefined, the only remaining qubits that are acted on by gates are those with H_{Start}.

12. Once all clock paths are reduced, check if any logical qubit has an H_{Start} in one path and a gate in another. If so, reject (Lemma 2). If we pass this step, then no defined qubit has gates in multiple paths.

13. Otherwise, we are left with clock paths that involve disjoint sets of defined logical qubits. These can be checked separately as in the single-path case:

14. If it is a reduced clock path, we know that it passes without further computation (since there is no final H_{End} check).

15. Otherwise, execute the circuit with a quantum computer and verify its output is $|0_L\rangle$ on all logical qubits with an H_{End}.

We have used classical preprocessing to separate out the system into a number of single clock-path subsystem $S_1 \dots S_\nu$ with no interactions between subsystems. We have a zero-energy ground state $|\Psi\rangle$ for the whole system iff we have zero-energy ground states $|\Psi_\nu\rangle$ for each subsytem. Each $|\Psi_\nu\rangle$ can be decomposed into a history state $\sum_{i=1}^{N} \alpha_i |\psi_i\rangle_L |i\rangle_C$. Each propagator at time t implies $\alpha_t = \alpha_{t-1}$ and $|\psi_t\rangle = U_t |\psi_{t-1}\rangle$, up until it acts on an undefined input at some time t_U, where we take $\alpha_i = 0$ for all $i \geq t_U$. If no undefined inputs are acted on, we end up with a uniform superposition of time, and H_{End} validates the result.

To see completeness, consider a YES instance. It will first pass each step of the classical preprocessing (all of our rejection reasons in the preprocessing are justified). The final quantum circuit(s) have a unique execution history, that accept with probability 1; when we run those circuits, we accept the instance with probability 1 as well, so we get perfect completeness.

To see soundness, consider a NO instance. It is either a NO instance because it fails the structural requirements that we check in the preprocessing – and we reject it then – or because the only valid execution history has a nonzero amplitude in the $|1_L\rangle$ state on some H_{End} check. By the promise, it must be at least a $1/p(n)$ amplitude, otherwise its execution history would form a state with energy

$< 1/p(n)$. So our execution of the circuit has at least a $1/p(n)^2$ chance of rejecting. This is soundness.

Part 2: Quantum Clock-Horn-SAT is BQP_1-Hard

Let U_X be the uniform quantum circuit that solves some given BQP_1 problem. Let m be the number of qubits that U_X acts upon, and k be the number of gates applied. Without loss of generality, we assume the gates are applied one at a time, one per time step. Then take a Quantum Clock-Horn-SAT problem on $m + k + 4$ sites. m will be used for the qubits, k for the time steps, and 4 for the first and last clock sites and the endpoints. Label the logical qubit sites as Q_i, the timestep sites as T_i, and the other four as S, T_0, T_{k+1}, and E. Then the clauses in the problem are:

1. For each $i \in [m]$: an H_{Start} clause, acting on S, T_0, T_1, Q_i.

2. For each $t \in [k]$, where U_t is the t-th unitary of U_X acting on bits j and k: an $H_{prop,U}$ clause, acting on $Q_j, Q_k, T_{t-1}, T_t, T_{t+1}$.

3. A final H_{End} clause, acting on E, T_{k+1}, T_k, Q_1

Then we claim this problem, denoted $\mathcal{P}(U_X)$, is satisfiable iff $U_X \in L_{yes}$. First observe that each qudit can be identified as a logical, clock, or endpoint qudit, and so any satisfying assignment must be in the appropriate subspaces. We can now break each qudit into the subspaces of $|0_C A\rangle$, $|0_C L\rangle$, and so on. The entanglement

qubits are now an entirely separate problem from the logical and "logical clock" qubits, and they can be easily recognized to be satisfied by Bell pairs. Since all qubits are initialized in the $|0_L\rangle$ state, it can be seen inductiviely that they can never enter the undefined $|U_L\rangle$ state without polynomially large energy penalty. This lets us restrict to our attention to the behavior when all logical qubits are in the $|0_L\rangle/|1_L\rangle$ subspace. We can remove the Bell pairs, which are otherwise uncoupled, from the system. What remains is precisely Kitaev's clock Hamiltonian from [81], up to a renaming of the basis states, which is shown there to be frustration-free iff the modelled circuit accepts. Thus the our proof of correctness requires very little direct algebraic manipulation: after throwing away the extra subspaces, we are left with the same verbatim Hamiltonian, and we can rely on that result. Thus the majority of the arguing above is about under exactly what circumstances we can discard certain subspaces or summarily reject invalid instances.

We briefly summarize Kitaev's proof here for completeness. The history state $|\Psi\rangle$ can be decomposed in the clock basis as $|\Psi\rangle = \sum_{i=0}^{k} \alpha_i |\psi_i\rangle_L |i\rangle_C$ with α_i real and positive. $|\psi_0\rangle$ is necessarily the initializing $|0\rangle^n$ state on some qubits. By construction of this hard instance, there are no undefined qubits. The propagator at time t can only be in its ground state if $\alpha_{t-1} = \alpha_t$, and by induction all α's are equal to $1/\sqrt{k}$. Propagator t also can only be in its ground state if $|\psi_t\rangle =$

$U_t |\psi_{t-1}\rangle$, implying $|\psi_k\rangle = U_{circuit} |\psi_1\rangle$. The H_{End} then will produce an energy penalty unless the result of the circuit in Q_1 is a $|0\rangle$ state at step $t = k$.

If the original problem was in the language, the BQP_1 circuit always accepts, and the QCSP is exactly satisfiable. If the original problem was not in the language, the BQP_1 circuit rejects with probability at least 1/2, and so there is no approximately satisfying assignment in the QCSP. □

2.6 QCMA completeness

There is a straightforward modification of the above that shows the existence of QCMA-complete local Hamiltonian problems. In fact, we will construct a QCMA_1-complete problem, then use the result of [75] that QCMA_1=QCMA. The class QCMA_1 is similar to BQP_1, except that the input is allowed to be a *classical* proof string. The construction remains largely the same as above, so we will not reiterate all of it, only the relevant modifications. We need to permit the logical bits to be in either the $|0\rangle$ or $|1\rangle$ state – but not any superposition of the two, nor any entangled state.

One strategy to do this, if someone gave us a proof state to verify, would be to simply measure the proof state in the classical basis before proceeding with the computation. If they tried to decieve us by giving us anything other than a

classical basis state, this measurement would collapse it into a classical basis state – or, more precisely, a mixture of classical states.

But a constraint problem has no notion of "measurement". What we can do is request two copies of the state, and then perform a Swap Test on each pair of qubits. That is, for each qubit input to the problem, we take two copies of the qubit, and impose a clause

$$H_{Commit} = I - |00\rangle\langle 00| - |11\rangle\langle 11| \tag{2.24}$$

on them. This will force that measuring the second qubit in the classical basis puts the first qubit in a pure state; the first qubit cannot be in the $|+\rangle$ state or entangled with any other qubits in the input. The second qubit acts as a commitment qubit.

Again, we can't actually measure the second qubit, but if we make sure never to touch them for the duration of the computation, they are effectively removed from the problem, and we can trace them out to understand the computation. To prevent ourselves from using the commitment qubits in the computation, we can create a separate subspace of states for them, so that we can't get them confused with our actual logic qubits. So the construction is:

First, define two more basis states, $|0_P\rangle$ and $|1_P\rangle$, the commitment bit states. In addition to the H_{Start} clause that forces a bit to start off as zero, we define a 5-local clause $H_{Start-Unk}$ that starts a bit off in either $|0_L\rangle$ or $|1_L\rangle$. It classically

copies the logical bit to the commitment bit:

$$H_{Commit} = I - |0_L 0_P\rangle \langle 0_L 0_P| - |1_L 1_P\rangle \langle 1_L 1_P| \tag{2.25}$$

and then

$$\begin{aligned} H_{Start-Unk} = &(I - (|0_{EC}0_{CA}\rangle + |1_{EC}1_{CA}\rangle)(\langle 0_{EC}0_{CA}| + \langle 1_{EC}1_{CA}|))_{12} \\ &+ (|0_{CL}\rangle \langle 0_{CL}|) \otimes (I - |0_L\rangle \langle 0_L| - |1_L\rangle \langle 1_L|)_{24} \\ &+ ((I - |0_{CL}\rangle \langle 0_{CL}|) \otimes H_{Commit})_{135} \\ &+ \Big((|0_{CL}\rangle \langle 0_{CL}| + |1_{CL}\rangle \langle 1_{CL}| - |0_{CL}\rangle \langle 1_{CL}| - |1_{CL}\rangle \langle 0_{CL}|) \otimes |0_{CL}\rangle \langle 0_{CL}|\Big)_{23} \\ &+ H_{BP,23} \\ &+ H_{E,1} + H_{C,2} + H_{C,3} + H_{L,4} \end{aligned} \tag{2.26}$$

All lines except the second and third identical to H_{Start}. The second line requires that, at $t = 0$, the bit must be in only logical states 0 or 1, and not undefined. This is contrasted with H_{Start}, where we require that it be 0 specifically. The third line of (2.26) requires that, at $t = 0$, the bit must be equal to the commitment bit.

The commitment effectively turns our computation's initial state into a partial trace that removes the commitment qubits, which is a mixture of classical states. So we know that if the final computation succeeds, it is only because the input was a mixture of suceeding classical bitstrings – which would imply that there was indeed at least one valid proof string.

An equivalent setup (although not easily built as a QCSP) would be the having some k extra logical qubits next to some $k < n$ we normally work with, with a requirement that the first time uses gates to copy (with CNOT) from the originals to the extras; and the extras are never touched again. This setup is easy to analyze in the sense that we can discuss its history state. If the logical state of the verifying circuit at $t = 0$ is $|\psi_0\rangle = \sum_{b=0}^{2^k} \beta_b |0\rangle^{n-k} |b\rangle$, then the initial state of the enlarged system would be

$$\sum_{b=0}^{2^n} \beta_b |0\rangle^{n-k} |b\rangle |b\rangle$$

and the full history state is

$$\frac{1}{\sqrt{T}} \sum_{t=0}^{T} U_{1\ldots t} \sum_{b=0}^{2^n} \beta_b |0\rangle^{n-k} |b\rangle |b\rangle = \sum_{b=0}^{2^n} \beta_b \left(\frac{1}{\sqrt{T}} \sum_{t=0}^{T} U_{1\ldots t} |0\rangle^{n-k} |b\rangle \right) \otimes |b\rangle . \quad (2.27)$$

When the first half is examined in isolation, and the half consisting of the extra copies is removed, what remains is a classical mixture of history states run on different classical input strings $|b\rangle$.

Given a verifier circuit for a QCMA problem, we can embed it the same way we embedded in Part 2 for BQP problems. The witness bits (which are absent in BQP) get $H_{Start-Unk}$ instead of H_{Start}. Iff there is a frustration-free state, the bitstrings $|b\rangle$ in its support produce a passing output in the verifying circuit, in the same way that $|0\rangle$ produces a passing output in a BQP circuit. This correctness shows that the problem is QCMA_1-hard.

To see that the problem is also in QCMA_1, the proof is largely the same as the BQP_1 problem. In order to solve it in QCMA_1, we will accept a classical proof string built from concatenating the proof strings to each circuit chunk in the problem, and running each verifying circuit. The arguments about identifying chunks, clock paths, and undefined bits all go through as before. There is the new case where a single commitment bit is a copy of several distinct input bits, or even input bits in separate circuits. But this merely imposes the restriction that the classical proof string has equal bits in those two locations, which does not make the problem any harder to verify: when a proof string is provided, check for any commitment clauses that share the same logical qubits, and if those two bits in the proof string differ, reject. Otherwise, the problem is identical to the problem in BQP, except with this alternate provided input state, and the same algorithm can be used with this alternate initial state. This modified verification algorithm implies that the problem is in QCMA_1, and so is QCMA_1-complete.

2.7 coRP completeness

The reduction to a coRP-complete problem is even simpler, since we build on the idea of proving MA-hardness from [30]:

> Any classical MA verifier V can be transformed into a quantum verifier V' which uses a quantum circuit U involving only classical reversible gates (for example, the 3-qubit Toffoli gates) together with ancillary

states $|0\rangle$, $|+\rangle$, and measures one of the output qubits in the $|0\rangle$, $|1\rangle$ basis.

Note that MA is just the version of coRP that is allowed to have a proof string provided. (Really, this is MA_1; but $\mathsf{MA} = \mathsf{MA}_1$.) Thus, we can modify the BQP_1-complete problem as follows:

- Replace the universal quantum 2-qubit gate set with a universal classical reversible 3-bit gate set. This changes from H_{prop} from being 5-local to being 6-local.

- In addition to H_{Start}, we have a $H_{Start-Rand}$. They are identical except in that $H_{Start-Rand}$ initializes in the $|+\rangle$ state instead of $|0\rangle$.

The remainder of the proof holds just as before. The preprocessing is doable classically, and so can be executed by even a simple coRP machine. The final circuit to evaluate is a classical probabilistic verifier and so can be done by the coRP machine as well.

To be precise, the BQP proof is completely agnostic to the gate set, in the sense that it shows the clauses built from gate set $\mathcal{G}$ is a complete problem for computations with the gate set $\mathcal{G}$. It is also agnostic to how the initial state is specified, as long as there is no entanglement. Separately, it was shown in [30] that the gate set of the 3-qubit Toffoli gates is universal for classical probabilistic computation,

given access to initial states $|0\rangle$ and $|+\rangle$. Since the BQP-completeness proof is agnostic to these modifications, the resulting set of clauses is complete for coRP.

In [5] a coRP-complete problem involving local Hamiltonians is also constructed, using their notion of "pinned" Hamiltonians: these are problems with the promise that the ground state $|\psi\rangle$ has nonzero overlap with the all-zero state, i.e. $\langle 0|\psi\rangle > 0$. This pinning promise allows the verifier to assume that $|0\rangle$ is a functioning 'witness', reducing the complexity from MA to coRP. The "pinning" promise cannot be expressed in terms of local constraints, though, so it is not a QCSP in the sense we have defined it.

2.8 Weak QCSPs

As noted in section 3.1, there is the detail of gate sets and exact results. For classical computations, there are finite universal gate sets (such as $\{CCNOT\}$). Even for probabilistic computations, we only need uniformly random bits (or a 50-50 bit flip gate) to build a robust definition of BPP, RP, MA, AM, PP... and so on – as uniformly random bits can be used to closely approximate other probabilities, and we only need certain bounds on the probability of a given trial.

For quantum circuits, phases can cancel out exactly, so that a single phase gate or qubit distribution cannot suffice for exact computations. Since different

circuits might need all variety of phases, we cannot keep perfect exactness with a finite set while allowing all the circuits we might want; at the same time, allowing an infinite set of gates creates issues such as uncomputable amplitudes, and raises questions of how we are to encode the circuit.

The Solovay-Kitaev theorem gives a weaker, but arguably more natural, notion of universal gate set, one that allows us to approximate a given gate to exponentially good accuracy in polynomial time. This is not useful to use in studying constraint problems as we have defined them together, as we have required that all ground states be exact. To remedy this, we present the notion of a *weak quantum CSP*, which will allow exponentially small errors (energy) in the ground state.

Definition 2.9 (Weak Quantum CSP)**.** *A weak QCSP has a domain size d, a set of clauses $C = \{\mathcal{H}_i\}$, and constants a, b, and c, with $b > a > 0$. Each clause $\mathcal{H}_i$ of arity m is a Hermitian projector on $(\mathbb{C}^d)^{\otimes m}$.*

An instance of this weak QCSP is given by an integer n indicating the number of d-qudits, and a list of clauses that apply to n qudits. The instance is satisfiable if the ground state energy is less than a/n^c, and is unsatisfiable if the ground state energy is greater than b/n^c, and we are promised one of these is the case. (So, weak QCSPs are a class of promise problems.)

It is worth emphasizing that the constants a, b, and c are not allowed to vary with the instance, but are rather part of the language itself. This restriction means that every weak QCSP is in QMA, because we only need to measure the energy to within a polynomial gap: after $O(n^b)$ measurements we can have a probability of error bounded away from $1/2$. We choose to fix these constants for a whole constraint problem class, in contrast to k-Local-Hamiltonian in [78], where they are parameters of an instance.

At the same time, weak QCSPs are generally independent of gate set. While strong QCSPs naturally describe complexity classes with one-sided error, weak QCSPs naturally describe complexity classes with two-sided bounded error. If we have a quantum algorithm that uses a gate set $\mathcal{G}_1$ to solve a problem with two-sided bounded error in polynomially many gates, we can also solve with two-sided bounded error in polynomially man gates using any other universal gate set $\mathcal{G}_2$. This is shown by applying the Solovay-Kitaev theorem, to simulate the first gate set using the second one. The Solovay-Kitaev theorem provides $O(2^{-k})$ precision per gate with only $O(k)$ times more gates. If the original algorithm uses $f(n)$ gates, we can always solve the weak QCSP in $O(f(n)\log(f(n))$ gates with bounded error. The fact that we don't have to worry about exact gate set will make several complexity classes more natural, at the expense of a more complicated notion of "problem".

2.8.1 Results on weak problems

First, it is straightforward to verify that there are still P, NP, MA, and QMA complete problems among weak QCSPs. P-complete and NP-complete problems like Horn-SAT and 3-SAT are described by commuting projectors, so the ground state energy of an unsatisfiable instance is always at least 1. Thus taking $a = 1/3$, $b = 2/3, c = 0$ suffices to formulate those problems as weak QCSPs.

To see that we still have a QMA-complete problem, refer to the specifics of the proof that 3-Local-Hamiltonian is QMA-complete in [78]: their construction yields a problem whose ground state energy is below $\frac{c_1\varepsilon}{n}$ if satisfiable, or above $\frac{c_2}{n^3}$ is unsatisfiable, where ε is the algorithm's allowed probability of error. By running the algorithm $O(n)$ times in series, ε becomes $O(2^{-n}) < \frac{1}{n^2}$, and so it suffices to take $a = c_1$, $b = c_2$, $c = 3$. Then their construction permits the rewriting of any QMA circuit into an instance of this weak QCSP. We could also do this with QMA_1-complete $3 - \mathsf{QSAT}$ to arrive at QMA-complete weak QCSP.

This shows the scheme by which a $1/poly(n)$ gap promise translates into the weakness of the QCSP. [30] have the same gap for their their MA-complete problem, and so it also can be described as weak QCSP. Interestingly, in this case, $\mathsf{MA}=\mathsf{MA}_1$, and so the complexity class does not change; the same holds for $\mathsf{QCMA}=\mathsf{QCMA}_1$.

The other two cases we gave above, coRP and BQP_1 also all have polynomially small gaps, and are designed to emulate a circuit that accepts perfectly on accepting instances. If we instead allowed our original circuit to have a polynomially small *two*-sided error, we could build a weak QCSP instead, that is complete for the corresponding two-sided error complexity class. If the original QCSP had a minimum unsatisfiable ground-state energy of $O(n^{-p})$, then whatever circuit we are embedding as QCSP, let us repeat it enough times that its error is $O(n^{-p-1})$. Then choosing $c = p + 0.5$ and taking any positive a and b describe it as a weak QCSP. This gives the following result.

Corollary 2.1. *There are weak QCSPs that are complete for the classes* BPP *and* BQP*.*

This means that any putative dichotomy-like theorem for weak QCSPs would need to at least account for the seven cases of P, BPP, NP, MA, BQP, QCMA, QMA – or show that some pair of these are equal. The insensitivty to choice of gate set could make this type of result more appealingly natural, and indeed more physical, as no real-world gate set can be realized with zero error.

2.9 Universality of qubits for QCSPs

We can show that any QCSP can be reduced to another problem using only only qubits, with little computational power. This may initially sound unsurprising, as operations on qubits are certainly universal for quantum computation. But it is perhaps surprising, in light of the fact that the analogous statement is believed to be *false* in the classical world!

Among classical constraint problems, it is believed that distinct complexity classes arise for different size domains. For boolean constraint problems, it is known [7] that every problem is either coNLOGTIME, L-complete, NL-complete, $\oplus$L-complete, P-complete, or NP-complete. This is a refinement of the dichotomy theorem specialized to boolean problems, as the first five classes are all contained within P. Among ternary constraint problems though, there are new classes that appear, such as Mod_3L-complete, which are not expected to be equal to any of the six the previously listed [32, 14].

When constructing a circuit for a quantum computer, we can emulate a d-qudit with a $\lceil \log_2(d) \rceil$ qubits, and carry out unitaries on those qubits. We can certainly try the same thing for a QCSP, turning (for instance) each 4-qudit into 2 qubits, and a k-local clause becomes $2k$-local. The issue arises that we cannot ensure that the $2k$-local clauses are applied to qubits in a consistent fashion. One

clause might treat a particular qubit as "bit 1" of an 4-qudit, while another clause might use that same qubit as "bit 2". This would lead to constraints that were previously unrealizable. Clauses could also "mix and match", combining "bit 1" from one 4-qudit with "bit 2" from another 4-qudit. The exact same problems exist in the classical setting.

In the quantum world, we can fix this, again by using monogamy to bind together our constituent qubits into ordered, entangled larger systems. Each clause in the resulting problem is given a projector that forces this particular ordering of qubits, and any two clauses that try use the same qubits in multiple ways are frustrated. Formally,

NameIgnored 2 (Theorem 2.8 (formal))**.** *For any QCSP $\mathcal{C}$ on d-qudits, there is another QCSP $\mathcal{C}'$ on qubits, and AC^0 circuits f and g, such that f reduces $\mathcal{C}$ to $\mathcal{C}'$, and g reduces $\mathcal{C}'$ to $\mathcal{C}$. If $\mathcal{C}$ is k-local, then $\mathcal{C}'$ can be chosen to be $4 \cdot 2^{\lceil \log_2(\lceil \log_2(d) \rceil) \rceil} k$ local (that is, $O(\log(d))$ times larger.)*

Proof. First we will show that for any d-qudits, we can reduce to 4-qudits; after that we will reduce to qubits. Finally we show that the reduction is in AC^0.

We will view 4-qudit as the product of a "data" qubit and an "entanglement" qubit. A d-qudit will be replaced by $n = \lceil \log_2(d) \rceil$ many 4-qudits, and the state of the d-qudit will be encoded in the product of all the data qubits. If $d < 2^{\lceil \log_2(d) \rceil} = 2^n$, that is, if d is not exactly a power of 2, we will have a

Hamiltonian term T_1 in our clauses to ensure that the last $2^n - d$ states are not used. Acting on entanglement subspaces of the 4-qudits, consider a term T_2 whose nullspace consists of just the vector

$$\left(1 \otimes X^{\theta} \otimes X^{2\theta} \otimes \ldots X^{n\theta}\right) \frac{|0\rangle^n + |1\rangle^n}{\sqrt{2}} \tag{2.28}$$

where $\theta = \frac{1}{2(n+1)}$. This is a kind of GHZ state, which uses a slightly different pair of basis states (instead of just $|0\rangle$ and $|1\rangle$) on each separate qubit. Any bipartition of this state is impure, but since T_2 has a one-dimensional nullspace, it cannot be satisfied by any mixed state. Thus the sum of two T_2 on any two overlapping but distinct sets of 4-qudits will be frustrated. If two copies of T_2 act on the same 4-qudits in a different order, they will apply the wrong angles $X^{k\theta}$ at those sites, and the ground states do not align – also leading to frustration.

Each clause H of $\mathcal{C}$ is mapped to a new clause H' that acts as H on the data subspaces of each set of 4-qudits; that has $T_1 = 1 - \sum_i^d |i\rangle \langle i|$ on each clumping of 4-qudits, to ensure that only the first d states are used; and T_2 on each clumping of 4-qudits, to ensure that they will only be clustered to each other and in a particular order.

Then we want to reduce this from 4-qudits to qubits. As we show in Appendix A, there exists a particular Hamiltonian $H_{4\to 2}$ that satisfies certain properties: it acts on four qubits; it has a nullspace of dimension 4; if two copies of $H_{4\to 2}$ are applied to sets of qubits that overlap but are not equal, the result is frustrated; if

two copies of $H_{4\to 2}$ are applied to four qubits but in a different order, the result is frustrated. This lets us "fix" the groups of 4 qubits and they act as $d = 4$ qudits.

In the $\mathcal{C}'$ QCSP, any problem where qubits are mixed or applied in inconsistent orders, can immediately be rejected. Some qubits may not be acted on by any clause, and so not correspond to a d-qudit in $\mathcal{C}$, but then those qubits can simply be ignored. This leaves us with only correctly grouped qubits, in a certain subspace, that thus function equivalently to the d-qudits.

Combined, this gives a faithful reduction from d-qudits to qubits, and back again. It turns a k-local Hamiltonian into a $4\lceil \log_2(d) \rceil k$ local Hamiltonian, which is the majority of the theorem. It remains to check the complexity of the reductions, that the reduction can be carried out in AC^0.

In the above description, the expansion factor is $4\lceil \log_2(d) \rceil$. To get a low circuit complexity, we want the expansion to be a power of two. Thus we round this up to $4 \cdot 2^{\lceil \log_2(\lceil \log_2(d) \rceil) \rceil}$, which will denote x; the QCSP $\mathcal{C}'$ is augmented through just adding more qubits to increase the subspace dimension, and then T_1 is modified again to prevent those states from being occupied.

An instance of a k-local QCSP $\mathcal{C}$ can be written down as a list of integers, each given by an integer clause type, and k many integers representing the qudits they act on. The clause types of $\mathcal{C}'$ are in one-to-one correspondence with the clause types of $\mathcal{C}$, so those integers remain unchanged. Each clause acting on qudit i now

instead acts on qudits $[xi, xi+1, \ldots xi+(x-1)]$ in that order. Thus a reduction circuit f only needs to be able to replicate an integer several times, multiply by a constant power of 2, x, and add a number $i \in [0, x)$. This is in AC^0 (in fact it requires no gates at all).

For a circuit g to convert back from $\mathcal{C}'$ to $\mathcal{C}$, we need to map qubit numbers back to qudits numbers, and check that no qubits are used in inconsistent fashion. For each collection of qubits $[a_1, a_2, \ldots a_x]$ that a clause in $\mathcal{C}'$ is acting on, we can map that to the d-qudit number a_1. Thus, many qudit numbers will go unused, but this doesn't affect the correctness: as long as all collections use the same numbers in the same order, they will all be mapped to a_1. To check that all qudits are used in a consistent fashion, we need to check for each pair of collections $([a_1, \ldots a_x], [b_1, \ldots b_x])$ that they do not use qubits in inconsistently. Logically, this reads:

$$\big(((a_i = b_i) \wedge (a_j = b_j)) \vee ((a_i \neq b_i) \wedge (a_j \neq b_j))\big) \wedge (a_i \neq b_j) \wedge (a_j \neq b_i)$$

and this must be checked for every collection, for every i and j, and then combined by an unbounded fan-in AND. The integer equalities $a_i = b_i$ and $a_i \neq b_i$ can be evaluated with unbounded fan-in AND and OR respectively. This check is all in AC^0. If the check fails, the circuit outputs some fixed clause(s) that are unsatisfiable, otherwise it outputs a repetition of the first clause. This AC^0 circuit

checks that the qudits are used consistently, and if they are, gives an equivalent instance in the original language. □

If we don't care about having the reductions be in AC^0, and instead allow P-reductions, then $4\lceil \log_2(d) \rceil k$ locality suffices. This reduction is optimal within a factor of 4, in the sense that encoding one d-qudit in several qubits requires at least $\lceil \log_2(d) \rceil$ many qubits. In section 4 we showed that there is a 5-local 13-qudit problem that is BQP-complete. Together with Theorem 2.8, we have:

Corollary 2.2. *There is a* BQP*-complete QCSP on qubits, with 80-local interactions.*

In practice the locality could be reduced quite a bit, likely below 20 with work.

2.10 Future directions

The seven complexity classes that are known to occur as strong QCSPs are now, in rough order of difficulty:

1. P: Classical, no proof, deterministic checks.

2. coRP: Classical, no proof, probabilistic checks.

3. NP: Classical, classical proof, deterministic checks.

4. MA: Classical, classical proof, probabilistic checks.

5. BQP_1: Quantum, no proof, probabilistic checks.

6. QCMA: Quantum, classical proof, probabilistic checks.

7. QMA_1: Quantum, quantum proof, probabilistic checks.

Are there obvious omissions we should expect to look in, or does this list seem complete? It seems natural in one way: we can have classical or quantum verifiers; we can have no proofs, classical proofs, or quantum proofs; and we can have deterministic or probabilistic verification. This would produce 12 classes in total. But we cannot have a quantum proof for a classical verifier, bringing us down to 10 classes. And it is very hard to force deterministic verification on a quantum verifier; EQP is a difficult class to study. Any construction of an EQP-complete constraint problem would likely require knowledge of particular forms of circuits that are powerful enough to capture the full power of EQP, while constrained enough to guarantee that they always produce deterministic results.

It is clear, though, that the set of gates can be freely exchanged. There is the question if there are gate sets that are more powerful than classical computation but weaker than universal quantum computation. This is plausibly the case for nonstandard models of computation, such as sampling problems [25, 82] or one-clean-qubit models [54, 115], but is less clear for the standard quantum circuit

model. But if it was discovered that there was such a set of intermediate-power gates, we would immediately have a corresponding constraint problem class which captures its difficulty.

The same questions linger for weak QCSPs. EQP-complete problems are extremely unlikely to exist among weak QCSPs, as weak QCSPs do not depend on gate sets and EQP problems genearlly do. We can also ask about exotic gate sets yielding two-sided error complexity classes as weak QCSPs.

The constructions given herein is certainly very large, both in arity of the clauses and local dimension of the qudits. It is natural to ask if there are smaller and more natural constraint problems that could realize the classes we discussed. In particular, finding a 3- or even 2-local constraint would be exciting. This does not seem implausible given that 2-local Hamiltonian problems are in general QMA-hard. Ideally we would even have something analogous to Theorem 8 for locality reduction, that all QCSPs could be made 2-local using high-dimensional qudits. In classical CSPs these are well-studied under the name of *binary* constraint problems [38].

Finally, regarding coRP and BPP, it is somewhat surprising that these easy classical complexity class arises from a purely quantum Hamiltonian. It would be good to reformulate the problem as closely as possible in purely classical terms. Aharanov and Grilo have recently performed a reformulation like this for MA,

transforming the language of stoquastic Hamiltonian constraints into a much simpler classical problem [44].

Chapter 3

Quantum Bayesian Inference and Positive Semidefinite Permanents

[1]

[1] Based on the work in 10.1109/FOCS54457.2022.00013

3.1 Background: Bayesian Inference

Below is perhaps the most basic scientific task for finite probabilistic systems:

> A probabilistic process P produces samples from d distinct possible outcomes. Nothing is otherwise known about P. An experimenter takes n many samples from P, giving observation counts $(k_1, k_2, \ldots k_d)$, with $\sum n_i = n$. Given this experimental data, what should we estimate the distribution P to be?

The statement that "nothing is otherwise known" is a large weasel word. In principle, we need a *prior* belief of plausible different distributions P should be. A typical starting point is a uniform prior over P, which can be informally described as giving each distribution $(p_1, p_2, \ldots p_d)$ the same initial probability density, although other popular priors exist[76].

The question of "what should we estimate the distribution" is also not precise. Possible questions include, "What is the probability of the next sample being an m?", "What distribution P is most likely is given the data?", or "What is the probability that the true distribution P is within distance ϵ of my hypothesis P_0?". Under very reasonable and modest assumptions, these can all be computed efficiently[2]. This is the pleasant world of classical Bayesian statistics.

The quantum analog of this scenario would replace the d-case discrete distribution with a quantum state ρ in a d-dimensional Hilbert space. If all measurements

[2] Here "efficiently" means "in polynomial time in n and d". The relevant assumption is that the posterior is convex. This is always satisfied by the uniform prior, and it is satisfied by the Jeffreys prior described in [76] as long as each of the d possibilities has been observed at least once.

are taken in the standard basis (or are all in some other shared basis), this reduces to the classical case, just as we would want for a quantum generalization. When the measurement data come from different bases, the question is more complex. The three types of questions above now become, "What is the probability of observing m on a subsequent measurement M?", "What mixed state ρ best explains the data?", and "What is the probability that the true state ρ is within distance ϵ of my hypothesis ρ_0?". This is an idealized form of quantum state tomography, where we try to estimate a state by repeatedly preparing and measuring it.

As it turns out, the quantum nature of the problem makes it much more difficult, even for the simple case where all states are pure and all measurements are projective. The likelihood functions are no longer convex, and so the classical algorithms based on convexity don't apply, but this on its own does not preclude the possibility that some other algorithm could quickly reconstruct ρ. The main work of this chapter is giving complexity-theoretic reasoning to prove that this problem is difficult.

3.2 Background: Matrix Permanents

A different problem, at first glance unrelated, is that of computing of matrix permanents. The matrix permanent is a classical problem of intense interest in

the study of counting problems. For a matrix $M \in \mathbb{C}^{n\times n}$, the permanent is defined as

$$\mathrm{Perm}(A) = \sum_{\sigma\in S_n} \prod_{i=1}^{n} a_{i,\sigma(i)} \tag{3.1}$$

summing over all products of permutations of rows and columns. This equation bears a close resemblance to that of the determinant, which has an additional factor $(-1)^{\mathrm{sign}(\sigma)}$ within the sum. It is known that matrix permanents are hard to compute or even estimate in general, and we will review what is known about their hardness. It has been previously suggested that the permanents of positive semidefinite (PSD) matrices may have efficient approximations. By algebraically connecting these permanents to the quantum state estimation problem described above, we also show the hardness of PSD permanent approximation.

How difficult is it compute a permanent? The expression in Eq 3.1 can be directly evaluated in $O(n!)$ time, and this gives an exact answer. Ryser's formula[106] improves this to an $O(2^n n)$ time algorithm. Valiant showed in 1989 that computing the permanent exactly is #P-hard[3], even for 0-1 matrices[118, 18]. However, it is amenable to efficient approximation in particular settings, such as when all elements are nonnegative (and thus the permanent is as well). In 2001,

[3] #P is a complexity class far harder than NP, and is comparable to the difficulty of PH, the polynomial hierarchy. #P asks "How many solutions are there to this easily-checkable problem?". Finding just one solution is the task of NP, but #P requires us to reliably distinguish between the case of 10^{100} solutions or $1 + 10^{100}$ solutions. PH can be briefly described as "Can I guarantee a win in a board game in k moves?", for an arbitrary but fixed k.

Jerrum, Sinclair and Vigoda[72] gave a fully-polynomial randomized approximation scheme (FPRAS) for permanents of nonnegative matrices. In 2002, Gurvits and Samorodnitsky[58] gave a polynomial time e^n multiplicative approximation to PSD mixed discriminants, which included permanents of nonnegative matrices as a special case. (This special case was not useful in isolation, as the 2001 result gave arbitrarily good approximations.)

When the matrix is Hermitian positive semidefinite (HPSD, or if purely real, PSD), the permanent is again necessarily nonnegative (even though the individual entries are not), and this offers hope of efficient multiplicative approximation. HPSD permanents are of particular interest to the quantum information community - for reasons unrelated to quantum state tomography, but rather related to thermal BosonSampling experiments[116, 104, 79]. Computing PSD permanents exactly remains #P-hard[57]. It is known that by Stockmeyer counting[57, 104, 114] computing multiplicative approximations to PSD permanents is contained in $\mathsf{FBPP}^{\mathsf{NP}}$[4]. In 1963, Marcus[94] observed that the product of the diagonal of a PSD matrix immediately gives an $n!$ approximation ratio to the permanent. In 2017, [8] gave a polytime approximation to PSD permanents

[4] $\mathsf{FBPP}^{\mathsf{NP}}$ can be defined as follows: Given access to a special oracle that will instantly solve any NP problem for you, and a source of randomness, $\mathsf{FBPP}^{\mathsf{NP}}$ is the set of numeric functions you can compute the answer to in polynomial time with high confidence. The ability to repeatedly solve NP hard problems makes this class generally infeasible, but it is still much easier than #P or PH in general.

within a ratio of c^n with $c = e^{1+\gamma} \approx 4.85$. [130] described a similar approach with the same approximation ratio. [34] and [16] gave algorithms for approxmation when the spectrum of the matrix is small in radius, that is, when $\lambda_{min}/\lambda_{max}$ is not too small.

3.3 Statement of Main Results

Our main result is to show that there is no efficient approximation of PSD permanents. Stated precisely, we show that it is NP-hard to approximate within a particular subexponential factor:

Theorem 3.1 (*Thm 3.5, restated*)**.** *For any constant $\epsilon > 0$, it is NP-hard to approximate the permanent of $n \times n$ HPSD matrices within a factor of $2^{n^{1-\epsilon}}$.*

Note that the constant 2 can be exchanged for any $c > 1$ without changing the theorem. In Section 3.5.5, we show that this theorems also holds for the case of purely real PSD matrices.

Our work provides a lower bound on the difficulty of approximating PSD permanents, that almost matches known upper bounds. The algorithm of [8] shows that the singly exponential approximation ratio 4.85^n is possible within polynomial time, while we show that subexponential approximation ratio $2^{n^{1-\epsilon}}$ is intractable. This primarily leaves the question whether $(1+\epsilon)^n$ is polynomial-

time computable for any $\epsilon > 0$. The algorithms of [34] and [16] fail on the hard instances that we construct: the matrices we construct are highly rank deficient, and therefore have $\lambda_{min} = 0$.

We arrived at our hard instances via a problem in quantum state tomography. If a matrix M is positive semidefinite, then it has a matrix square root $VV^\dagger = M$. If M is $n \times n$, and rank d, then V is $n \times d$. We show that

$$\mathrm{Perm}(M) = \frac{(d+n-1)!}{2\pi^n} \int_{\vec{v}\in\mathbb{C}^n,\,|v|=1} d\vec{v} \prod_{i=1}^{d} |\vec{v}\cdot V_i|^2 .$$

Here V_i are the rows of V, and the integral is the Haar measure over the unit complex sphere. As we will see, this integral occurs naturally in the context of Bayesian inference, where the rows V_i correspond to an observation history. We analyze the problem by first establishing a concentrating construction (Lemmas 4 and 5). Informally, when V_i contains many copies of basis vectors $\vec{e}_j$ and vectors of the form $\frac{\vec{e}_j \pm i\vec{e}_k}{\sqrt{2}}$, the integral concentrates at the points (up to a phase) of an appropriately scaled hypercube:

$$\int_{\vec{v}\in\mathbb{C}^n,\,|v|=1} d\vec{v} \prod_{i=1}^{d} |\vec{v}\cdot V_i|^2 \approx C \sum_{\vec{v}\in\{-1,+1\}^d} \prod_{i} |\vec{v}\cdot V_i|^2$$

for some constant C that depends only on d and n. This concentration will let us relate permanents to combinatorial problems (Lemma 6), specifically counting solutions to Not-All-Equal-3SAT, and ultimately let us prove hardness.

The connection to quantum state tomography means we also get results about the hardness of estimating quantum states given measurements. For a quantum system with Hilbert space dimension n and poly(n) observations, the *maximum pure state likelihood* is the highest likelihood of those observations attainable over any pure state $|\psi\rangle$.

Theorem 3.2 (*Thm 3.9, informal*)**.** *For any constant $\epsilon > 0$, the following task is NP-hard: given a series of quantum observations, find a pure state with likelihood at least $2^{-n^{1-\epsilon}}$ times the maximum pure state likelihood.*

Unless RP=NP, this implies that there is no PRAS for maximum likelihood estimation (MLE) quantum state tomography (in fact, it is not even in APX). We have similar statements about the NP-hardness of computing the Bayesian average state and Bayesian average observables (Theorem 3.8). These results are unusual in that they imply exponential difficulty in dimension n in the Hilbert space $\mathbb{C}^n$. Most quantum problems are only considered tractable if they have efficient algorithms in the number of particles $q = \log(n)$, and have trivially polynomial solutions in n; whereas we show that (assuming ETH[70]) quantum state tomography takes time exponential in n.

We stress that although our work has connections to quantum information through BosonSampling and tomography, our discussion of complexity is focused on classical computers. The NP-hardness are statements about classical hardness,

and the algorithm described in section 3.6.3 for tomography in fixed dimension is a polynomial time *classical* algorithm. Unless $\mathsf{NP} \subseteq \mathsf{BQP}$ however, our results rule efficient permanent computations on quantum computers as well.

3.4 Key ideas of the proof

We start with a lemma relating symmetric, multilinear functions to permanents. Similar lemmas have appeared in [15, 16], and they can broadly be viewed as alternate forms of Wick's Theorem [133].

Theorem 3. Suppose $f : (\mathbb{C}^d)^{2n} \to \mathbb{R}$ is a function of $2n$ vectors of dimension d, with the properties:

- Multilinear in its first n arguments:

$$f(v_1, \ldots, \alpha v_i + \beta v_i', \ldots) = \alpha f(v_1, \ldots, v_i, \ldots) + \beta f(v_1, \ldots, v_i', \ldots)$$

- Conjugate multilinear in its latter n arguments:

$$f(v_1, \ldots, \alpha v_{n+i} + \beta v_{n+i}', \ldots) = \alpha^* f(v_1, \ldots, v_{n+i}, \ldots) + \beta^* f(v_1, \ldots, v_{n+i}', \ldots)$$

- Symmetric in its first n arguments, and its latter n arguments:

$$f(v_1, v_2, \ldots; v_n, v_{n+1}, \ldots) = f(v_{\sigma(1)}, v_{\sigma(2)}, \ldots, v_{\tau(n)}, v_{\tau(n+1)})$$

- Invariant under unitary change of basis: for any unitary $U \in \mathbb{C}^{d\times d}$,

$$f(v_1, v_2, \dots; v_n, v_{n+1}, \dots) = f(Uv_1, Uv_2, \dots; Uv_n, Uv_{n+1}, \dots)$$

Then f is determined up to an overall constant C by the formula,

$$f(v_1, \dots; v_n, \dots) = C \operatorname{Perm}(A_{ij}), \quad \text{where } A_{ij} = v_i \cdot v_j^* \tag{3.2}$$

and the constant C can be determined by

$$C = \frac{f(\vec{e}_1, \vec{e}_1, \vec{e}_1, \dots)}{n!} \tag{3.3}$$

where $\vec{e}_1$ is the unit basis vector in the first coordinate.

Proof. Because f is invariant under a unitary change of basis, f can only depend on its inputs through inner products of vectors, $\langle v_i, v_j \rangle$. Since f is multilinear, it can be written as a sum of terms t_k, where each t_k is a product of terms from the vectors. The separate linearity and conjugate linearity means that the only permitted inner products are of covariant (first n) and contravariant (latter n) vectors. This means every term in the sum must be some product of the form $\prod_{i\in[n]} v_i \cdot v^*_{n+\sigma(i)}$ for some permutation of n. Then by symmetry of the arguments, all pairs must occur in the same relation to either, so all pairings must occur equally. This leaves only a single form, the result above.

Computing C can be found by substituting in $\vec{e}_1$ in 3.2 so that all dot products become 1. The permanent of the all-1's matrix is just $n!$, so this becomes the normalizing factor. □ □

This lets us relate the permanent to a particular integral over unit-norm complex vectors:

Theorem 3.3. *For any $L, R \in \mathbb{C}^{d\times n}$ be complex matrices, denoting the kth row as L_k,*

$$\int_{\vec{v}\in\mathbb{C}^n,\,|v|=1} d\vec{v} \left(\prod_{k=1}^{d} \vec{x}^\dagger L_k\right)\left(\prod_k R_k^\dagger \vec{x}\right) = \frac{2\pi^n \operatorname{Perm}(LR^\dagger)}{(d+n-1)!} \tag{3.4}$$

Note that when $L = R$, the product in the integral becomes $\prod_k |\langle L_k, \vec{x}\rangle|^2$, and the product $M = LL^\dagger$ is PSD.

Proof. Viewing the left side as a function f of the n rows of each L and R, we can see that it satisfies all the hypotheses of Lemma 3. It is linear in each row of L, conjugate linear in each row of R, and symmetric under permuting the rows of L or the rows of R. It is also invariant under a unitary change of basis:

$$\begin{aligned} f(UL, UR) &= \int_{\vec{v}\in\mathbb{C}^n,\,|v|=1} d\vec{v} \prod_k \vec{v}^\dagger (UL_k) \prod_k (UR_k)^\dagger \vec{v} \\ &= \int_{\vec{v}\in\mathbb{C}^n,\,|v|=1} d\vec{v} \prod_k (U^\dagger \vec{v})^\dagger L_k \prod_k R_k^\dagger (U^\dagger \vec{v}) \\ &= \int_{\vec{u}\in\mathbb{C}^n,\,|u|=1} d\vec{v} \prod_k \vec{u}^\dagger L_k \prod_k R_k^\dagger \vec{u} = f(L, R) \end{aligned}$$

so that we've used the symmetry of the unit sphere in $\mathbb{C}^n$ to remove the unitary via $\vec{u} = U\vec{v}$. Setting each $L_k = R_k = \vec{e}_1$, the spherical integral can be computed with standard formulae (e.g. [53]) to find the normalizing constant

$$C = \frac{1}{n!}\int_{\vec{v}\in\mathbb{C}^n,\,|v|=1} d\vec{v}\ (\vec{v}^\dagger \vec{e}_1)^n (\vec{e}_1^\dagger \vec{v})^n = \frac{1}{n!}\cdot\frac{2\pi^d n!}{(d+n-1)!} = \frac{2\pi^n}{(d+n-1)!}. \tag{3.5}$$

□ □

This formula is similar to another well-known expression for the permanent involving Gaussian integrals, and can be understood as a version of Wick's theorem. [15, 133]

3.4.1 Construction Details

Before diving into the proof of hardness itself, we aim to provide some intuition of the construction. We focus on the integral $F = \int_{\vec{x}} \prod_k |\langle V_k, \vec{x}\rangle|^2$ over a the sphere of unit (complex) vectors, and build up a set of vectors V with desirable properties. The proof will involve gradually adding vectors to a list V_k, in turn modifying the integrand $I_V(\vec{x}) = \prod_k |\langle V_k, \vec{x}\rangle|^2$. This integrand $I_V(\vec{x})$ is nonnegative, so there cannot be any cancellation in the integral. Our goal will be only showing that certain regions have exponentially small magnitude, so that only particular regions with appreciable contribution remain, and they are primarily responsible for the overall value of F. Then, the magnitude of F will be used to understand the value of I on those particular regions, where large values of F indicate solutions to an NP-hard problem. And since F can be computed by a HPSD permanent, computing that permanent must be hard as well.

How are we to choose the V in order to make an interesting function I_V? Each vector V_k introduces zeroes on the sphere at all vectors orthogonal to V_k.

All points approximately orthogonal to V_k will have a very small magnitude, and so contribute very little to the integral. We will start our collection of vectors includes many copies of each standard basis vector $\vec{e}_k$. This creates high-degree zeros along each of d distinct perpendicular directions, slicing the sphere so that the only regions with appreciable magnitude form the corners of a cube.

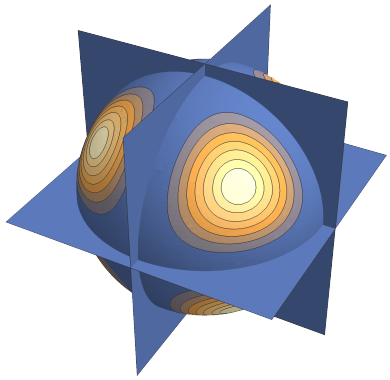

Figure 3.1: Schematic of how we can create "corners" on the sphere by repeatedly cutting with planes. Blue represents lower magnitude. This shows only purely real $\vec{x}$.

After adding one copy of each basis vector $\vec{e}_k$, the magnitude at a given point $\vec{x} = (\alpha_1, \ldots \alpha_d)$ is the product of the absolute values of its entries in that basis: $I_V(\vec{x}) = \prod_j |\alpha_j|^2$. This is maximized when $|\alpha_j| = |\alpha_k| = \frac{1}{\sqrt{d}}$ for all j, k. If we then subsequently add several vectors of the form $\frac{\vec{e}_j + i\vec{e}_k}{\sqrt{2}}$ and $\frac{\vec{e}_j - i\vec{e}_k}{\sqrt{2}}$, together these rule out a purely imaginary phase between the j and k components, so that the maxima are at $\frac{\vec{e}_j \pm \vec{e}_k}{\sqrt{2}}$. After adding these two for each $j \neq k$, $I(\vec{x})$ will peak near $\vec{x} = \frac{e^{i\theta}}{\sqrt{d}}(1, \pm 1, \pm 1 \ldots)$. Up to an overall phase of $\vec{x}$, we've focused I to a set of

2^{d-1} distinct points. These 2^{d-1} circles of "binarized" vectors form a set B_0. To get this, we had to put $d + 2\binom{d}{2} = d^2$ vectors into V_k. By analogy with quantum information, we will refer to these as the Z vectors and Y vectors respectively. Together, this set of d^2 vectors will form one "basic set" – "basic" in the set of "enforcing the basis".

Once we have our basic vectors to concentrate I at these binarized points B_0, we want to add vectors that will penalize some of these 2^{d-1} points, so that finding the optimum becomes a search problem over exponentially many points. Our functional I is only sensitive to the relative phase between components of a vector, and not to the signs of the components themselves. This leads us most naturally to the problem of Not-All-Equal 3-Satisfiability, or NAE3SAT.[107] So now consider the impact of adding a triple of "clause vectors",

$$\vec{v}_1 = \frac{\vec{e}_1 + \vec{e}_2 - 2\vec{e}_3}{\sqrt{6}}$$

$$\vec{v}_2 = \frac{\vec{e}_1 - 2\vec{e}_2 + \vec{e}_3}{\sqrt{6}}$$

$$\vec{v}_3 = \frac{-2\vec{e}_1 + \vec{e}_2 + \vec{e}_3}{\sqrt{6}}.$$

Each is orthogonal to $\frac{1}{\sqrt{3}}(\vec{e}_1 + \vec{e}_2 + \vec{e}_2)$, in which all the relative signs are positive (or equivalently, all negative). We call this collection of three vectors a "clause set". This effectively rules out the possibility of all signs being the same. There

are three not-all-equal points (up to phase):

$$\vec{p}_1 = \frac{1}{\sqrt{3}}(\vec{e}_1 + \vec{e}_2 - \vec{e}_3)$$

$$\vec{p}_2 = \frac{1}{\sqrt{3}}(\vec{e}_1 - \vec{e}_2 + \vec{e}_3)$$

$$\vec{p}_3 = \frac{1}{\sqrt{3}}(-\vec{e}_1 + \vec{e}_2 + \vec{e}_3)$$

The three $\vec{p}_i$ all have the same squared inner products with the set of $\vec{v}_i$, those being $\{\frac{8}{9}, \frac{2}{9}, \frac{2}{9}\}$ in some order, and so all $\vec{p}_i$ have an equal $I_V(\vec{p}_i) = \frac{32}{729}$.

The total effect may be visualized in the following plot:

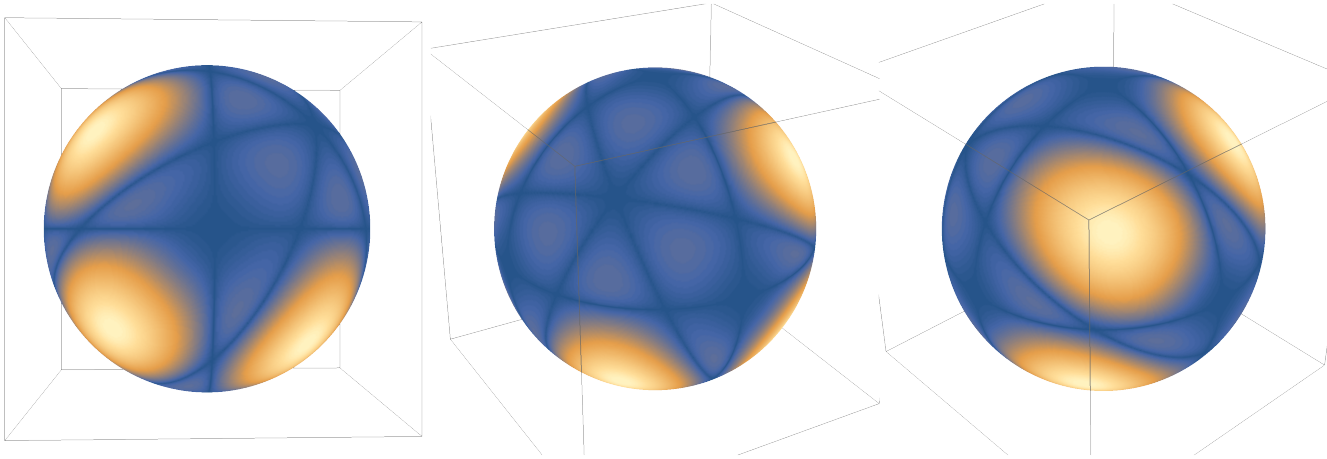

Figure 3.2: Three plots of $I_V(\vec{x})$. Only real points are plotted, smaller values are blue. As the integrand only depends on points up to an overall phase, all points appear effectively doubled, as $I_V(\vec{x}) = I_V(-\vec{x})$. There are twelve vectors in V. Nine come from a basic set: $\vec{e}_1$, $\vec{e}_2$, $\vec{e}_3$, and $(\vec{e}_j \pm i\vec{e}_k)/\sqrt{2}$ in six permutations (given by $j, k \in \{1, 2, 3\}$, $j < k$). The right-angled crosses are due to the first three vectors, dividing the space into eight corners (the Z vectors). The last three vectors in V are a clause set: $(\vec{e}_j + \vec{e}_k - 2\vec{e}_\ell)/\sqrt{6}$ (in all 3 permutations), creating the 6-way intersection shown in the second diagram, eliminating two opposing corners of the eight.

Here we look at $d = 3$ with one basic set and one clause set. The plot doesn't display states with complex coefficients, but it can be verified that the maxima have all real phases. Without the clause vectors, there would be eight high amplitude points. The first subplot shows the effect of the clause most directly: out of the four points (up to sign) in B_0, one of them – the top right corner – has been eliminated. The excluded option has three zero planes running through it.

By adding appropriate clause sets, the only remaining points with large values will be those satisfying an NAE3SAT problem, which is NP-hard. The other points will be too small to contribute to the integral, so that evaluating the integral tells us about the satisfiability of the NAE3SAT problem. With the outline complete, we now begin the steps of the proof, starting with the concentration.

3.5 Proof of Hardness

3.5.1 Concetration

After one basic set, each point in B_0 has a value $I_V(\vec{x})$ of $1/d^{d^2}$ (by direct calculation). We would like to show that any state far away from B_0 has a significantly lower value. For this reason, and with the intuition that the integrand I_V represents likelihood values, we talk about *relative* values of I_V. By the value of $I_V(a)$ relative to $I_V(b)$, we simply mean $I_V(a)/I_V(b)$.

Any unit vector $\vec{x} \in \mathbb{C}^d$ can be written as

$$\vec{x} = \frac{e^{i\Theta}}{\sqrt{d}} \sum_{k=1}^{d} \sqrt{\alpha_k} e^{i\pi(\theta_k + n_k)} \vec{e}_k$$

where Θ, $\vec{\alpha}_k$ and $\vec{\theta}_k$ are all real, $\alpha_k \geq 0$, $\sum_k \alpha_k = d$, $\theta_1 = 0$, and all Θ, $\theta_k \in [-1/2, 1/2])$, and $n_k \in \{0, 1\}$. The $\vec{\alpha}$, $\vec{\theta}$, and $\vec{n}$ respectively indicate the amplitudes, phases relative to the first component, and signs of the real part. This polar representation is unique except for when one of the $\alpha_0 = 0$, which is a measure-zero set. Accordingly, we can neglect this measure zero set in subsequent discussions of the integral $\int I_V(x)$ – as we could otherwise arbitrarily set $I_V(x) = 0$ there without modifying the integral.

Theorem 4. *Let $\vec{x}$ be a unit vector with polar representation Θ, $\vec{\alpha}$, $\vec{\theta}$, and $\vec{n}$. Let ϵ_α be the 2-norm distance of $\vec{\alpha} = (\alpha_1, \ldots \alpha_d)$ from $\vec{1}$. Then when V is one basic set, the value of $I_V(\vec{x})$ relative to any point in B_0, is at most $1 - \frac{\epsilon_\alpha^2}{4d}$. If $\epsilon_\alpha \leq 1/2$, then the likelihood is also at most $1 - 3\theta_i^2$ for all components θ_i of $\vec{\theta}$.*

Proof. Then B_0 consists of the points with $\alpha_k = 1$ and θ_k is an integer. If $\vec{x}$ has significant distance from all elements of B_0, then either the amplitudes α_k or phases θ_k must differ significantly from these conditions. The likelihood after of the measurements is

$$L(\psi) = \left(\prod_k \left| \sqrt{\frac{\alpha_k}{d}} \right|^2 \right) \left(\prod_{j \leq k} \left| \frac{\sqrt{\alpha_j} e^{i\pi(\theta_j + n_j)} + i\sqrt{\alpha_k} e^{i\pi(\theta_k + n_k)}}{\sqrt{2d}} \right|^2 \left| \frac{\sqrt{\alpha_j} e^{i\pi(\theta_j + n_j)} - i\sqrt{\alpha_k} e^{i\pi(\theta_k + n_k)}}{\sqrt{2d}} \right|^2 \right)$$

$$= \left(\prod_k \frac{\alpha_k}{d^d}\right)\left(\prod_{j\leq k} \frac{\alpha_j^2 + \alpha_k^2 + 2\alpha_j\alpha_k\cos(2\pi(\theta_j - \theta_k + n_j - n_k))}{4d^2}\right)$$

$$= \frac{1}{d^d(2d)^{d^2-d}}\left(\prod_k \alpha_k\right)\left(\prod_{j\leq k} \alpha_j^2 + \alpha_k^2 + 2\alpha_j\alpha_k\cos(2\pi(\theta_j - \theta_k))\right)$$

The first factor coming from the Z vectors $\vec{e}_k$ in the basic set, and the last two factors coming from the Y vectors $\frac{\vec{e}_j \pm i\vec{e}_k}{\sqrt{2}}$, for each $j < k$, in the basic set.

The first step is to bound the likelihood in terms of the magnitudes α_k. Looking at the effect of the Z vectors, $\prod_k^d \alpha_k$, we have a convex function on the standard $(d-1)$-simplex $\sum \alpha_k = d$. It is clearly maximized at $\vec{\alpha}_{opt} = (1, 1, 1, \ldots 1)$, where it evaluates to 1. Suppose that our state $|\psi\rangle$ has an associated α-vector, $\vec{\alpha} = (\alpha_1, \ldots \alpha_d)$ is a distance at least ϵ_α away from $\vec{\alpha}_{opt}$, and that $\epsilon_\alpha \leq 1$. Then one of the coordinates must be at least $\epsilon_\alpha/\sqrt{d}$ away from 1. With generality, let this coordinate be α_1. If $\alpha_1 \leq 1 - \epsilon_\alpha/\sqrt{d}$, then the greatest the likelihood could still be is when the other α_k are all equal at $1 + \epsilon_\alpha/\sqrt{d}(d-1)$. Multiplying these together, the resulting likelihood is upper-bounded by $1 - \frac{\epsilon_\alpha^2}{2(d-1)}$. If α_1 has instead been increased so that $\alpha_1 \geq 1 + \epsilon_\alpha/\sqrt{d}$, then the likelihood is maximized when the other α_k are all equal at $1 - \epsilon_\alpha/\sqrt{d}(d-1)$. Multiplying these together, the resulting likelihood is upper-bounded by $1 - \frac{\epsilon_\alpha^2}{4d}$. Since the latter of these bounds is looser, we see that any state whose $\vec{\alpha}$ is at least ϵ_α away from the all-ones vector has a likelihood at most $1 - \frac{\epsilon^2}{4d}$ in these measurements.

This gives bounds on the Z vectors' contribution to the likelihood. To keep this bound when the Y vectors are added, we need to check that they are also maximized at $\vec{\alpha} = \vec{1}$. Each factor

$$\prod_{j \leq k} \alpha_j^2 + \alpha_k^2 + 2\alpha_j \alpha_k \cos(2\pi(\theta_j - \theta_k))$$

is maximized when $\theta_j - \theta_k$ is an integer, at which point it becomes $\prod_{j \leq k} (\alpha_j + \alpha_k)^2 = (\prod \alpha_j + \alpha_k)^2$. This is in turn globally maximized by $\alpha_j = \alpha_k = 1$, so the error bound on $\vec{\alpha}$ holds.

The next step is to bound the likelihood in terms of the $\vec{\theta}$. We only care about the degree to which $\theta_i - \theta_j$ is not an integer, let $r_{ij} = \theta_i - \theta_j$ to the nearest integer, so $r_{ij} \in [-1/2, 1/2]$. Given that $\cos(2\pi r) \leq 1 - 8r^2$ for all $r \in [-1/2, 1/2]$, we have a relative likelihood of

$$\frac{I_V(\vec{x})}{I_V(B_0)} = \frac{\alpha_j^2 + \alpha_k^2 + 2\alpha_j \alpha_k \cos(2\pi r_{jk})}{\alpha_j^2 + \alpha_k^2 + 2\alpha_j \alpha_k} \leq \frac{\alpha_j^2 + \alpha_k^2 + 2\alpha_j \alpha_k (1 - 8r_{jk}^2)}{\alpha_j^2 + \alpha_k^2 + 2\alpha_j \alpha_k} = 1 - \frac{16\alpha_j \alpha_k}{(\alpha_j + \alpha_k)^2} r_{jk}^2$$

Let's assume that each α_j is in the interval $[1/2, 3/2]$ – which is implied by them being sufficiently close to the all-ones vector, that is, $\epsilon_\alpha \leq 1/2$. Then the expression $\frac{16\alpha_j \alpha_k}{(\alpha_j + \alpha_k)^2}$ is at least 3, so

$$\frac{I_V(\vec{x})}{I_V(B_0)} \leq 1 - 3r_{jk}^2$$

which tells us that every phase θ_i should be close to 0 for I_V to be large, or else suffer a $1 - 3r^2$ penalty in the likelihood. □ □

Later we will also need *lower* bounds on the likelihood, if we *are* in U.

Theorem 5. *If a state $\vec{x}$ is within distance $\epsilon \leq 0.1$ of some point b in B_0, and V is one set of basic vectors, then $\vec{x}$ has likelihood at least*

$$I_V(\vec{x}) \geq \frac{1 - 2\epsilon d^{5/2}}{d^{d^2}}.$$

or in terms of the relative value, $I_V(\vec{x})/I_V(B_0) \geq 1 - 2\epsilon d^{5/2}$.

Proof. We will again use polar representation for $\vec{x}$:

$$I_V(\vec{x}) = \frac{1}{d^d (2d)^{d^2-d}} \left(\prod_k \alpha_k \right) \left(\prod_{j \leq k} \alpha_j^2 + \alpha_k^2 + 2\alpha_j \alpha_k \cos(2\pi(\theta_j - \theta_k)) \right)$$

If our point $\vec{x}$ is within distance $\epsilon < 1$ of B_0, then each of the α_i must individually be within $\epsilon\sqrt{d}$ of 1, and each θ_i satisfies

$$\cos(\pi\theta_i) > \sqrt{1-\epsilon^2} \implies |\theta_i| < \epsilon/2$$

and so

$$\cos(2\pi(\theta_j - \theta_k)) \geq 1 - \frac{(2\pi(\theta_j - \theta_k))^2}{2} \geq 1 - \frac{(2\pi\epsilon)^2}{2}.$$

Then the likelihood is bounded by,

$$L(\psi) \geq \frac{1}{d^d (2d)^{d^2-d}} \left(\prod_k 1 - \epsilon\sqrt{d} \right) \left(\prod_{j \leq k} (1 - \epsilon\sqrt{d})^2 + (1 - \epsilon\sqrt{d})^2 + 2(1 - \epsilon\sqrt{d})^2 \left(1 - \frac{(2\pi\epsilon)^2}{2} \right) \right)$$

$$= \frac{1}{d^d (2d)^{d^2-d}} (1 - \epsilon\sqrt{d})^{(d^2+d)/2} \left(4 - 4\pi^2\epsilon^2 \right)^{(d^2-d)/2}$$

$$\geq \frac{1}{d^{d^2}} \left(1 - \epsilon\sqrt{d}(d^2+d)/2 - \pi^2\epsilon^2(d^2-d)/2 \right)$$

If $\epsilon < \sqrt{d}/\pi^2$, which is implied by $\epsilon < 0.1$, then the term $\pi^2\epsilon^2(d^2 - d)$ is smaller than $\epsilon\sqrt{d}(d^2 + d)/2$, so we can combine the two. We can also bound $d^2 + d < 2d^2$.

$$L(\psi) \geq \frac{1}{d^{d^2}} \left(1 - 2\epsilon\sqrt{d}(d^2 + d)/2\right) = \frac{1 - 2\epsilon d^{5/2}}{d^{d^2}}$$

□ □

Together, these two lemmas establish a form of concentration: points close to B_0 have large (lower-bounded) values of I_V, and points far from B_0 have small (upper-bounded) values of I_V.

3.5.2 Restricting to neighborhoods of G

Now we consider the effect of clause sets. A clause C is defined by a triple of integers $(\mathsf{C}_1, \mathsf{C}_2, \mathsf{C}_3)$. A point $b \in B_0$ with coordinates $(b_1, b_2, \ldots b_d)$, each $b_k = \pm e^{i\Theta}$, is "good" for the clause C if $\{b_{\mathsf{C}_1}, b_{\mathsf{C}_2}, b_{\mathsf{C}_3}\}$ are not all equal. A point in B_0 is "good" for a set of clauses if it is good for each of them, and a point is "bad" if it is not good. Each clause C has an associated set of three clause vectors

$$\vec{v}_1 = \frac{\vec{e}_{\mathsf{C}_1} + \vec{e}_{\mathsf{C}_2} - 2\vec{e}_{\mathsf{C}_3}}{\sqrt{6}}$$

$$\vec{v}_2 = \frac{\vec{e}_{\mathsf{C}_1} - 2\vec{e}_{\mathsf{C}_2} + \vec{e}_{\mathsf{C}_3}}{\sqrt{6}}$$

$$\vec{v}_3 = \frac{-2\vec{e}_{\mathsf{C}_1} + \vec{e}_{\mathsf{C}_2} + \vec{e}_{\mathsf{C}_3}}{\sqrt{6}}.$$

Theorem 6. *Take a clause* $\mathsf{C} = (\mathsf{C}_1, \mathsf{C}_2, \mathsf{C}_3)$ *and let* V *be its three clause vectors. Nowhere does* I_V *exceed 1. At any point* $\vec{x}$ *within a distance* ϵ *of a good point,* $I_V(\vec{x}) \geq \frac{32}{27d^3}\left(1 - 12\epsilon\sqrt{d}\right)$. *At any point* $\vec{x}$ *within a distance* ϵ *of a bad point,* $I_V(\vec{x}) \leq \frac{4096}{27}\epsilon^6$.

Proof. To see that 1 is an upped bound on I_V, note that I_V is a product of dot products of unit vectors, each of which is at most 1, so that $I_V \leq 1$.

For the second claim, we have a point $\vec{x}$ close to a good point $\vec{g}$. Since we only care about the value of I_V and the distance between $|\vec{x} - \vec{g}|$, we may adjust the phase of $\vec{x}$ and $\vec{g}$ jointly so that $\vec{g}$ is entirely real, and all of its entries are ± 1. We decompose $\vec{x}$ in the form

$$\vec{x} = \alpha\vec{e}_{\mathsf{C}_1} + \beta\vec{e}_{\mathsf{C}_2} + \gamma\vec{e}_{\mathsf{C}_3} + \Delta\vec{x}_\perp$$

Then the impact of the three clause vectors is,

$$I_V(\vec{x}) = \frac{1}{6^3}|\alpha + \beta - 2\gamma|^2 \cdot |\alpha - 2\beta + \gamma|^2 \cdot |-2\alpha + \beta + \gamma|^2$$

We seek to bound this value in the vicinity of good points. A good B_0 point has not all signs equal. Since we can permute the elements of C without affecting the value of I_V, a general good point $\vec{g}$ can be written as

$$\vec{g} = \frac{1}{\sqrt{d}}\left(-\vec{e}_{\mathsf{C}_1} + \vec{e}_{\mathsf{C}_2} + \vec{e}_{\mathsf{C}_3} + \sqrt{d-3}\,\vec{g}_\perp\right)$$

where $\vec{g}_\perp$ contains the support on all the other basis vectors. It has $I_V(\vec{g}) = \frac{32}{27d^3}$, by direct computation. Then for our other point $\vec{x}$ within a distance ϵ of $\vec{g}$, each coordinate must also be within ϵ of the corresponding coordinate in $\vec{g}$. So

$$\Re[\alpha+\beta-2\gamma] \leq \frac{1}{\sqrt{d}}\left((-1+\epsilon\sqrt{d})+(1+\epsilon\sqrt{d})-2(1-\epsilon\sqrt{d})\right) = -2(1-2\epsilon\sqrt{d})/\sqrt{d}$$

and similarly

$$\Re[-2\alpha+\beta+\gamma] \geq \frac{1}{\sqrt{d}}\left(-2(-1+\epsilon\sqrt{d})+(1-\epsilon\sqrt{d})+(1-\epsilon\sqrt{d})\right) = 4(1-\epsilon\sqrt{d})/\sqrt{d} \geq 4(1-2\epsilon\sqrt{d})/\sqrt{d}$$

Putting together the six factors,

$$\begin{aligned} I_V(\vec{x}) &= \frac{1}{6^3}|\alpha+\beta-2\gamma|^2 \cdot |\alpha-2\beta+\gamma|^2 \cdot |-2\alpha+\beta+\gamma|^2 && (3.6)\\ &\geq \frac{1}{6^3}\Re[\alpha+\beta-2\gamma]^2\Re[\alpha-2\beta+\gamma]^2\Re[-2\alpha+\beta+\gamma]^2 && (3.7)\\ &\geq \frac{1}{6^3}\frac{32}{27d^3} \times \left(1-2\epsilon\sqrt{d}\right)^6 && (3.8)\\ &\geq \frac{1}{6^3}\frac{32}{27d^3} \times \left(1-12\epsilon\sqrt{d}\right) && (3.9) \end{aligned}$$

which is the second claim. For the third claim, take a bad point $\vec{h}$ in B_0, for which we can correct the phase to put it in the form

$$\vec{h} = \frac{1}{\sqrt{d}}\left(+\vec{e}_{\mathsf{C}_1}+\vec{e}_{\mathsf{C}_2}+\vec{e}_{\mathsf{C}_3}+\sqrt{d-3}\,\vec{h}_\perp\right)$$

Then for a nearby point only ϵ away, each coordinate is at most ϵ away. This means

$$\Re[\alpha+\beta-2\gamma] \leq \left(\frac{1}{\sqrt{d}}+\epsilon\right)+\left(\frac{1}{\sqrt{d}}+\epsilon\right)+\left(\frac{-2}{\sqrt{d}}+2\epsilon\right) = 4\epsilon$$

$$\Im[\alpha + \beta - 2\gamma] \leq 4\epsilon$$

$$\implies |\alpha + \beta - 2\gamma|^2 \leq 32\epsilon^2$$

and similarly for the other two permutations, so that

$$I_V(\vec{x}) \leq \frac{1}{6^3}(32\epsilon^2)^3 = \frac{4096}{27}\epsilon^6.$$

□ □

3.5.3 $F = \int_x I_V(\vec{x})$ Approximates #NAE3SAT

With these bounds, we will be able to relate the number of solutions to a NAE3SAT instance to the integral $F = \int_x I_V(\vec{x})$.

Theorem 3.4. *Given an instance of NAE3SAT with d variables and k clauses, let the set of vectors V be given by $K_1 = 1600d^7 \ln^2(d)$ copies of basic vectors (Z and Y vectors), together $K_2 = d^2 \ln(d)$ copies of the clause vectors for each clause. For sufficiently large d, there is a function $p(n,k)$ such that, if there is at least one solution to the NAE3SAT, $F = \int_x I_V(\vec{x}) \geq pd^{-22d}$, and if there are no solutions, $F \leq pd^{-d^2}$.*

Proof. The theorem will hold if we take p as the value of I_V at a good point, or

$$p = d^{-K_1 d^2}\left(\frac{32}{27d^3}\right)^{K_2}.$$

If the original NAE3SAT instance has a satisfying assignment $(1, 0, 0, 1, \dots)$, there is a corresponding good point

$$\vec{g} = \frac{1}{\sqrt{d}} \left(+\vec{e}_1 - \vec{e}_2 - \vec{e}_3 + \vec{e}_4 \dots\right)$$

with a large value of $I_V(\vec{g})$. Each set of basic vectors introduces a factor of $1/d^{d^2}$ in I, and each set of clause vectors introduces a factor of $32/27d^3$. Thus

$$I_V(\vec{g}) = d^{-K_1 d^2} \left(\frac{32}{27d^3}\right)^{K_2} = p$$

Further, we want to show that around this good point $\vec{g}$, there is an appreciable volume with large I_V, that will contribute substantially to F. Around each good point, take the ball of radius

$$\epsilon_g = \frac{1}{3200d^9(1+d)}.$$

Then by Lemma 5, each set of basic observations gives a factor in I of at least

$$I_1 \geq \frac{1 - 2\epsilon_g d^{5/2}}{d^{d^2}}$$

and by Lemma 6, each set of clause observations gives a factor at least

$$I_2 \geq \frac{32}{27d^3}(1 - 12\epsilon_g\sqrt{d})$$

so that the final I_V value of each point in the ball is at least

$$I_0 = I_1^{K_1} I_2^{K_2} \geq p(1 - 2\epsilon_g d^{5/2})^{K_1}(1 - 12\epsilon_g\sqrt{d})^{K_2} \tag{3.10}$$

$$\geq p(1 - 2\epsilon_g K_1 d^{5/2})(1 - 12K_2\epsilon_g\sqrt{d}) \tag{3.11}$$

$$= p\left(1 - 2\frac{1}{3200d^9(1+d)}(1600d^7\ln^2(d))d^{5/2}\right)\left(1 - 12(d^2\ln(d))\frac{1}{3200d^9(1+d)}\sqrt{d}\right) \tag{3.12}$$

$$\geq p\left(1 - \frac{\ln^2 d}{\sqrt{d}}\right) \tag{3.13}$$

This means the total contributed to F by the ball around this good point is then at least $p(1 - \ln d/\sqrt{d})$ times the volume of this ball around $\vec{g}$. The ball is not actually a sphere in $\mathbb{R}^{2d}$, as it lies on the manifold of normalized states, which is curved; it's the intersection of a ball centered at $\vec{g}$ and the unit sphere. But since $\epsilon_g < 1/2$, this deformation reduces the volume by less than a factor of $1/2$, and then we can use the standard volume of the ball. So the volume obeys

$$\text{Vol} \geq \frac{1}{2} \cdot \frac{2(d-1)!(4\pi)^{(d-1)}}{(2d-1)!}\epsilon_g^{2d-1}$$

and a single good point contributes a total likelihood to p_{norm} at least

$$\text{Vol} \cdot I_0 \geq p c_1 c_2^{-d} d^9 d^{-21d}$$

for some particular constants $c_1, c_2 > 1$; the d^{-21d} term clearly dominates the scale for large d. For sufficiently large d then we can write

$$F \geq \text{Vol} \cdot I_0 \geq p d^{-22d}$$

which establishes the first claim. The second claim concerns when there are no good points. Suppose for contradiction that there is some point $\vec{x}$ (not necessarily in B_0) so that $I_V(\vec{x}) > p/d^{d^2}$. Applying Lemma 4, we know that it must have $\epsilon_\alpha = |\vec{\alpha} - \vec{1}| < 0.1/d^2$, otherwise it would have at most

$$I_V(\vec{x}) \leq \left(d^{-d^2}(1 - 0.1^2/4d^5)\right)^{K_1} < d^{-K_1 d^2} \exp(-K_1/400d^5) \tag{3.14}$$

$$= d^{-K_1 d^2} \exp(-4d^2 \ln^2 d) \tag{3.15}$$

$$< d^{-K_1 d^2} \exp\left(-4d^2 \ln^2 d + d^2 \ln d \ln(32/27)\right) \tag{3.16}$$

$$= d^{-K_1 d^2} \exp\left(-d^2 \ln^2 d + d^2 \ln d \ln(32/27d^3)\right) \tag{3.17}$$

$$= d^{-K_1 d^2} \exp\left(-d^2 \ln^2 d + \ln\left(\left(\frac{32}{27d^3}\right)^{K_2}\right)\right) \tag{3.18}$$

$$= d^{-K_1 d^2} \left(\frac{32}{27d^3}\right)^{K_2} / d^{d^2 \ln d} = p/d^{d^2 \ln d} \tag{3.19}$$

$$< p/d^{d^2} \tag{3.20}$$

Since $\epsilon_\alpha \leq 1/2$, we can also apply the second part of Lemma 4 and check that the all phases $|\theta_i| < 0.1/d$, otherwise our point would have I_V at most

$$\left(d^{-d^2}(1-3\theta_i^2)\right)^{K_1} < \left(d^{-d^2}(1-0.03/d^2)\right)^{K_1} \tag{3.21}$$

$$< \left(d^{-d^2}(1-0.1^2/4d^5)\right)^{K_1} \tag{3.22}$$

$$< p/d^{d^2} \tag{3.23}$$

Since the amplitudes are all within ϵ_a of $1/\sqrt{d}$, and the phases are all within $0.1/d$ of 0, the point's distance to the nearest point b in B_0 is at most

$$\begin{aligned}\text{dist}_{B_0} &\leq \sqrt{d}\left(\epsilon_a + \left(\frac{1}{\sqrt{d}} + \epsilon_a\right)\left((1-\cos(\theta_i))^2 + \sin^2(\theta_i)\right)\right) \\ &\leq \sqrt{d}\left(\frac{0.1}{d^2} + \left(\frac{1}{\sqrt{d}} + \frac{0.1}{d^2}\right)(2 - 2\cos(0.1/d))\right) \leq \sqrt{d}\left(\frac{0.1}{d^2} + \frac{2}{\sqrt{d}}(0.1/d)^2\right) \\ &\leq \frac{0.11}{d^{3/2}}\end{aligned}$$

If that point b is bad, then by Lemma 6 our point would have I_V at most

$$\begin{aligned}d^{-K_1 d^2}\left(\frac{4096}{27}\left(\frac{0.11}{d^{3/2}}\right)^6\right)^{K_2} &= d^{-K_1 d^2}\left(\frac{0.00023}{d^9}\right)^{K_2} = d^{-K_1 d^2}\left(\frac{32}{27d^3}\right)^{K_2}\left(\frac{0.00023}{(32/27)d^6}\right)^{K_2} \\ &< p \times 0.0002^{K_2} \leq p/5000^{d^2 \ln d} = p/(5000d)^{d^2} < p/d^{d^2}.\end{aligned}$$

We've shown that all points have $I_V \leq p/d^{d^2}$. The volume of integration is $S_{2n-1} < 1$, so the total integral F is less than p/d^{d^2}. □ □

3.5.4 NP Hardness

We can now prove our main result.

Theorem 3.5. *For any constant $C < 1$, it is NP-Hard to approximate the permanent of an $n \times n$ Hermitian positive semidefinite matrix within a factor of 2^{n^C}.*

Proof. We can reduce from NAE3SAT. Given an NAE3SAT instance on d variables, we can use the set of vectors V described in Theorem 3.4 and examine the resulting value F. As we have $O(d^9)$ vectors in V, the quantity F can be represented as a permanent of a matrix of size $O(d^9)$. The NAE3SAT instance is satisfiable if $F \geq pd^{-22d}$ and unsatisfiable if $F \leq pd^{-d^2}$, which can be distinguished if approximating within a factor of $d^{d^2-22d} = O(d^{d^2})$, and so $O(2^{d^2})$ will suffice. If we had an oracle that could approximate permanents of size n PSD matrices within a factor of 2^{n^C} for some $C < 1$, then we could do the replica trick: take the matrix corresponding to F, and repeat it $M = d^{(2-9C)/(1-C)}$ many times along the diagonal. The result is a matrix of size Md^9, which is then approximated within a factor of $2^{(Md^9)^C}$. The resulting matrix size Md^9 is still $poly(d)$ for any fixed C. Then we raise this approximate answer to the power $1/M$ to recover an approximation to the original permanent, and it has multiplicative error

$$\left(2^{(Md^9)^C}\right)^{1/M} = 2^{d^{10C}M^{C-1}} = 2^{d^{9C}d^{2-9C}} = 2^{d^2}$$

which is sufficient to distinguish between satisfiable and unsatisfiable instances. As NAE3SAT is NP hard, so is approximating HPSD permanents with this accuracy.

□

This result is complementary to one of Anari et al[8], where they show that one *can* approximate within a factor of $\exp((1+\gamma+o(1))n)$ where γ is the Euler-Mascheroni constant, while we showed that permanents cannot be approximated with subexponential error. Our hard instances circumvent the fast approximation schemes of [16] and [34], which both have requirements on the spectrum of the matrix, and perform more favorably when $\lambda_{max}/\lambda_{min}$ is smaller. Our instances are of low rank (only rank d, which is much larger than the matrix size n) so that $\lambda_{min} = 0$.

Finally, we conjecture that the reduction above is approximation preserving: that each good point contributes an equal amount of likelihood that can easily be estimated beforehand. Showing this would require tighter error bounds.

Conjecture 3.1. *With an appropriate choice of polynomial-scaling K_1 and K_2, the construction used in Theorem 3.4 is an approximation-preserving reduction from #NAE3SAT to HPSD permanents, such that approximating HPSD Permanents within a factor C is as hard as approximating #NAE3SAT (or #3SAT) within a factor C.*

It is known that by Stockmeyer counting[57, 104, 114] computing multiplicative approximations to PSD permanents is contained in $\mathsf{FBPP}^{\mathsf{NP}}$, and if it is indeed as hard as approximating #3SAT, it seems unlikely to be much easier than this.

3.5.5 Real Matrices

The arguments above all involve complex vectors, complex matrices, and integrals over the complex unit sphere. The arguments however can easily be adapted to show that PSD permanents remain hard even for purely real matrices. We could have proved the results only for the real case and this would of course imply hardness for the more general complex case, but the proof for the real case was less symmetric, asthetic, or inuitive than the complex case, which is why we delayed to this section.

Theorem 3.6. *For any constant $C < 1$, it is NP-Hard to approximate the permanent of an $n \times n$ real positive semidefinite matrix within a factor of 2^{n^C}.*

Proof. The construction proceeds very similarly to above, by reducing from NAE3SAT. However, we now use one dimension more in the space: a d-variable NAE3SAT problem is mapped to a $(d+1)$-dimensional spherical integral $\int I(\vec{x})$. The clauses are mapped, as before, with K_2 many sets of clause vectors, connecting the variables 1 through d in the original problem with dimensions 1 through d in the spherical integral $I(x)$. The "basic sets" still include K_1 many instances of the

unit vectors $\vec{e}_k$ in each basis direction $k \in [d+1]$, what we previously referred to as the Z vectors.

The Y vectors were, in preivous proofs, of the form $\frac{\vec{e}_j \pm i\vec{e}_k}{\sqrt{2}}$, for $j \neq k$. This was the sole source of complex terms in our vectors, and the reasons the resulting matrices were complex. Instead now we use four copies of each of $\frac{\vec{e}_j \pm \vec{e}_{d+1}}{\sqrt{2}}$. These each softly enforce the constraint that the component of $\vec{x}$ in the j direction and the $d+1$ direction have relative phase $\pm i$ (that is, $\pm\sqrt{-1}$). Since each j has $\pm i$ relative to $d+1$, this implies that each $j \neq k$ have relative phase ± 1.

To make this quantitative and precise, we refer to the proof of Lemma 4. The bound of $1 - \frac{\epsilon_\alpha^2}{4d}$ applies as before, since the $\vec{e}_k$ vectors occur just as before. As proved in Lemma 4, if θ_j and θ_{d+1} differ by a phase (up to ± 1) of $\Delta\theta_j = \theta_j - \theta_{d+1}$, then the likelihood $I(x)$ is reduced by a factor of $1 - 3\Delta\theta_j^2$; since we use each vector eight times, this becomes $(1 - 3\Delta\theta_j^2)^8$. Then for two $j \neq k$, $j, k \leq d$, the likelihood is at most

$$(1-3\Delta\theta_j^2)^4(1-3\Delta\theta_k^2)^4 \leq \left(1 - 3\left(\frac{|\Delta\theta_j| + |\Delta\theta_k|}{2}\right)^2\right)^8 \leq \left(1 - 3\left(\frac{|\theta_j - \theta_k|}{2}\right)^2\right)^8 \leq 1-3(\theta_j-\theta_k)^2$$

which gives us the same bound on the relative phases as before, so that an analogous statement to Lemma 4 for our new basis set. The proof of Lemma 5 holds with few modifications: in the proof above, the Y terms

$$\prod_{j \leq k} \alpha_j^2 + \alpha_k^2 + 2\alpha_j\alpha_k \cos(2\pi(\theta_j - \theta_k))$$

lead to the a penalty

$$\prod_{j\leq k}(1-\epsilon\sqrt{d})^2+(1-\epsilon\sqrt{d})^2+2(1-\epsilon\sqrt{d})^2\left(1-\frac{(2\pi\epsilon)^2}{2}\right)=\left((1-\epsilon\sqrt{d})\left(4-4\pi^2\epsilon^2\right)\right)^{(d^2-d)/2}$$

$$\geq 1-(\epsilon\sqrt{d}+\pi^2\epsilon^2)\frac{d^2-d}{2}.$$

Here instead we have four copies of each phase constraints, but only between $j \leq d$ and $d+1$. So the penalty from

$$\prod_{j\leq d}\left(\alpha_j^2+\alpha_{d+1}^2+2\alpha_j\alpha_{d+1}\cos(2\pi(\theta_j-\theta_{d+1}))\right)^4$$

becomes

$$\prod_{j\leq d}\left((1-\epsilon\sqrt{d})^2+(1-\epsilon\sqrt{d})^2+2(1-\epsilon\sqrt{d})^2\left(1-\frac{(2\pi\epsilon)^2}{2}\right)\right)^4=\left((1-\epsilon\sqrt{d})\left(4-4\pi^2\epsilon^2\right)\right)^{4d}$$

$$\geq 1-(\epsilon\sqrt{d}+\pi^2\epsilon^2)(4d)\geq 1-(\epsilon\sqrt{d}+\pi^2\epsilon^2)\frac{d^2-d}{2}$$

as before, as long as $d \geq 9$. The resulting conclusion of the lemma that the relative value $I_V(\vec{x})/I_V(B_0) \geq 1-2\epsilon d^{5/2}$ thus still holds.

Finally, Lemma 6 remains umodified in this setting, as the form of the clause vectors is unchanged. As all the necessary lemmas hold as before, and the proofs of Theorems 3.4 and 3.5 only care about relative values, they will all hold in the real-valued PSD setting. □ □

3.6 Quantum State Tomography

The author initially found the above construction while investigating the worst-case hardness of quantum state tomography, and the hardness implies that several problems in the context of tomography are NP-hard as well.

Quantum State Tomography (QST) is the procedure of estimating an unknown quantum state from a set of measurements on an identically prepared ensemble. The procedure can encompass both the choosing of measurement bases as well as estimating the resulting state from the measurements; in adaptive settings, the running estimate is also used to inform future measurement choices[68, 103]. We focus on the latter task, of building an estimate of the state. We look at four related forms of what "estimation" can qualify as:

1. Finding the Maximum Likelihood Estimator (MLE): the pure state ρ most likely to produce the observations.

2. Finding the Bayesian expected state ρ_{Avg}: assuming a prior over the possible pure states, finding the mixed state presenting the mixture of appropriately weighted possible states.

3. Computing the expectation value of some future observation(s).

4. Finding the probability that the unknown state is in fact some particular ρ_0. (As there are infinitely many different pure states, we are actually asking for the probability *density* at ρ_0.)

The first three estimations problems have all been extensively studied with various heuristics. MLE can be attempted by linear inversion[102, 40], iterative search[92, 69], or even neural networks[117]. Bayesian estimation can be accomplished by direct numerical integration[22] or particle based sampling[68], possibly with neural networks guiding the particles[103]. Directly estimating future samples has also been attempted with neural networks[112] or classical shadows[1, 2, 66]. The author is not aware of any prior work on computing estimation problem 4.

We can show that estimation problems 2, 3, and 4 are essentially as hard as approximating PSD permanents, and that task 1 is also NP-hard. The exponential difficulty (assuming ETH[70]) is in fact in the dimension d of the underlying Hilbert space. Many questions in quantum information appear to be "exponentially" hard, in the sense that it is hard to analyze a system q qubits faster than $O(2^q)$. But here $d = 2^q$, so that even when the number of qubits is a logarithmically small $q = \log(d)$, the problem of state estimation remains exponentially hard.

3.6.1 Outline of Tomography Results

Of the four forms above, we focus first on estimation problem 4. Although it is likely the question least relevant to experiment, it is the easiest to manipulate algebraically. We call it Quantum-Bayesian-Update, or simply QBU, and define it in Section 3.6.2. In Section 3.6.3, we give an exponential time algorithm for QBU, showing that it is at least possible. In Section 3.6.4 we show that estmation problems 2, 3, and 4 are equivalent. In Section 3.6.5 we explain QBU's connection to HPSD permanents, and show it is NP-Hard to approximate within subexponential error. In Section 3.6.6 we show how the construction of difficult PSD permanents can also be modified shows that the MLE problem (estimation problem 1) is also NP-hard to approximate: it is NP-hard to check the existence of a state with likelihood within a subexponential factor.

3.6.2 Quantum Bayesian Update

We define the QBU problem as follows: given a series of observations $\mathcal{O}_i$ each taken from a copy of ρ, and a guess ρ_0, what is the probability density that $\rho = \rho_0$? The actual probability of equality is zero – unless we have some other powerful information about the state – which is why we ask for the probability density in the space of candidate density matrices.

Bayes' theorem lets us compute the probability density of a true state ρ in terms of the likelihood of the observations $P(\mathcal{O}|\rho)$, a prior belief distribution $P(\rho)$, and the total probability of the sequence of observations $P(\mathcal{O})$. It reads,

$$P(\rho_0|\mathcal{O}) = \frac{P(\mathcal{O}|\rho_0)P(\rho_0)}{P(\mathcal{O})}$$

In order for the equation to be meaningful and not identically zero on both sides, we can read ρ as representing a small volume in the space of density matrices. While there are many natural priors on the space of density matrices, we focus on the case where we know the unknown state ρ is pure. This models, for instance, where we are trying to identify the output of a unitary quantum channel. The most natural prior is then the uniform distribution over all pure states, given by the Haar measure. Then all $P(E)$ are equal. The likelihood of a given observation $\mathcal{O}_i$ is simply $\mathrm{Tr}[\mathcal{O}_i\rho]$, so our goal is to compute

$$P(\rho|\mathcal{O}) = \frac{\prod_{i\in[n]} \mathrm{Tr}[\mathcal{O}_i\rho]}{P(\mathcal{O})}$$

In general $\mathcal{O}_i$ could be operators of any rank, and could belong to POVMs. For hardness, it will suffice it consider only observations with rank 1 and trace 1, but for now we allow them to be general. For any particular ρ and sequence $\mathcal{O}_i$, the likelihood $\prod \mathrm{Tr}[\mathcal{O}_i\rho]$ can be evaluated directly in $O(nd^2)$ operations. The difficulty then lies in the normalizing factor,

$$p_{norm} = P(\mathcal{O})$$

so that

$$P(\rho|\mathcal{O}) = p_{norm}^{-1} \prod_{i \in [n]} \mathrm{Tr}[O_i \rho]$$

This indicates the probability of an entire sequence of observations. While a single observation has the simple form of $P(\mathcal{O}_i) = \mathrm{Tr}[\mathcal{O}_i]$, the expression rapidly becomes more complicated as we consider sequences of observations.

A brief example is useful for understanding what p_{norm} represents. Suppose that we measure a qubit 1000 times along each of the X, Y, and Z axes: we expect to see a particular amount of bias. Observing 1000 results each of +X, +Y, and +Z would be very unlikely, as the qubit cannot be in the +1 eigenstate of all three axes at once. It would be similarly surprising to see exactly 500 counts each of +X, -X, +Y, -Y, +Z, and -Z: this state shows no tendency of a particular orientation, but a pure qubit state must show a bias towards some orientation. This would have a small value of p_{norm}, as there is no good state to explain the sequence observed. A sequence of 1000 +Z observations, and 500 each of +X, -X, +Y, and -Y is much more likely, as it can be well explained by the $|\uparrow\rangle$ state, and so has a larger value of p_{norm}.

As we just saw, computing $P_{density}(\rho = \rho_0|\mathcal{O})$ is easy if p_{norm} is known, and conversely p_{norm} can be easily computed from the probability density. p_{norm} is

a more attractive goal for our problem, as it doesn't depend on ρ_0. It can be computed by summing up all unnormalized probabilities:

$$p_{norm} = \int_{\vec{x}\in\mathbb{C}_1^d} \prod_{i\in[n]} \mathrm{Tr}[\mathcal{O}_i x x^\dagger]\, dx$$

where the integral is over the Hilbert space $\mathbb{C}^d$ restricted to length-1 vectors. This leads to the definition,

Definition 3.1 (Quantum-Bayesian-Update)**.** *Given a collection of observations $\mathcal{O} = (\mathcal{O}_1, \ldots \mathcal{O}_n)$ in a Hilbert space of dimension d, compute*

$$p_{norm} = \frac{\int_{\vec{x}\in\mathbb{C}_1^d} \prod_{i\in[n]} \mathrm{Tr}[\mathcal{O}_i x x^\dagger]\, dx}{\mathrm{Tr}[\mathcal{O}_i]} \tag{3.24}$$

3.6.3 Polynomial time QBU for fixed d

This space of state vectors $\mathbb{C}_1^d$ has the geometry of a real $(2d-1)$-sphere, and the entries of ρ are quadratic in the Cartesian coordinates for this sphere. Thus, p_{norm} becomes a integral over a $(2d-1)$-sphere of a homogeneous $2n$ degree polynomial in the $2d$ variables. The expansion of the polynomial into monomials takes $O((2n)^{2d})$ time, and each monomial can then be immediately integrated over the sphere using the formula[53]

$$\int_{S^k} x_1^{\alpha_1} x_1^{\alpha_2} \ldots x_k^{\alpha_k} = \begin{cases} 0 & \text{if any } \alpha_i \text{ are odd} \\ \frac{2\prod_i \Gamma(\frac{1}{2}(\alpha_i+1))}{\Gamma(\sum_i \frac{1}{2}(\alpha_i+1))} & \text{if all } \alpha_i \text{ are even} \end{cases} \tag{3.25}$$

where Γ is gamma function, $\Gamma(\frac{1}{2}(\alpha+1)) = \sqrt{\pi 2^{\alpha}}(\alpha-1)!!$. This gives a polynomial time algorithm for evaluating p_{norm} when d is fixed.

3.6.4 Relationship between estimation problems

Since QBU is not of particular interest to actual tomography tasks, we show it is equivalent (under polynomial many-one reductions) to the more realistic tasks 2 and 3 above, of estimating observables or the state itself. We can show that these are just as difficult (or, just as easy) as the Bayesian update step.

Computing ρ_{Avg}

Given that there will always be room for uncertainty, we cannot meaningfully ask for a single pure state as an answer, but we can ask for ρ_{Avg}: the mixed state representing the correctly updated mixture over all the possible true states, given by $\int P(\rho)\rho\, d\rho$. The impure ρ_{Avg} reflects the expectation of all observables given our current information.

We parameterize the space of density matrices by a single vector $\psi \in S^{2d-1}$, and given some completed observations $\mathcal{O}$, the Bayesian expected state is

$$\rho_{Avg} = \int_{\psi \in S^{2d-1}} P\Big(|\psi\rangle\langle\psi| \,\Big|\, \mathcal{O}\Big)\, |\psi\rangle\langle\psi|\, d\psi$$

$$= \int_{\psi \in S^{2d-1}} p_{norm}^{-1}\Big(|\psi\rangle\langle\psi| \Big) \prod_{O \in \mathcal{O}} \langle\psi|O|\psi\rangle\, d\psi$$

whose individual matrix elements are

$$\langle i|\rho_{Avg}|j\rangle = p_{norm}^{-1} \int_{\psi \in S^{2d-1}} \langle i|\psi\rangle \langle \psi|j\rangle \prod_{O\in\mathcal{O}} \langle \psi|O|\psi\rangle \; d\psi$$

We have already discussed computing p_{norm}, as a spherical integral of a polynomial. For any given i and j, the remaining integral is also a spherical integral of a polynomial, and can be computed in the same fashion. In fact we can re-use the results from the large product excluding the i and j, and so ρ_{Avg} can be recovered in $O(n^d)$ time.

On the other hand, a diagonal element $\langle i|\rho_{avg}|i\rangle$ gives

$$\langle i|\psi\rangle \langle \psi|i\rangle \prod_{O\in\mathcal{O}} \langle \psi|O|\psi\rangle = \langle \psi| \Big(|i\rangle \langle i| \Big) |\psi\rangle \prod_{O\in\mathcal{O}} \langle \psi|O|\psi\rangle = \prod_{O\in(\mathcal{O}\cup\{|i\rangle\langle i|\})} \langle \psi|O|\psi\rangle$$

which is the same integrand as for p_{norm}, only with one additional observation $|i\rangle \langle i|$ added.

If we had an algorithm compute ρ_{Avg} efficiently, we could use it to solve the Bayesian update problem on a set of observations $\mathcal{O}$, by discarding the last observation O_{last}, computing ρ_{Avg}, decompose O_{last} into a scaled sum of projectors $\sum_i \lambda_i |i\rangle \langle i|$, and then evaluate the sum of matrix elements $\sum_i \lambda_i \langle i| \rho_{Avg} |i\rangle$. This shows that state estimation is at least as hard as Bayesian updating.

Computing observable expectations

We could try to only find the expectation of a particular observable A, and not the whole state ρ_{Avg}, conditioned on our observations. We can write this as $E[A|\mathcal{O}]$. This is also just as hard: density matrices as a d^2-1 linear space, and expectations of observables are linear in ρ, so by computing the exact expectation of d^2-1 independent obsevables, we can find ρ_{Avg} exactly. This is of course precisely the idea behind least-squares quantum state estimation, and it shows that computing expectation values is as hard as ρ_{Avg}.

Finally, if we could compute a Bayesian update, we could compute the expectation values of observables. Just as before, write our desired obsevable as $A = \sum \lambda_i |i\rangle \langle i|$, and evaluate

$$E[A|\mathcal{O}] = \sum \lambda_i E[|i\rangle \langle i| \, |\mathcal{O}] = \sum \lambda_i p_{norm}^{-1} \int_{\psi \in S^{2d-1}} \prod_{O \in (\mathcal{O} \cup \{|i\rangle\langle i|\})} \langle \psi | O | \psi \rangle \; d\psi$$

Computing p_{norm} and each of the d many spherical integrals is a Bayesian update problem. We have reductions (Bayesian update) $\rightarrow$ (Compute ρ_{Avg}) $\rightarrow$ (Compute $E[A|\mathcal{O}]$) $\rightarrow$ (Bayesian update), so these are equivalent in difficulty. Note that these are many-one reductions, which is unavoidable as ρ_{Avg} is a matrix-valued function problem while the two are scalar-valued.

3.6.5 NP-Hardness of QBU and ρ_{Avg}

We now state the main hardness results on quantum tomography.

Theorem 3.7. *For any $C < 1$, it is NP-hard to compute the value p_{norm} for Quantum-Bayesian-Update with an approximation factor of at most 2^{n^C}.*

Proof. When $\mathcal{O}_i$ are all rank-1 operators, the numerator in Eq. 3.24 is of the form in Theorem 3.3, and the denominator in Eq. 3.24 can be efficiently computed by direct calculation. Thus any PSD permanent can be efficiently reduced to a problem of computing p_{norm} with an approximation-preserving reduction, and QBU is NP-Hard to approximate to the same degree. □

Theorem 3.8. *For any $C < 1$, it is NP-hard to compute a diagonal matrix entry of ρ_{Avg}, in any basis, with an approximation factor of at most 2^{n^C}. It is also NP-hard to compute the expectation of a positive semidefinite operator $\mathcal{O}$ with an approximation factor of at most 2^{n^C}.*

Proof. A diagonal element of ρ_{Avg} is the expectation value of the rank-1 PSD operator projecting onto that element, so the first statement is a special case of the second. As described above, both of these quantities then also take the form of a PSD permanent, and any PSD permanent can be turned into these problem by taking the desired matrix element (in the first case) or observabe $\mathcal{O}$ (in the

second case) to be the first vector $V_1^\dagger V_1$. These are also approximation preserving reductions, so these are also NP-hard to approximate. □

3.6.6 NP-Completeness of Maximum Likelihood Estimation

In the case of MLE state tomography, we are not so demanding that we require knowledge of the full average state, and we are content with just finding one good explanatory state $|\psi\rangle$. Accordingly, we do not consider a permanent $\int_x I_V(x)$ (a problem of counting solutions to 3-SAT), but just the question of maximizing $I_V(x)$ (a problem of finding a solution to 3-SAT). This allows to show that the problem is actually lies in NP, while this is unlikely to be true for the other problems described above, unless $\mathsf{BPP}^{\mathsf{NP}} = \mathsf{NP}$ (widely presumed to be false).

Formulating the MLE problem as a decision problem:

Definition 3.2 (C-Approximate-Quantum-MLE)**.** *Given a collection of observations $\mathcal{O}_i$ of an unknown quantum state $|\psi\rangle$, and a real number p, decide whether there is a $|\psi\rangle$ whose likelihood $L(\psi) = \prod_i \langle\psi|\mathcal{O}_i|\psi\rangle$ is at least p, or if $L(\psi) < p/C$ for all ψ, being promised that one of these is the case.*

We will show that even the approximate problem is NP-hard, for any C.

Theorem 3.9. *For any $C > 1$, the C-Approximate-Quantum-MLE problem is NP-complete.*

Proof. Containment in NP is straightforward, as one can supply a description of the state $|\psi\rangle$, which requires only d many real numbers, and then $L(\psi)$ can be directly evaluated.

To show hardness, we use the same NAE3SAT construction as in Theorem 3.4. As was shown in the proof of that theorem, any good point (thus, a solution to the underlying NAE3SAT problem) has

$$L(\psi) = I_0(x) \geq p\left(1 - \frac{\ln^2 d}{\sqrt{d}}\right).$$

We also show in that proof that, if there are no good points (and thus no solutions) then

$$L(\psi) = I_0(x) \leq p/d^{d^2}$$

for all points. Thus, the existence of a high likelihood point even within $C < d^{d^2}$ implies the existence of a solution. □

3.6.7 Practical Difficulty of Tomography

Although the above results imply that several approaches to quantum state tomography may be difficult to compute exactly, these difficult instances are somewhat artificial and unlikely to occur in practice. Additionally, difficult instances

such as the one constructed in the above proofs could be readily addressed in practice by the addition of measurements in e.g. the X measurement basis, which would directly probe the relative signs in the state vector and allow relatively efficient readout of the state. Additionally, the constraint that we only search for pure states – while a useful prior that could be relevant once high-fidelity quantum computer exists – makes a highly nonconvex search space. If we relax this and take a prior with uniform measure over the space of density matrices, then the resulting likelihood function is logarithmically convex and the resulting MLE problem can be solved in polynomial time in d. Thus, these results should not be taken as a statement that quantum state tomography is actually exponentially hard in the Hilbert space dimension d. Rather, any analysis of quantum state tomography procedures will need at least one of: careful choice of measurement basis, only probabilistic guarantees on convergence, or (if doing MLE) a convex prior.

Chapter 4

Linear-time generalized Hartree-Fock algorithm for quasi-one-dimensional systems [1]

[1] Based on the work in 10.1103/PhysRevResearch.4.023128

4.1 Introduction

While the quantum many-body problem for fermions can in general not be solved numerically efficiently, a plethora of approximate computational approaches have been developed that are able to capture relevant properties of many-fermion systems in certain limits. A widely used class of such methods are Hartree-Fock and self-consistent Bardeen-Cooper-Shrieffer (BCS) theory [12, 13]. These can be viewed as variational mean-field methods: they find the best approximation to the exact ground state within the space of non-interacting fermion states. Another powerful approach is density-functional theory, which expresses the total energy of the fermionic system as a – generally unknown – functional of the electron density [64]; while finding this functional is difficult, decades of numerical experience have shown that relatively simple approximations to this functional, such as the local density approximation, can successfully describe materials where the effect of interactions is moderate [73].

Common to these approaches is that the most computationally expensive step in the numerical simulation is finding the ground state of a system of fermions coupled to their self-consistent potential (the mean-field potential in the case of HF and BCS [62], and the Kohn-Sham potential in the case of DFT [83]). Without further approximation, this step scales cubically in the number of degrees

of freedom and as such becomes prohibitively costly for systems in excess of a few thousand degrees of freedom.

Here, we will demonstrate that in low spatial dimensions and for local Hamiltonians, this step can be accelerated significantly by using tensor network states (TNS) [125, 128, 51, 101, 111, 96, 97, 98, 120, 74, 121, 122] (for recent reviews, see Refs. [100, 99, 31]). Such states are known to be able to compactly represent weakly entangled quantum many-body states, such as the ground states of local Hamiltonians. In many cases, the computational scaling of these approaches is empirically found to be approximately linear in the size of the system and exponential in its bipartite entanglement. However, in the case of free fermions, this can be reduced further to a polynomial scaling in the entanglement by using so-called Gaussian fermionic tensor network states [87, 49, 110, 45, 59, 52, 50, 60, 71]. Recently, a particular variant, Gaussian fermionic matrix-product states (GFMPS) [110], was used as basis for efficient computational methods for non-interacting fermions in quasi-one-dimensional systems. These methods are able to compute equilibrium and non-equilibrium properties for systems order of magnitudes larger than naive approaches [108].

In this chapter, we focus on accelerating the self-consistent generalized Hartree-Fock (gHF) iteration [11, 23, 43, 41] using Gaussian fermionic matrix-product states. We begin by reviewing the gHF approach, which should be viewed as the

most general variational method using states of non-interacting fermions and elegantly unifies Hartree-Fock and self-consistent BCS theory. We then rederive the self-consistency equations for gHF, review key properties of GFMPS and discuss how to efficiently implement the gHF iteration using GFMPS. Finally, we demonstrate the approach on an example of interacting fermions in a quasi-dimensional geometry in the presence of an inhomogeneous trapping potential.

We note that other numerical approaches to solve the gHF problem have been discussed in Ref. [85, 86]. In particular, Ref. [85] discusses how to perform real- and imaginary-time evolution in the gHF setting, and pursues imaginary-time evolution as an approach to find the ground states. We focus here instead on the self-consistent field approach, which is often faster but may be more prone to becoming trapped in local minima. We note that the time evolution described in Ref. [85] could similarly be accelerated using GFMPS techniques [108]. A different approach to improve the performance of gHF based on highly scalable methods for solving the fermionic problem for sparse systems was discussed in Ref. [90], reaching remarkably large systems by parallelizing the computation on several thousand computational cores.

4.2 Methods

4.2.1 Gaussian fermionic states

We consider a lattice of fermions, where the operators $a_i^\dagger$ and a_i create and annihilate a fermion on the i'th site, respectively. For our purposes, it will be convenient to introduce a basis of self-adjoint Majorana fermion operators, which we denote c_i, and which satisfy the commutation relations $\{c_i, c_j\} = 2\delta_{ij}$. They are related to the standard creation and annihilation operators by $\hat{c}_{2i-1} = \hat{a}_i + \hat{a}_i^\dagger$, $\hat{c}_{2i} = -i(\hat{a}_i - \hat{a}_i^\dagger)$. In this way, any system with N fermionic modes can be rewritten as one with $2N$ Majorana fermions.

For any density operator $\hat{\rho}$ of a system of fermions, we can associate a real anti-symmetric so-called covariance matrix Γ:

$$\Gamma_{ij} = \frac{i}{2} \operatorname{Tr}(\hat{\rho}[\hat{c}_i, \hat{c}_j]). \tag{4.1}$$

For pure states, which we focus on here, Γ has to satisfy $\Gamma^2 = -\mathbb{1}$. Gaussian states are furthermore characterized by the property that this covariance matrix contains a full description of the density operator, i.e. the expectation value of *any* operator can be computed from it [28] (for a more formal definition of Gaussian states, see Ref. [26]). Specifically, the expectation value of a Majorana monomial

$\prod_{x\in\mathcal{X}} \hat{c}_x$, where $\mathcal{X}$ denotes some set of lattice sites, is given by

$$\mathrm{Tr}\left(\hat{\rho}\prod_{x\in\mathcal{X}}\hat{c}_x\right) = \mathrm{Pf}(\Gamma_{\mathcal{X}}), \tag{4.2}$$

where $\mathrm{Pf}(\cdot)$ denotes the Pfaffian and $\Gamma_{\mathcal{X}}$ the covariance matrix restricted to the sites in $\mathcal{X}$. This should be viewed as embodiment of Wick's theorem in the covariance matrix formalism.

For Hamiltonians quadratic in creation and annihilation operators (alternately, quadratic in Majorana operators), i.e. of the form

$$\hat{\mathcal{H}} = -i\sum_{i,\,j} H_{ij}\hat{c}_i\hat{c}_j \tag{4.3}$$

where H is real and anti-symmetric, the ground state is always Gaussian. H can be diagonalized and has purely imaginary eigenvalues. The minimum energy state is determined by the covariance matrix

$$\Gamma = i(V_-V_-^\dagger - V_+V_+^\dagger) \tag{4.4}$$

where V_+ (V_-) are the normalized eigenvectors of H corresponding to eigenvalues with positive (negative) imaginary parts. In this way, Γ and the ground state can be determined through a diagonalization of H. Such diagonalization takes $\mathcal{O}(N^3)$ time.

Every pure Gaussian fermionic has well-defined fermion parity, i.e. any pure Gaussian state $|\gamma\rangle$ satisfies $\prod \hat{c}_i\,|\gamma\rangle = e^{i\pi\sum\hat{a}_i^\dagger\hat{a}_i}\,|\gamma\rangle = p\,|\gamma\rangle$ with $p = \pm 1$. (See

Appendix B.2.) In terms of the covariance matrix γ corresponding to this state, the parity is given by $\mathrm{Pf}(\gamma)$, where Pf denotes the Pfaffian. An important subset of states within the class of Gaussian fermionic states are those with a well-defined particle number, i.e. that satisfy $\sum \hat{a}_i^\dagger \hat{a}_i \left|\gamma\right\rangle = n \left|\gamma\right\rangle$ for some integer n. These are traditionally referred to as Slater determinants and can be written as $\prod_{i=1}^{n} \hat{d}_i^\dagger \left|0\right\rangle$, where the $\hat{d}_i^\dagger$ are a new set of n fermionic creation operators that are related to the original $\hat{a}_i^\dagger$ by a unitary transformation, and $\left|0\right\rangle$ is again the fermionic vacuum.

4.2.2 Generalized Hartree-Fock

For Hamiltonians that are not quadratic, finding the ground state is in general exponentially difficult. However, the solution can be approximated using a variational approach, i.e. finding the state within some efficiently parametrized variational class that minimizes the expectation value of the energy. If a sufficiently powerful class of variational states is chosen, this approximates physical properties of the true ground state accurately. Choosing this variational class to be the Slater determinants, i.e. Gaussian fermionic states with fixed particle number, leads to the well-known Hartree-Fock approach. By considering the entire set of Gaussian fermionic states, i.e. including those with fluctuating particle number, one arrives at a generalized Hartree-Fock approach that is also able to capture superconductivity at the mean-field level, i.e. contains the ground states of

BCS theory where the superconducting order parameter has no quantum fluctuations [11]. It is known that there exist systems that are much better approximated by generalized Hartree-Fock than by non-generalized Hartree-Fock [29].

We now review this generalized Hartree-Fock approach, rederive the self-consistent iteration for its numerical solution, and clarify its relation to better-known approaches. Our starting point is a Hamiltonian that is quartic in the fermion operators:

$$\hat{\mathcal{H}} = -i \sum T_{ij} \hat{c}_i \hat{c}_j + \sum U_{ijk\ell} \hat{c}_i \hat{c}_j \hat{c}_k \hat{c}_l. \tag{4.5}$$

Here, T_{ij} is real and antisymmetric, while $U_{ijk\ell}$ is real and antisymmetric under exchange of any two indices, i.e. $U_{ijk\ell} = -U_{jik\ell} = U_{jki\ell} = \ldots$. Any quartic fermion Hamiltonian can be written in this form, including physically relevant cases such as the Hubbard Hamiltonian and the Coulomb interaction (see Appendix B.1 for details).

The energy for a Gaussian state $|\Gamma\rangle$ with corresponding covariance matrix Γ is easily evaluated using Eqn. (4.2) by recognizing that the expectation value of

the 4-fermion term is given by

$$\begin{aligned}\langle\Gamma|\hat{c}_i\hat{c}_j\hat{c}_k\hat{c}_l|\Gamma\rangle &= \mathrm{Pf}\begin{pmatrix} 0 & \Gamma_{ij} & \Gamma_{ik} & \Gamma_{il} \\ -\Gamma_{ij} & 0 & \Gamma_{jk} & \Gamma_{jl} \\ -\Gamma_{ik} & -\Gamma_{jk} & 0 & \Gamma_{kl} \\ -\Gamma_{il} & -\Gamma_{jl} & -\Gamma_{kl} & 0 \end{pmatrix} \quad (4.6)\\ &= \Gamma_{ij}\Gamma_{kl} - \Gamma_{ik}\Gamma_{jl} + \Gamma_{il}\Gamma_{jk}. \quad (4.7)\end{aligned}$$

We can now use the identities (repeated indices are summed over):

$$\begin{aligned}-\sum U_{ijkl}\Gamma_{ik}\Gamma_{jl} &= -\sum U_{ikjl}\Gamma_{ij}\Gamma_{kl} \\ &= \sum U_{ijkl}\Gamma_{ij}\Gamma_{kl} \quad (4.8)\end{aligned}$$

and

$$\begin{aligned}\sum U_{ijkl}\Gamma_{il}\Gamma_{jk} &= \sum U_{iljk}\Gamma_{ij}\Gamma_{kl} \\ &= \sum U_{ijkl}\Gamma_{ij}\Gamma_{kl} \quad (4.9)\end{aligned}$$

to arrive at the final expression

$$\langle\Gamma|\,\hat{\mathcal{H}}\,|\Gamma\rangle = \sum T_{ij}\Gamma_{ij} + 3\sum U_{ijk\ell}\Gamma_{ij}\Gamma_{kl}. \quad (4.10)$$

The factor of 3 can be viewed as counting the Hartree term, Fock term, and BCS term each, which are traditionally viewed as distinct. However, due to the symmetries of the Majorana representation, here these three terms all appear symmetrically.

We now have to find the pure-state covariance matrix Γ (i.e. satisfying the non-linear constraint $\Gamma^2 = -\mathbb{1}$) that minimizes the above expression. We can recover the typical self-consistent HF iteration by starting with an initial guess Γ^0 (satisfying $(\Gamma^0)^2 = -\mathbb{1}$) and expressing the new state as $\Gamma = \Gamma^0 + \delta\Gamma$ (where $\delta\Gamma$ is not by itself a valid covariance matrix). In terms of this new Γ and the starting point Γ^0, the energy is given by

$$\begin{aligned}
\langle\Gamma|\hat{\mathcal{H}}|\Gamma\rangle &= \sum T_{ij}\Gamma^0_{ij} + 3\sum U_{ijk\ell}\Gamma^0_{ij}\Gamma^0_{k\ell} && (4.11)\\
&\quad + \sum T_{ij}(\Gamma_{ij} - \Gamma^0_{ij}) + 3\sum U_{ijk\ell}(\Gamma_{ij} - \Gamma^0_{ij})\Gamma^0_{k\ell}\\
&\quad + 3\sum U_{ijk\ell}\Gamma^0_{ij}(\Gamma_{k\ell} - \Gamma^0_{k\ell}) + O(\left\|\Gamma - \Gamma^0\right\|^2)\\
&\approx \text{const.} + \sum \left[T_{ij} + 6U_{ijk\ell}\Gamma^0_{k\ell}\right]\Gamma_{ij} && (4.12)
\end{aligned}$$

Here, we have made the key approximation to neglect terms of order $\|\Gamma - \Gamma^0\|^2$ in order to arrive at a linear functional of Γ. Furthermore, we have used the same symmetries as in Eqns. (4.8), (4.9) to collect different terms together. We note that the final expression can be viewed as an effective quadratic Hamiltonian acting on Γ,

$$F_{ij} = T_{ij} + 6\sum_{k\ell} U_{ijk\ell}\Gamma^0_{k\ell}, \tag{4.13}$$

which is commonly referred to as Fock matrix. Its ground state is by construction a valid covariance matrix that satisfies $\Gamma^2 = -\mathbb{1}$. When the system has local

hopping and interaction terms T and U, they are sparse and will have only $O(N)$ entries, so the Fock matrix can be computed from Γ in $O(N)$ time. The iteration now proceeds by solving for the ground state of F, then replacing Γ^0 by that new state and recomputing the Fock matrix F, and repeating this procedure until convergence.

While it is known that this iteration cannot find the lowest-energy state in all cases [109], for many systems, especially those in which $\|U\| \ll \|T\|$, it is empirically known to converge rapidly and reliably to a global minimum energy. In other cases, there can be local minima or stable oscillations. The Optimal Damping Algorithm attempts to remediate this by choosing the minimum-energy convex combination $t\Gamma^{\text{new}} + (1-t)\Gamma^{\text{old}}$ [33]. Since the energy is a scalar quadratic function of t, this can be directly minimized through evaluation at any three values of t.

Computing Γ from H requires an eigenvalue decomposition, an operation which scales as $O(N^3)$ in general. In many cases, this will become the computational bottleneck and limit the system size for which Hartree-Fock can be used to several fermionic degrees of freedom. It is worth noting, however, that there are important use-cases for Hartree-Fock where this is not the bottleneck. For example, in quantum chemistry the Hamiltonian is non-local and the basis is not a real-space grid, such that there are $\mathcal{O}(N^4)$ terms in the Hamiltonian that need to

be computed as multi-dimensional integrals over the basis functions. In this case, computing the terms of the Hamiltonian is the bottleneck of the Hartree-Fock simulation. However, as we will see in the next section, when the Hamiltonian is local and the system quasi-one-dimensional, Gaussian fermionic tensor networks offer a more time- and memory-efficient approach.

4.2.3 Gaussian fermionic tensor networks

While a generic state on N particles can have as many as $N/2$ bits of entanglement across a cut, obeying what is commonly referred to as a volume law, the entanglement in low-energy states is typically much less. The situation is best understood for gapped, local Hamiltonians in one spatial dimension, which are known to have area-law entanglement in their ground state [46], i.e. the entanglement is bounded by a constant regardless of system size. The same behavior is expected for most systems also in higher dimensions [119]. The area law is typically violated in gapless systems; however, in many cases this violation is mild. For example, conformal field theories in 1D have only $O(\log(N))$ entanglement [65, 123], i.e. the area law is violated by a logarithmic correction.

Tensor networks [100, 99, 31] make use of the entanglement properties of low-energy states to represent them more efficiently. Matrix product states (MPS) are a particular class of tensor network states [51, 126, 101] that is known to be able to

efficiently represent the ground states of gapped one-dimensional Hamiltonians. Furthermore, MPS can be manipulated efficiently and the variational problem can in many cases be solved efficiently using the density-matrix renormalization group (DMRG) method [125]. The accuracy of the approximation can be controlled systematically using the *bond dimension* M of the MPS, which is the size of the matrices associated with each site in the lattice; as such, the computational cost scales with the third power of the bond dimension. The maximum bipartite entanglement that can be captured in an MPS is bounded by $\log(M)$, and therefore M needs to grow exponentially with the entanglement in the system. While rigorous bounds for the scaling of MPS simulations are available [88, 67], heuristically one often finds an approximately linear scaling of the computation time with system size for gapped systems.

However, if the underlying Hamiltonian is quadratic, this exponential scaling in entropy can be improved further [87, 49, 110, 45, 59, 52, 50, 60, 71, 108]. A conventional tensor network can be understood as associating a quantum state with the vertices of a graph (which may or may not be the underlying lattice); the degrees of freedom on these states are associated with edges of the graph and can be physical or auxiliary. The physical quantum state is recovered by projecting the auxiliary degrees of freedom on each edge of the graph onto a maximally entangled state. One can now choose these quantum states associated with the vertices of

the graph to be Gaussian states, i.e. states satisfying Wick's theorem, and choose the maximally entangled state that the edges are projected onto as a Gaussian state as well. In this case, the physical state being represented is Gaussian as well, and the entire computation can be performed in terms of covariance matrices of the states. This representation inherits most properties of general tensor network states; however, the exponential scaling with the entanglement entropy is replaced by a polynomial scaling, i.e. the ansatz is exponentially more efficient in terms of its scaling with entanglement entropy. This construction was used to obtain practical, efficient algorithms for one-dimensional systems of free fermions using Gaussian fermionic matrix-product states (GFMPS) in Ref. [108]; these methods form the basis of the efficient gHF calculations presented in this chapter.

On a technical level, a GFMPS is obtained by associated to each site i on a chain a pure Gaussian state $|\gamma_i\rangle$ with covariance matrix γ_i. The fermionic modes on each state can be assigned to three groups: physical modes and auxiliary modes connecting to the left and right. These auxiliary degrees of freedom can be thought of as capturing entanglement to the left and right of the system, respectively, and the physical state is obtained by projecting the right auxiliary modes on site i with the left auxiliary modes on site $i+1$ onto a maximally mixed state (often referred to as "tracing out" or "contracting"), so that only physical modes are left. The number of auxiliary modes on each bond, which we denote

as χ and which should be viewed as hyperparameter refining the ansatz similar to the bond dimension M for conventional MPS, bounds the bipartite entanglement in the state by $S \leq \chi \log \sqrt{2}$, i.e. the maximal entanglement is linear rather than logarithmic in the case of conventional MPS.

As a crucial ingredient for practical calculations, Ref. [108] describes the canonical form for GFMPS, efficient computation of the total energy as well as a way to express the total energy as a linear function of a local tensor. These components together allow for a straightforward generalization of standard MPS techniques, such as the DMRG algorithm, which (starting from an initial guess for the state, for example a completely random state) finds an approximation of the ground state of the system by iteratively optimizing each tensor (or pairs of tensors) in the MPS. This optimization is swept back and forth across the system until convergence is reached. While a detailed review of the technical aspects of GFMPS calculations is beyond the scope of this manuscript, we review some key aspects of the GFMPS method in Appendix B.3.

For the discussion of our numerical results below, an important practical difference between conventional MPS and GFMPS is that in the latter case, it can be advantageous to group several physical sites together and form a lattice of such blocks. We will therefore typically refer to a block of B sites, which is a single site in the GFMPS but encompasses B physical sites. Choosing χ and B must

be done carefully, and one must generally ascertain convergence with respect to χ. For a given χ, it is typically close to optimal to choose blocks of size 2χ (if the goal is to minimize memory) or χ (if the goal is to minimize computation time).

4.2.4 gHF using GFMPS

In the full solver, the gHF iteration forms an outer loop; its pseudocode is shown in Fig. 4.1. In each iteration, it queries the covariance matrix Γ^0 from the underlying GFMPS representation, builds an effective potential F from the covariance matrix Γ^0, and then passes this new potential to the DMRG solver to obtain an updated Γ in GFMPS form. In the inner loop, several sweeps of the DMRG optimization are performed to obtain the lowest-energy GFMPS for a given Fock matrix F. The GFMPS is re-used as the initial state for the DMRG solver in the next gHF iteration in order to speed up convergence.

In principle, the entire dense covariance matrix Γ can be extracted from a GFMPS. However, this would require $O(N^2)$ memory and negate the time and memory savings of the GFMPS approach. We focus on a local Hamiltonian, where T and U connect each block to only a small number of other blocks. Then we only need to extract a block-space Γ, populating the blocks that are connected by T and U. Strictly local models like the Hubbard model have only intra-block quartic terms (U), and inter-block quadratic terms (T), so that a sparse Γ can be extract

Algorithm 1 GFMPS gHF

function GFMPS-GHF(T, U)

gfmps ← random initial state

GFMPSDMRG(T, *gfmps*);

Γ = EXTRACTGAMMA(*gfmps*);

$E_0 \leftarrow \infty$

for $s \leftarrow 1$ to $maxIter$ **do**

$F \leftarrow 6U_{ij,kl}\Gamma_{k,l} + T_{i,j}$;

GFMPSDMRG(F, *gfmps*);

$E_{\text{new}} \leftarrow (F_{i,j} + T_{i,j})\Gamma_{j,i}/2$

$\Delta E \leftarrow E_0 - E_{\text{new}}$

If $|\Delta E| < \Delta E_{\text{target}}$, **break**;

Γ = EXTRACTGAMMA(*gfmps*);

$E_0 \leftarrow E_{new}$

end for

end function

Figure 4.1: Pseudo-code description of the gHF iteration using a GFMPS-based solver.

in $O(N)$ time; the same asymptotic scaling will be preserved for models with finite but bounded range (e.g., next-nearest-neighbor interactions). Conversely, a non-local interaction such as an unscreened Coulomb interaction with $1/r$ decay would add terms between all pairs of blocks, and thus require computing all elements of Γ and lead to a dense Γ and F. In this unfavorable case, the GFMPS-based approach would recover the computational cost of the dense approach. In some cases, for example for screened Coulomb interaction of the form $e^{-r/\xi}/r$, it may be possible to introduce a sharp cutoff and set all interaction terms beyond this distance to zero in order to recover the linear scaling.

Pseudocode for the subroutines GfmpsDmrg and ExtractGamma can be found in Appendix C. It is important to note that the same *gfmps* object is being used across iterations, and the previous state computed by GfmpsDmrg is used as input to the next GfmpsDmrg. After the first one or two iterations, the effective potential F will not change much, so the previous state of *gfmps* is a good initial state for the DMRG solver.

4.3 Results

4.3.1 Model

To demonstrate the efficacy of the approach, we study the Hubbard model on a two-dimensional rectangular lattice with a quadratic anisotropic trapping potential, loosely modeling trapped quantum gases [36, 42, 113]. The number operator at a site $\mathbf{r}$ is $\hat{n}_{\mathbf{r}\sigma} = \hat{a}^\dagger_{\mathbf{r}\sigma}\hat{a}_{\mathbf{r}\sigma}$, and the electron density as its expectation $\langle \hat{n}_{\mathbf{r}\sigma} \rangle$. The superfluid density is $\frac{1}{2}\langle \hat{a}_{\mathbf{r}\uparrow}\hat{a}_{\mathbf{r}\downarrow} + \hat{a}^\dagger_{\mathbf{r}\uparrow}\hat{a}^\dagger_{\mathbf{r}\downarrow} \rangle$. The Hamiltonian is given by

$$\hat{\mathcal{H}} = \hat{\mathcal{H}}_0 + \hat{\mathcal{H}}_{\text{int}} + \hat{\mathcal{H}}_{\text{trap}} \tag{4.14}$$

$$\hat{\mathcal{H}}_0 = -t \sum_{\langle \mathbf{r}_1, \mathbf{r}_2 \rangle, \sigma} \hat{a}^\dagger_{\mathbf{r}_1\sigma}\, \hat{a}_{\mathbf{r}_2\sigma} - \mu \sum \hat{n}_{\mathbf{r}_1\sigma} \tag{4.15}$$

$$\hat{\mathcal{H}}_{\text{int}} = U \sum_{\mathbf{r}} \left(\hat{n}_{\mathbf{r}\uparrow} - \frac{1}{2} \right) \left(\hat{n}_{\mathbf{r}\downarrow} - \frac{1}{2} \right) \tag{4.16}$$

$$\hat{\mathcal{H}}_{\text{trap}} = \sum_{\mathbf{r},\sigma} (V_x r_x^2 + V_y r_y^2) \hat{n}_{\mathbf{r}\sigma}, \tag{4.17}$$

where $\hat{a}^\dagger_{\mathbf{r}\sigma}$ creates a fermion of spin σ on site $\mathbf{r} = (x, y)$ of the lattice, by $\langle \mathbf{r}_1, \mathbf{r}_2 \rangle$ we denote pairs of nearest-neighbor pairs, and the spin index σ runs over $\{\uparrow, \downarrow\}$. Here, t denotes the hopping strength, μ sets the chemical potential and thus controls the filling of our system (noting that due to the Majorana representation being used in our method, we don't fix the particle number), U is the strength of the on-site Hubbard interaction, and V_x and V_y control the properties of the harmonic trap. We quote all energy scales below in units of the hopping t.

We note that in the metallic phase, the Hubbard model exhibits a mild violation of the area law: for a system of width W and length L, with $L \gg W$, the entanglement entropy will scale as $W \log L$. However, as previously demonstrated in Ref. [108], the scaling of the GFMPS approach remains cubic in W and only slightly higher than linear in L.

In addition to the parameters of the physical model, there are the parameters of the method. Both dense gHF and GFMPS are run until $\Delta E < 10^{-3}$, where ΔE is the energy difference after subsequent iterations. The GFMPS has additional parameters for the bond dimension χ and block size B. Lengths were picked to always be multiples of B, so that all blocks were equal size. We generally performed 4 GFMPS DMRG sweeps per gHF iteration and use the single-site DMRG algorithm [127].

4.3.2 Square systems

While the GFMPS approach is much better suited to quasi-one-dimensional systems, i.e. where the length L far exceeds the width W, we first test the accuracy of the approach for a square system with $W = L = 32$. Although this not in the $L \gg W$ regime we are primarily interested in, this allows us to ascertain what convergence rates or hyperparameters we might expect, and the $O(L^3)$ scaling of the GFMPS on square systems still outperforms the $O(L^6)$

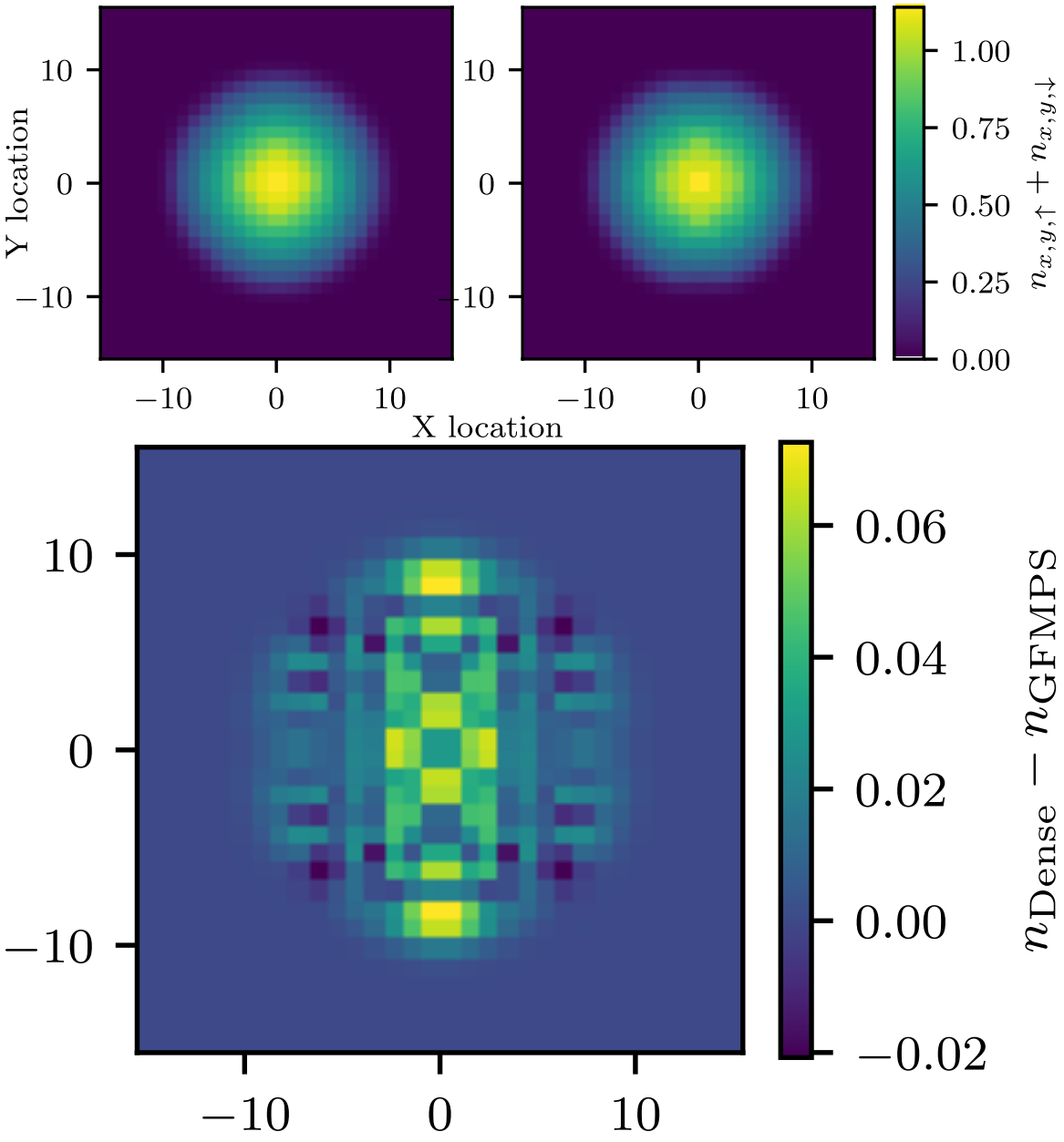

Figure 4.2: Top-left: filling fraction over space with standard Hartree-Fock. Top-right: with GFMPS accelerated method. Bottom: difference between top two, contrast enhanced 13x.

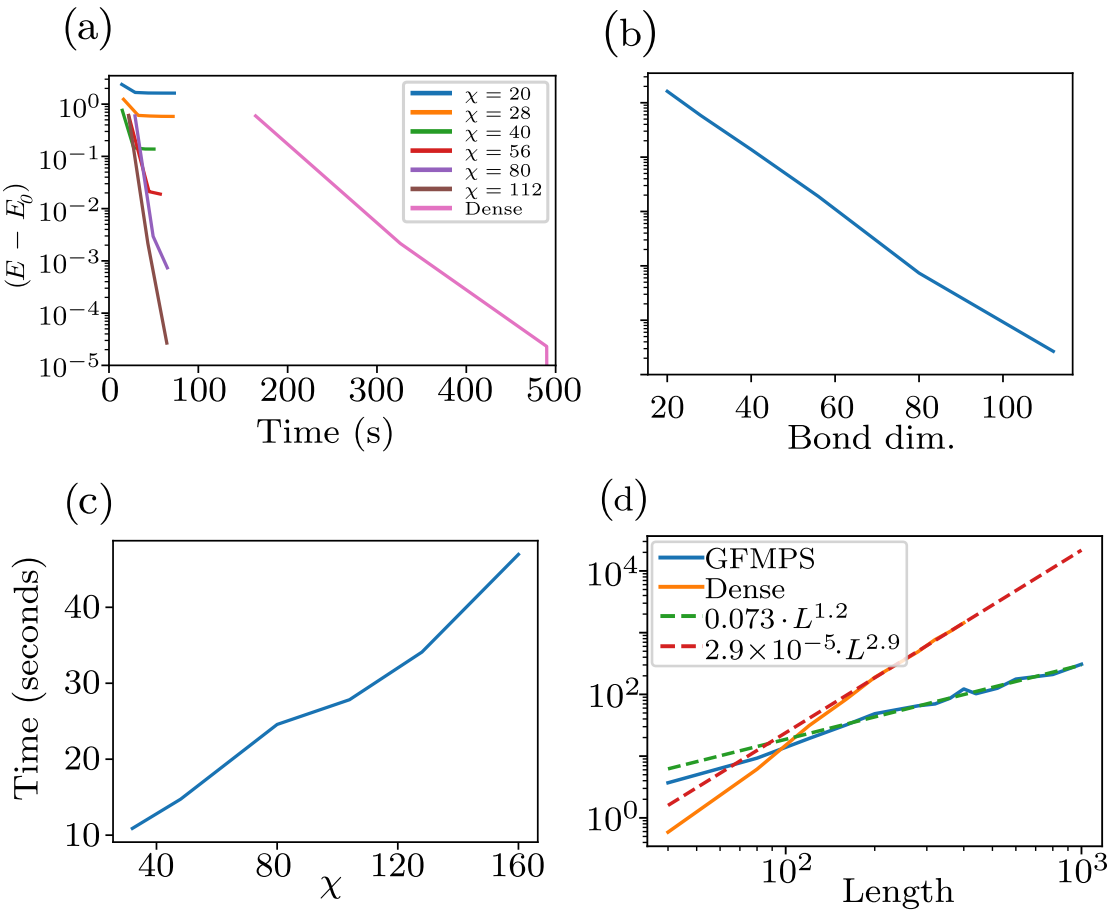

Figure 4.3: (a) Convergence of the energy as function of CPU time for various bond dimensions χ as well as for the dense Hartree-Fock solver. Here we have used $L = 280, U = 0.4t$. A bond dimension of $\chi = 80$ sufficed to achieve similar energy to the dense solver ($< 10^{-3}$ error), but ran 7.5x faster. Energy is relative to the final result of the dense computation, which converged on the 4th iteration after 653 s. (b) Convergence of the energy estimate as function of bond dimension. (c) Comparing the time for one Hartree-Fock iteration at different bond dimensions, for $L = 400t, U = 0.4t$. (d) Time required to run Hartree-Fock to convergence ($\delta < 0.001$) on varying system lengths. The standard dense approach displays roughly $O(n^3)$ time, while the GFMPS scales close to linearly. Dashed lines are the lines of best fit (power law fits).

traditional Hartree-Fock. The parameters of the Hamiltonian were chosen as $U = 0.4t$, $V_x = V_y = 0.02t$, $\mu = 0.3t$, $\chi = 32$, $B = 8$. This puts it in the weakly repulsive regime $0 < U/t < 1$. The GFMPS DRMG method found a state of energy -6791.37 and peak filling 1.140, while the full dense method found -6793.47 and peak filling 1.169. The largest difference in filling was just off-center, with 0.072. This gave agreement within a relative error of 10^{-4} for the energy and about $6 \cdot 10^{-2}$ for the filling. Shown in Fig. 4.2 are the densities of GFMPS and dense solution in the top two panels, and the difference between the two in the bottom panel. We can see that the 90-degree rotational symmetry of the physical system is broken by choosing how the physical system is mapped onto the one-dimensional arrangement of the GFMPS. In Fig. 4.2, the sites of the GFMPS are arranged along the horizontal direction of the system. The vertical and horizontal reflection symmetry still remains in the GFMPS solution.

4.3.3 Computational performance for quasi-one-dimensional systems

To evaluate the performance benefit of the GFMPS approach for quasi-one-dimensional systems, we turn our attention to systems of fixed width $W = 4$ and varying length L. For the other parameters, we choose $U = 0.4t$, $V_x = V_y = (6/L^2)\, t$, $\mu = 0.3t$, $\chi = 4$, $B = 8$. Each run of the DMRG used 4 sweeps. The

potential was chosen to scale $V_x \sim L^{-2}$ so that the fraction of the trap occupied by particles stays roughly constant, with the potential rising from 0 in the center to 1.5 at the edges.

The first test was to see how our method compares to dense generalized Hartree-Fock. Because it is a variational approach, the minimum energy attained is our primary figure of merit. We also want to ensure that we see the linear scaling of computation time with the length of the trap for the GFMPS-based calculations, as compared to a cubic scaling for conventional, dense Hartree-Fock.

First, we held $L = 280$ fixed and observed accuracy and runtime with different bond dimensions. We found that at $\chi = 80$, the error in energy was < 0.001, which represents capturing almost all the energy that Hartree-Fock can. The GFMPS computation took only 65 seconds, as opposed to 490 seconds with dense Hartree-Fock. This represents a 7.5x speedup. Results with other bond dimensions are shown in panels (a)-(c) of Fig. 4.3.

The same value $\chi = 80$ was then used across a broad span of lengths to see how the computation time scaled with system size for otherwise fixed parameters. Our results are shown in Fig. 4.3(d). We expect that the GFMPS scales approximately linearly with length, while the dense method, which requires diagonalizing an $O(L)$ size matrix, would scale cubically. Fitting power laws $t = a \cdot L^p$ to each yielded exponents of 1.20 for the GFMPS and 2.95 for the dense methods, in

good agreement with expectations. Despite holding χ fixed, the error in energy did not increase significantly, staying below 10^{-3}.

4.3.4 Repulsive case

Having established the improved performance of the GFMPS-based solver for weakly interacting systems, we now investigate whether the expected behavior is found also in more strongly interacting cases, starting with the case of repulsive interactions. To this end, we increase the interaction strength to $U/t = 3.0$. The phase diagram of the translation-invariant Hubbard model is well-understood [89, 48] and it is known that as μ varies there are separate partially-filled (compressible) and half-filled (incompressible) phases. The half-filled phase occurs in a region of chemical potential $\mu \in [-\mu_0, \mu_0]$ centered around the half-filled point $\mu = 0$. The critical value μ_0 is given by [48, 89]

$$\mu_0 = -2 + \frac{u}{2} + 2\int_0^\infty \frac{d\omega}{\omega} \frac{J_1(\omega)e^{-\omega u/4}}{\cosh(\omega u/4)} \tag{4.18}$$

$$\approx \exp(-6/u) \tag{4.19}$$

with $u = U/t$ and the approximation good for $u \ll 1$. By creating an effective $\mu > -\mu_0$ in the center of our trap and $\mu < -\mu_0$ on the edges, we should see distinct regions appear in the same system.

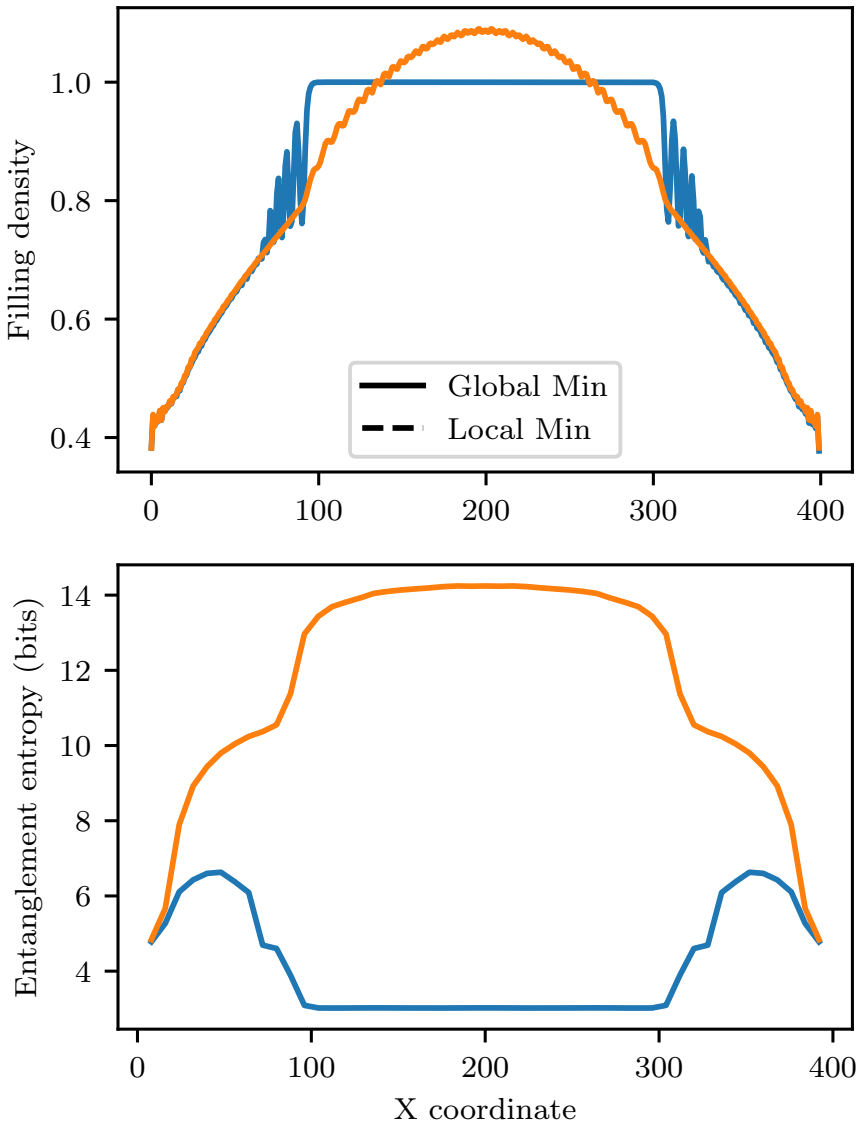

Figure 4.4: Comparison of the metastable state found by standard Hartree-Fock iteration, and the true global minimum HF configuration. Top: Filling fraction along the length of the system. Bottom: entanglement entropy across different cuts of the system. Orange line is the local minimum which fails to avoid the repulsive energy penalty, and has accordingly higher entanglement entropy in middle of the system. Runs from several different initial random states reliably converged to these same two cases.

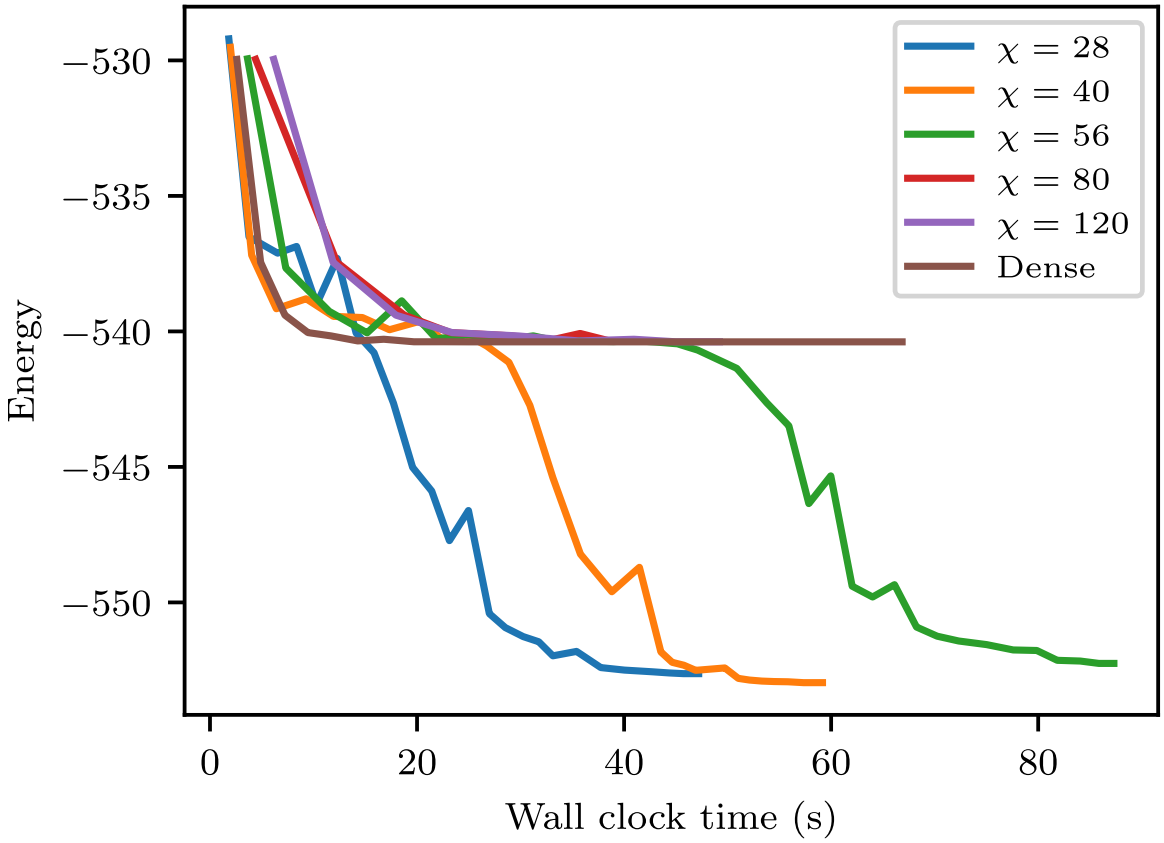

Figure 4.5: Evolution of energy over time as GFMPS DMRG is run with different bond dimensions. It was expected that smaller bond dimensions would fall more quickly, but bottom out at higher energy. We found instead that higher bond dimensions became stuck at a higher energy, due to a local energy minimum.

We ran with system parameters of $U = 3.0t$, $V_x = V_y = 6/L^2\, t$, $\mu = 0.3t$, $\chi = 40$, $B = 8$. The GFMPS produced a solution in line with the two-phase result we expected. Its electron density is shown as the solid line in the top panel of Fig. 4.4. As anticipated, we find an expected region of unit filling in the center of the trap and a continuous decay to zero filling towards the edge.

Comparing the result of the GFMPS-based calculation with the result of the dense solver, we find that the latter converged to a state with considerably higher energy. Furthermore, as shown in the dashed line in the top panel of Fig. 4.4, the state did not exhibit the extended plateau of unit filling in the center of the system. This was surprising as we generally view the GFMPS as a more restricted ansatz, and therefore expect the dense solver to produce lower energies. In this case, however, it turns out that the dense solver becomes trapped in a stable yet unphysical fixed point of the Hartree-Fock iteration, a local minimum which is avoided by the GFMPS-based solver.

To understand why the GFMPS is able to avoid this fixed point, we compared the entanglement entropy of the global minimum and the local minimum, as shown in the bottom panel of Fig. 4.4. The entanglement of the local minimum is much larger than that of the global minimum, which in the center of the trap potential corresponds to an incompressible and thus weakly entangled state. This hints at why the GFMPS is able to avoid this local minimum: like all tensor-network

based approaches, it is (at finite bond dimension) biased towards low-entanglement solutions, thus making it more likely to find the incompressible state.

To confirm this, we reran the GFMPS solver with much larger bond dimensions, where its behavior should more closely resemble that of the dense solver. Results are shown in Fig. 4.5. We find that indeed for bond dimensions $\chi \geq 80$, the GFMPS-based solver becomes trapped in the same local minimum as the dense solver. Intermediate bond dimensions may become trapped in this local minimum for a few sweeps, but eventually find the global minimum. Overall, this suggests that it may in some situations be beneficial to limit the bond dimension of the GFMPS at least in initial sweeps of the self-consistent iteration.

4.3.5 Attractive case

When $U < 0$, the interaction between fermions is attractive, and we expect the appearance of finite superconducting pairing as measured by $\hat{a}_{i,\uparrow}\hat{a}_{i,\downarrow} + \hat{a}^\dagger_{i,\uparrow}\hat{a}^\dagger_{i,\downarrow}$. At the half-filled point $\mu = 0$, the Hubbard model on bipartite lattices has a $U \to -U$ symmetry corresponding to applying $(\hat{a}_{i,\uparrow} + \hat{a}^\dagger_{i,\uparrow})(\hat{a}_{i,\downarrow} + \hat{a}^\dagger_{i,\downarrow})$ at every other site. To study specifically the behavior of the superconducting phase, we thus study densities away from half-filling.

We simulate at $\mu = 0.5$, $L = 2000$, varying U/t in order to see both the weakly- and strongly-interacting cases.

As expected, we observe that the convergence is fast for weak interactions like $U = -0.5t$, where the Hartree-Fock procedure does not modify the potential as greatly between iterations. $U = -4t$ converged more slowly in the middle, as discussed in the previous section, likely for similar reasons of gradually adjusting the potential. All solutions showed electron density oscillations, especially pronounced in the vicinity of the $\langle n_i \rangle = 1$ point (the x location with a filling of approximately one fermion per site). Running the Hartree-Fock iteration for many more steps gradually reduced the amplitude of the oscillations, but they did not go away, suggesting that these Friedel-like oscillations are genuine physical effects, but that the search procedure may be prone to overestimating them.

It is well-known that many qualitative features of the superconducting phase are well-captured by the BCS [13] mean-field solution, which can be viewed as a more restricted version of the gHF ansatz. For the translationally-invariant case, the mean-field solution can be obtained semi-analytically by solving the gap equation, which relates the filling fraction n, superconducting gap Δ and the interaction strength U, and in one dimension takes the form

$$\begin{aligned} \xi(k) &= -2\cos(k) - \mu - U\left(n - \frac{1}{2}\right) && (4.20) \\ n &= \frac{1}{2\pi}\int_0^{2\pi} \Theta\left(-\xi(k)\right)\, dk && (4.21) \\ \frac{2}{|U|} &= \int_0^{k_F} \frac{dk}{\sqrt{\xi(k)^2 + \Delta^2}}. && (4.22) \end{aligned}$$

To compare to our numerical gHF solution, we can consider what we refer to as "local gap approximation" (following the "local density approximation" widely used in density-functional theory calculations), where we approximate the solution at each point in space by the solution of the (translationally-invariant) gap equation for the parameters at that point in space. This should be appropriate in the limit where the potential varies very slowly compared to the coherence length of the superconductor.

At $U = -2$, where the coherence length is on the order of a few lattice sites, this local gap approximation accurately reproduces the numerical gHF solution, as shown in the top panel of Fig. 4.6. The deviation is most pronounced at the edges of the system, where the local gap approximation has the superfluid density drop sharply to zero, while in the true inhomogeneous problem it tapers off over a few sites. At the much weaker interaction strength of $U = -0.65$, which is much more representative of real-world conditions, where superconducting pairing is a weak effect compared to the Fermi energy, a very distinct picture emerges. Our results for this regime are shown in the bottom panel of Fig. 4.6. We find that the local gap approximation is much less accurate at capturing the inhomogeneous solution, underestimating the strength of superconducting pairing by approximately a factor of 2 in the center of the system. Our GFMPS-based approach is still able

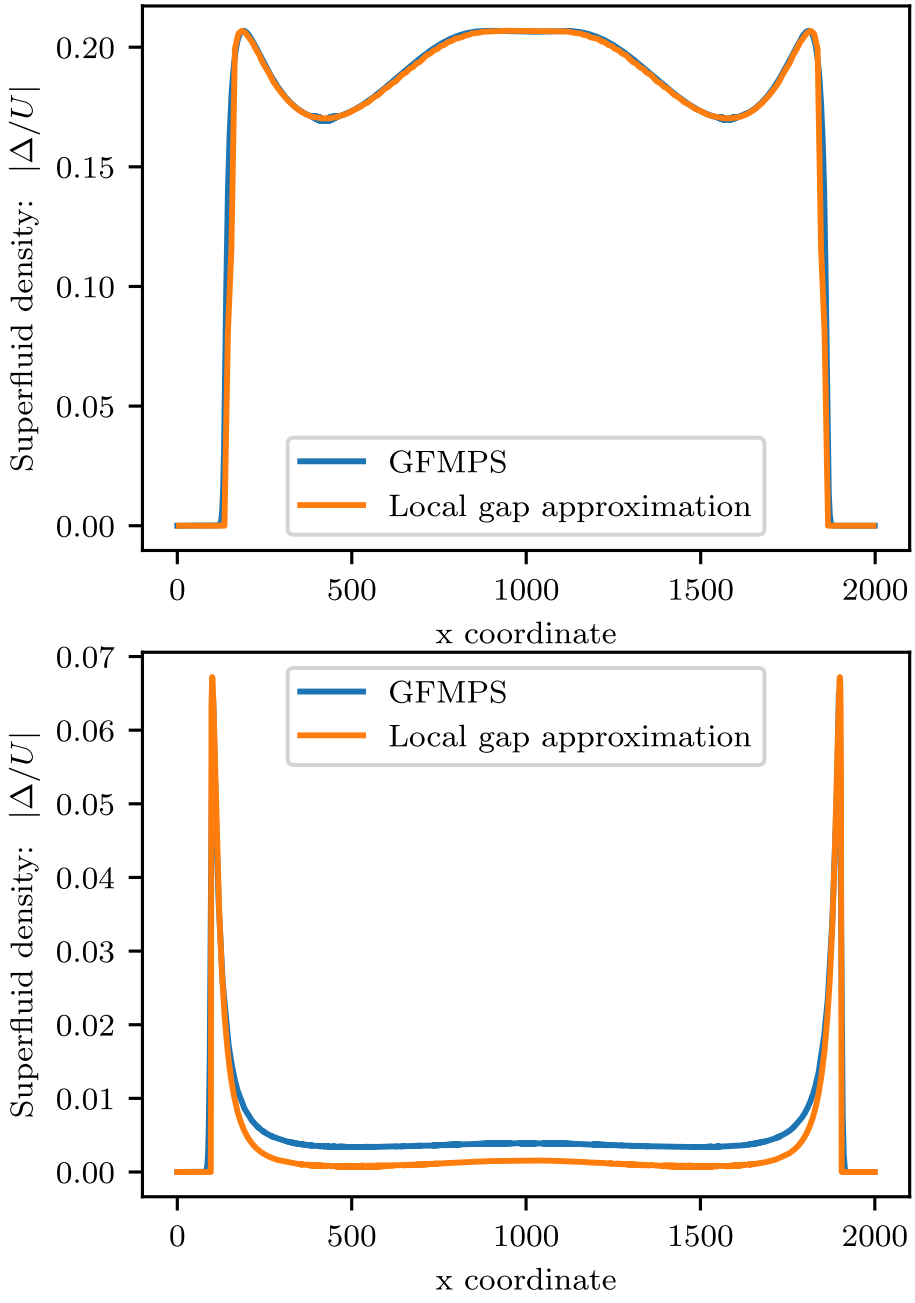

Figure 4.6: Comparison of gHF-GFMPS calculation of superfluid density with predictions of superfluid density from the BCS gap equation. Both have $V = 6t/L^2$. Top figure: $\mu = -1$, $U = -2$. Bottom figure: $\mu = -0.75$, $U = -0.65$. As the local gap approximation becomes more accurate as L increases, a larger L of 2000 was chosen for the bottom figure to show how the differences persist.

to solve this inhomogeneous system of several thousands of degrees of freedom in 113 seconds.

4.4 Outlook

In this chapter, we have demonstrated that using numerical methods based on Gaussian tensor networks can accelerate computational methods that map many-body electron problems onto effectively non-interacting problems. This family includes Hartree-Fock and self-consistent BCS, which can be unified into the generalized Hartree-Fock method used here, but also other widely used approaches such as density functional theory. We thus expect this general approach to be applicable to a wide array of problems.

Application areas where an inhomogeneous real-space solution may be particularly important include systems with very large unit cells, which is a feature typically found in Moiré materials such as twisted bilayer graphene (tBLG) [9]. Our approach seems suitable, for example, to extend recent studies of tBLG using hybrid Wannier orbitals to larger systems [63]. Similarly, in mesoscopic device physics, inhomogeneities in the system often play an important role, and the methods put forward here suggest a pathway to realistic simulations of such structures.

Chapter 5

Green's Function Estimation

5.1 Introduction

For a Hamiltonian H, a unitary excitation[1] operator $\mathcal{O}$, and a state $|\psi\rangle$, we define the real-time Green's function as

$$G(t) = \langle\psi| \exp(iHt)\mathcal{O}^\dagger \exp(-iHt)\mathcal{O}|\psi\rangle \tag{5.1}$$

and the real-frequency Green's function as the Fourier transform,

$$\tilde{G}(i\omega) = -i \int_0^\infty G(t)e^{-t\omega}\, dt$$

The imaginary-frequency Green's function is also of interest, but cannot be implemented directly on a quantum computer, as the corresponding imaginary-time evolution is not unitary. With the scheme described in [17], many samples of $G(t)$ are taken with a smaller quantum computer are collected into a (classical) computer's database, where the signal is then Fourier transformed to produce an estimate of $\tilde{G}(i\omega)$. While the quantum hardware requirements are modest, the number of samples required for chemical accuracy is high.

However, we have other information about the signal $G(t)$ that should, in principle, allow us to better estimate it from fewer samples. We describe and compare several schemes for better estimation of $\tilde{G}(i\omega)$, and show that they should lead to polynomial reduction in sample complexity, with an advantage of roughly 100x

[1] The most familiar Green's function G' uses $\mathcal{O}' = a^\dagger$, a (nonunitary) creation operator. Defining $\mathcal{O}_\pm = a^\dagger \pm a$, with Green's functions $G_\pm$, we can recover $G' = \frac{G_+ + G_-}{2}$, so restricting ourselves to unitary excitations here suffices.

at chemical accuracy scales. Our first main approach uses Gaussian processes, and we discuss choices of kernel function and complex phases. Our second main approach uses the L_1 boundedness of its Fourier transform $\tilde{G}(\omega)$ to impose constraints that greatly enhance convergence. We discuss uncertainty estimation and adaptive sampling.

5.2 Measurement Model

We assume that the state $|\psi\rangle$ can be efficiently prepared on the quantum computer, and that the (unitary) operator $\mathcal{O}$ is easily implemented by gates as well. The time evolution $\exp(iHt)$ cannot in general be carried out exactly, but a Trotter decomposition of the Hamiltonian allows for very efficient approximation. We will neglect the Trotter error and focus on the noise inherent in the sampling and reconstruction process. We also assume that the gate error and decoherence rate of the quantum hardware is negligible.

The real or imaginary part of the Green's function can then be estimated by a one-bit phase estimation algorithm, see Appendix D. A real-part measurement of $G(t)$ has result of a Bernoulli distribution with $p = \frac{1+\Re(G(t))}{2}$, resp. the imaginary part. The Green's function obeys $|G(t)| \leq 1$, so this probability is always in the range $[0, 1]$. Other measurements are possible: $|G(t)|^2$ can be estimated by undo-

ing the preparation of the state $|\psi\rangle$, or more accurate estimates of the phase $G(t)$ through more quantum phase estimation. These are typically less informational however, so we concentrate on direct measurements of the real and imaginary parts of $G(t)$.

5.3 Simple Reconstruction Methods

We start by discussing baseline methods for reconstructing G. We will then move on to Gaussian process regression and Fourier-constrained fitting.

5.3.1 Simple Baseline

Given the ability to sample $\Re(G(t))$ and $\Im(G(t))$, we would like to estimate $G(t)$ so that we can numerically integrate against it; the importance of $G(t)$ at a particular t falls off exponentially with t. In the simplest approach, we discretize time to N points t_n, sample each S many times in the real and imaginary parts, and estimate that $G(t_n)$ is given by the sample mean at that point:

$$G(t_n) \approx \frac{k_{\Re,pos} - k_{\Re,neg}}{S/2} + i\frac{k_{\Im,pos} - k_{\Im,neg}}{S/2}$$

The uncertainty at each point will be approximately $\sim \sqrt{1/3S}$. The exact constant depends on prior distribution of $G(t)$. The sample points t_n can then be

used to build a proxy $G(t)$ for all $t \geq 0$, for example by linear interpolation, which is then integrated to compute $\tilde{G}(t)$. This will form our baseline method.

The sample points t_n can also be used directly as the sample points for the numerical integration. As nearly all non-adaptive quadrature methods derive their weights by first fitting a function to several points, this is merely offloading the construction of the $G(t)$ to the quadrature algorithm. We can frame the problem entirely in terms of estimating $G(t)$, then, and this allows us to compare other reconstruction approaches on equal footing. In [17] for instance, the authors used Simpson's rule for integration, which is equivalent to piecewise quadratic interpolation.

It can be shown that the optimal allocation of S samples for piecewise linear interpolation has a density of $O(S^{1/5})$ different t values, and the average L^2 error will scale as $S^{-0.4}$, assuming bounded first derivatives of $G(t)$.

5.3.2 Per-Point Bayes

The first improvement is to look at each point $G(t)$ more intelligently. Each sample tells us about $\Re(G(t))$ or $\Im(G(t))$ individually, but they are not independent. Since $|G(t)| \geq 1$, these two components must lie in a unit disk; we can take some Bayesian prior over this disk, such as the uniform prior, to account for this relationship. The resulting estimator does not have a closed form and requires

numerical integration. In the typical case where $|G(t)| \neq 1$ however, this estimator provides no asymptotic advantage, as the real and imaginary parts converge independently as quickly as they do together.

5.4 Gaussian Processes and BLUPs

The first asymptotic sampling advantage we can gain comes from the fact that $G(t)$ is a smooth function, so that our prior expects the values $G(t)$ and $G(t+\delta t)$ to be highly correlated. If the sampled times are significantly denser than the scale over which G varies, $1/|G'(t)|$, then this can reduce our sample complexity by a corresponding factor.

Effectively using this information requires a prior over the possible functions G. Although at each point $G(t)$ there is one true complex value, we can associate to it a pair of Bernoulli random variables $X(t)$ and $Y(t)$, such that $G(t) = \langle \mathbf{X}(t) \rangle + i \langle \mathbf{Y}(t) \rangle$. By some abuse of notation, we will use $G(t)$ to also represent the variable $\mathbf{X}(t) + i\mathbf{Y}(t)$. Then the problem of reconstructing G is equivalent to building a good estimator for the mean of the associated random variable, G.

In general, any stochastic process $\mathbf{X}(t)$ has some mean $M(t) = \langle \mathbf{X}(t) \rangle$ and covariance $K(t_1, t_2) = \mathrm{CoV}(\mathbf{X}(t_1), \mathbf{X}(t_2))$. If we assume (or know) the functions M and K, we can compute the Best Linear Unbiased Predictor (BLUP), which

produces the optimal prediction of the form

$$\tilde{X}(t) = c_0 + \sum c_i k_{t_i},$$

where k_{t_i} are the sampling results and the c_i depend only on the sampling times $\{t_i\}$. Gaussian processes [105] are the special case where all finite marginals of $\mathbf{X}(t)$ are multivariate Gaussian distributions. In this case, the prior is completely determined by M and K, and the BLUP is also the optimal predictor (even among nonlinear predictors).

5.4.1 Ordinary Kriging

A reasonable approach is to then approximate $G(t)$ with a Gaussian process and use the BLUP; this technique is also known as "Kriging". The two complex components, $\Re(G(t))$ and $\Im(G(t))$ become two separate random variables, that can have zero covariance (in which case the real and imaginary parts can be estimated separately) or nonzero (resp. jointly), depending on the form of K. This approximation makes some drastic concessions: for one, $G(t)$ has bounded support on the complex unit disk, quite different from a Gaussian. Second, $G(t)$ has nonlocal constraints on its functional form, such as analyticity, that cannot be captured in two-site correlations K. Nonetheless, this approach can be proven to asymptotically improve sample efficiency. The standard equations for ordinary kriging in a single variable are reviewed in Appendix E.

When our function $G(t)$ has separate real and imaginary parts, it is easier to express the kriging process in terms of the two correlated functions $X(t)$ and $Y(t)$. The covariance function K now has four components, indicating which pairs of observations we're comparing:

$$K(x,y) = \begin{bmatrix} K_{1,1}(x,y) & K_{1,2}(x,y) \\ K_{2,1}(x,y) & K_{2,2}(x,y) \end{bmatrix}$$

If we have a phase-invariant prior (that is, a prior that assigns an equal likelihood to a function $f(x)$ as to $e^{i\theta}f(x)$), we must have

$$K_{1,1}(x,y) = K_{2,2}(x,y)$$

$$K_{1,2}(x,y) = -K_{2,1}(x,y).$$

The second equation implies that $K_{1,2}(x,x) = 0$. If furthermore the prior is stationary, then

$$K_{1,1}(x,y) = K_{2,2}(x,y) = K_1(|x-y|)$$

$$K_{1,2}(x,y) = -K_{2,1}(x,y) = \text{sign}(x-y)\, K_2(|x-y|)$$

Ordinary kriging allows a nonzero mean, which now means the two values c_{re} and c_{im}. For Green's functions as in Eq. 5.1, one can show that $c_{im} = 0$. The resulting unbiasedness condition actually cannot be correctly handled as a particular case of ordinary kriging. In the real-valued, single-function setting, the unbiasedness condition gives the constraint that the sum of weights $\sum w_i = 1$.

In this complex-valued setting, when we producing a prediction for the real part $X(t)$, the weights on our real measurements must sum to one, but the weights on imaginary measurements are not constrained. When producing a prediction for the imaginary part $Y(t)$, the weights on real measurements must instead sum to zero.

The resulting augmented kernel $\mathbf{K}$, and weights $\mathbf{w}$, are

$$\mathbf{K}_{aug} = \begin{bmatrix} K(x_1,x_1) & \cdots & K(x_1,x_n) & s_1 \\ \vdots & \ddots & & \vdots \\ K(x_n,x_1) & \cdots & K(x_n,x_n) & s_n \\ s_1 & \cdots & s_n & 0 \end{bmatrix} \quad \mathbf{w} = \mathbf{K}_{aug}^{-1} \begin{pmatrix} K(x_1,x^*) \\ \vdots \\ K(x_n,x^*) \\ s_* \end{pmatrix}.$$

Here s_i is 1 if the sample x_i is real, and 0 otherwise; and s_* is 1 if the value being predicted at x_* is real, and 0 otherwise. It is understood that each $K(\cdot,\cdot)$ is taking the correct component $K_{a,b}$ depending on the type of measurements at each argument.

5.4.2 Complex Kernels

It remains to choose an appropriate kernel function K. In univariate (i.e. real) kriging, typical kernels are things like $K(x) = a\exp(-x^2/b)$, $a/(b+x^2)$, $a\exp(-|x|/b)$, or even $a\delta(x)+b$. In general, a continuous (resp, differentiable) kernel is appropriate if the underlying function is expected to be continuous (resp,

differentiable). Not all differentiable, monotonic, symmetric functions are appropriate, as the $\mathbf{K}$ matrices should be positive definite; for instance, $K(x) = \{1 \text{ if } |x| < 1; \text{ else } 0\}$ produces indefinite matrices, and so is inappropriate.

Gaussian kernels, $a\exp(-x^2/b)$, can be motivated with a model where the true function is a mixture of sines and cosines of different frequencies. The standard modelling is to assume

$$f(t) = \sum_{j=1}^{M} w_{j,1}\sin(f_j t) + w_{j,2}\cos(f_j t)$$

with some $2M$ unknown Gaussian weights $w_{j,1}$ and $w_{j,2}$, and some randomly selected frequencies f_j. The frequencies are chosen from a normal distribution $N(0,\rho^2)$, setting a characteristic length scale on t of $1/\rho$. If $M \to \infty$ while the weights w decrease as $1/\sqrt{M}$ to compensate, the resulting model has a simple covariance function

$$\begin{aligned} K(x,y) &= \int_{-\infty}^{+\infty} \frac{\exp(-f^2/2\rho^2)}{\sqrt{2\pi}\rho}[\sin(fx)\sin(fy) + \cos(fx)\cos(fy)] \\ &= \exp(-\rho^2(x-y)^2/2). \end{aligned}$$

This pleasing model adapts naturally to the complex setting, where we replace sin and cos with complex exponentials:

$$f(t) = \sum_{j=1}^{M} w_j \exp(if_j t)$$

so that

$$C_{1,2}(x,y) = \int_{-\infty}^{+\infty} \frac{\exp(-f^2/2\rho^2)}{\sqrt{2\pi}\rho}[\sin(fx)\cos(fy) - \cos(fx)\sin(fy)] = 0$$

and, in fact, the real and imaginary parts are entirely uncorrelated. This is also immediate through the fact that our model of $f(t)$ has conjugacy symmetry, $f_j \to -f_j$. In fact, correlations will only appear if our frequencies f_j are chosen from an *asymmetric* distribution, rather than a normal distribution.

This asymmetric distribution happens naturally when computing Green's functions of excitations around an (approximate) ground state, as all frequencies $f_j \geq 0$. A few possible distributions of frequencies include:

1. A normal distribution constrained to be positive. In this case,

$$C_{1,2}(x) = \frac{2}{\sqrt{\pi}}F(x) = \exp(-x^2)\text{erfi}(x)$$

   using the Dawson F function or the imaginary error function $\text{erfi}(z) = -i\,\text{erf}(iz)$. Then $C_{2,2}(x) = \exp(-x^2)$.

2. A normal distribution constrained to be greater than f_0. Then $C_{1,2}(x) = \exp(-x^2)(i\,\text{erf}(f_0 - ix) - i\,\text{erf}(f_0 + ix))$.

3. A Cauchy disribution constrained to be positive. Then $C_{1,2}(x) = xG^{2,1}_{1,3}\left(\begin{matrix} 0 \\ 0, 0, -\frac{1}{2} \end{matrix}\middle| x^2\right)$, where G is the Meijer G function.

4. A distribution $p(f) = \frac{4f}{\pi(1+f^2)^2}$ on $f \geq 0$. Then $C_{1,2}(x) = x\exp(-x)$, which is simple to compute, but $C_{1,1}(x) = \frac{2}{\sqrt{\pi}} G^{2,1}_{1,3}\left(\begin{array}{c|c} 0 & \\ & \frac{x^2}{4} \\ 0, 1, \frac{1}{2} & \end{array} \right)$

5. A distribution $p(f) = \frac{f}{2\sqrt{\pi}} \exp\left(\frac{-f^2}{4}\right)$ on $f \geq 0$. Then $C_{1,2}(x) = x\exp(-x^2)$, but $C_{1,1}(x) = \frac{1-2xF(x)}{\sqrt{\pi}}$ again involving the Dawson F function.

6. An exponential distribution. Then $C_{1,2}(x) = \frac{x}{1+x^2}$, and $C_{1,1}(x) = \frac{1}{1+x^2}$.

All of these have the desired odd symmetry $K(x) = -K(-x)$. With the exception of the second one, all are unimodal on $x \geq 0$. The sixth option is the easiest to evaluate, but has very long tails for the correlations, which may be an unreasonable prior or could make it harder to approximate K by a sparse matrix. The first option reproduces the Gaussian correlation $K(x) = \exp(-x^2)$ on the real-real parts, which makes it attractive provided that an implementation of the F, erfi, or the complex erf function is available. All of these functions can of course be arbitrarily scaled in width and height, $K(x) = aK_0(x/b)$, to fit the variance and length scales appropriate to the problem.

In terms of fixing the magnitude on the covariance, e.g. the a parameter in $a\exp(-(x-y)^2/b^2)$, we know that $X(t)$ lies in the range $[-1, 1]$. If we take a uniform prior over this interval, we would have an variance of $1/3$. Interpreting this value too seriously is impossible, as we have already approximated $X(t)$ by a

Gaussian distribution. A choice of timescale b is inherently tied to the timescale of the dynamics of the underlying quantum system; if typical excitations have energy ΔE, then a choice of $b = \hbar/\Delta E$ is reasonable. Having some rough estimate of b is not a new requirement of this method, as it was already required that $G(t)$ should be sampled with a spacing $\ll b$.

We report the performance of ordinary kriging in the results section below.

5.5 Gaussian Processes as Classifiers

Approximating the marginals of $G(t)$ by multivariate Gaussians is drastic. Luckily, we can go beyond this. Gaussian processes have been frequently applied to *classification* problems, where an unknown probability $\pi(t) \in [0, 1]$ is to be estimated. $\pi(x)$ is reparameterized by $\pi(x) = \sigma(f(x))$, where σ is a *link function* (e.g. logistic or probit), and the *nuisance function* f is instead given the Gaussian process prior [105].

5.5.1 Real Classifiers

Given the earlier data points, computing the prediction $\mathbb{E}[\pi(x^*)] = \mathbb{E}_f[\sigma(f(x^*))]$ at a new point is intractable analytically, which has led to the development of several approximate computations. The maximum likelihood estimator (MLE)

simplifies computing $\mathbb{E}_f[\pi(x*)]$, an integral over all f, by instead computing the the *most likely likelihood* f. This is the $\hat{f}$ which most plausibly explains the observations under the Gaussian prior. The function $\hat{f}$ is represented by its estimated values $f(x_i)$ at the location of each sample x_i. At a new point to evaluate, these stored values can be used to compute a maximum likelihood $\hat{f}(x^*)$, and then $\sigma(\hat{f}(x^*))$ is interpreted as a prediction.

A next-order approximation is the Laplacian regression, which takes into account not only the maximum likelihood $\hat{f}$, but also the Hessian of the likelihood around $\hat{f}$. This Hessian is used to approximate the variance in $f(x_i)$; the true distribution of f conditioned on the samples is not Gaussian, but the second-order Taylor expansion lets us fit an approximation at the peak. The resulting Gaussian distribution over possible f's marginalizes to a Gaussian distribution on $f(x^*)$, which is integrated to approximate $\mathbb{E}[\pi(x^*)]$.

Finding the maximum likelihood estimator is typically done by gradient ascent or, more commonly, Newton iteration. In this context it is referred to as Iteratively Reweighted Least Squares (IRLS) [105], with the iteration defined by

$$\hat{f}_{\text{new}} = \hat{f} + (K^{-1} - \nabla_f^2 \pi(f))^{-1}(\nabla_f \pi(f) - K^{-1} f).$$

5.5.2 Complex Classifiers

We can use the methods above to estimate the real and imaginary parts separately, treating each as a Bernoulli variable $\frac{1+X(t)}{2}$ and $\frac{1+Y(t)}{2}$. While this works fine, this does discard the fact that $|G(t)| \leq 1$, and any correlations between real and imaginary parts at different times. We can repair this with a complex link function $\sigma_{\mathbb{C}} : \mathbb{C} \to D_1$, that will map a complex-valued link function $f(x) = f_R(x) + if_I(x)$ to the unit disk.

If our prior is to be phase-invariant, our link function should be rotationally symmetric, not giving higher prior to α than to some other $e^{i\theta}\alpha$. For example, this would be a problem with $\sigma_{\mathbb{C}}(f_R, f_I) = \text{logit}(f_R) + i\sqrt{1 - \text{logit}(f_R)^2}\,\text{logit}(f_I)$.

This means we can think of the link function as acting only on the magnitude of the complex vector $f_R + if_I$, leaving the phase unchanged: $\sigma_{\mathbb{C}}(f_R, f_I) = \frac{f_R+if_I}{|f_R+if_I|}(2\sigma(|f_R + if_I|) - 1)$. Then the magnitude link function σ can be a logit or probit model, such as $\sigma(x) = (1 + \exp(-x))^{-1}$. Given a $+1$ observation of the real part at x_i, this contributes to the log-likelihood with a term

$$L = \log\left(\frac{1 + \Re[\sigma_{\mathbb{C}}(x_i)]}{2}\right) = \log\left(\frac{1}{2} + \frac{f_R(x_i)}{|f_R(x_i) + if_I(x_i)|}\left(\sigma(|f_R(x_i) + if_I(x_i)|) - \frac{1}{2}\right)\right)$$

which produces a nonzero gradient (and Hessian) in both f_R and f_I. "-1" observations are identical but with the $\frac{1}{2}+$ replaced by $\frac{1}{2}-$, and measurements of f_I have the real and imaginary parts exchanged. With a logit link function, the

derivatives are all rational functions of f_R, f_I, $\sqrt{f_R^2 + f_I^2}$, and $\sigma(\sqrt{f_R^2 + f_I^2})$, a useful fact for fast numerics.

The likelihood contribution from the Gaussian prior on f_R and f_I is an unchanged K^{-1} from univariate regression, and Newton iteration can then be performed to find the maximum likelihood f. The Hessian $\nabla_f^2 \pi(f)$ was originally a diagonal matrix; it is now a 2x2 block diagonal, as it connects the real and imaginary f values at each point to each other. Otherwise, the IRLS iteration to find the maximum likelihood estimator is unchanged.

5.5.3 Complex Laplacian Approximation

With the maximum likelihood $\hat{f}$, the Laplacian approximation consists of finding the mean and covariance of f_R and f_I at a new point x^* to be predicted, and integrating that Gaussian distribution. The means are given by normal Gaussian prediction,

$$\mathbb{E}[f_R^*] = \hat{f}^\top K^{-1} \mathbf{k}_R(x^*), \quad \mathbb{E}[f_I^*] = \hat{f}^\top K^{-1} \mathbf{k}_I(x^*)$$

where K is the covariance matrix of the observations and $\mathbf{k}$ is the covariance vector between the observations and the new point.

In Rassmussen and Williams [105], equation (3.24) gives the variance in f^* in the univariate case. In our bivariate case where we need the 2x2 covariance matrix

of f_R and f_I, the computations are similar except that our covariance vectors $\mathbf{k}$ are replaced with 2xN matrices, one row for each prediction:

$$\mathrm{Var}[f^*_{R,I}] = k(x^*, x^*) - \mathbf{k}_{R,I}(x^*)^\top (K + W^{-1})^{-1} \mathbf{k}_{R,I}(x^*)$$

$$= \begin{bmatrix} C_{11}(x^*, x^*) & C_{11}(x^*, x^*) \\ C_{21}(x^*, x^*) & C_{22}(x^*, x^*) \end{bmatrix} - \begin{bmatrix} \mathbf{k}_R(x^*)^\top \\ \mathbf{k}_I(x^*)^\top \end{bmatrix} (K + W^{-1})^{-1} \begin{bmatrix} \mathbf{k}_R(x^*) & \mathbf{k}_I(x^*) \end{bmatrix}$$

Then the expected value of the true complex function at the point x^* is given by integrating over the possible $f_R(x^*)$ and $f_I(x^*)$ values, multiplied of their likelihood of coming from our estimated Gaussian distribution:

$$\mathbb{E}[z^*] = \int_{f_R, f_I \in \mathbb{R}^2} \sigma_{\mathbb{C}}(f_R, f_I)\ p\big((f_R, f_I) \sim \mathcal{N}(\mathbb{E}[f^*_{R,I}],\ \mathrm{Var}[f^*_{R,I}])\big)$$

With the maximum likelihood $\hat{f}$, the Laplacian approximation consists of finding the mean and covariance of f_R and f_I at a new point x^* to be predicted, and integrating that Gaussian distribution. The means are given by normal Gaussian prediction,

$$\mathbb{E}[f^*_R] = \hat{f}^\top K^{-1} \mathbf{k}_R(x^*), \quad \mathbb{E}[f^*_I] = \hat{f}^\top K^{-1} \mathbf{k}_I(x^*)$$

where K is the covariance matrix of the observations and $\mathbf{k}$ is the covariance vector between the observations and the new point.

In Rassmussen and Williams, equation (3.24) gives the variance in f^* in the univariate case. In our bivariate case where we need the 2x2 covariance matrix of

f_R and f_I, the computations are similar except that our covariance vectors $\mathbf{k}$ are replaced with 2xN matrices, one row for each prediction:

$$\text{Var}[f^*_{R,I}] = k(x^*, x^*) - \mathbf{k}_{R,I}(x^*)^\top (K + W^{-1})^{-1} \mathbf{k}_{R,I}(x^*)$$

$$= \begin{bmatrix} C_{11}(x^*, x^*) & C_{11}(x^*, x^*) \\ C_{21}(x^*, x^*) & C_{22}(x^*, x^*) \end{bmatrix} - \begin{bmatrix} \mathbf{k}_R(x^*)^\top \\ \mathbf{k}_I(x^*)^\top \end{bmatrix} (K + W^{-1})^{-1} \begin{bmatrix} \mathbf{k}_R(x^*) & \mathbf{k}_I(x^*) \end{bmatrix}$$

Then the expected value of the true complex function at the point x^* is given by integrating over the possible $f_R(x^*)$ and $f_I(x^*)$ values, multiplied of their likelihood of coming from our estimated Gaussian distribution:

$$\mathbb{E}[z^*] = \int_{f_R, f_I \in \mathbb{R}^2} \sigma_{\mathbb{C}}(f_R, f_I)\ p\big((f_R, f_I) \sim \mathcal{N}(\mathbb{E}[f^*_{R,I}],\ \text{Var}[f^*_{R,I}])\big)$$

5.6 Fourier Constrained Methods

Setting aside the Gaussian process approach, there is a different powerful prior we can put on G. In the L^∞ norm, $||G||_\infty = 1$. Its Fourier transform, $\hat{G}$, must then also be bounded by 1 in the dual L^1 norm. As mathematicians, this is a result from harmonic analysis; as physicists, this is time evolution $\exp(iHt)$ splitting $|\psi\rangle$ into several components with different frequencies and then adding them back together. For a finite quantum system, this means that $\hat{G}$ is a sum of finitely

many frequencies adding to 1:

$$\hat{G}(\omega) = \sum_k m_k \delta(\omega - \omega_k), \quad \sum_k |m_k| = 1.$$

Since our form of Green's functions in Eq. 5.1 also has $G(0) = 1$, we know that the m_k are purely real and positive. This is a significant advantage of focusing on Green's functions of inverse operators as in Eq. 5.1, as opposed to a more general form $\langle\psi| \exp(-iHt) U_2 \exp(iHt) U_1 |\psi\rangle$ with $U_2 U_1 \neq 1$.

Knowing that $\hat{G}$ must be of this form, we know $G(t)$ takes the form[2]

$$G(t) = \sum_k m_k \exp(i\omega_k t), \quad \sum m_k = 1, \quad 0 \leq m_k \leq 1 \tag{5.2}$$

The oscillatory nature of $\exp(i\omega t)$ makes this a highly nonlinear regression problem in m_k and ω_k, and any local search in those variables will invariably get trapped in bad local minima. There is also the issue that we do not know how many different values of k – how many different frequencies – are meaningfully contributing. A simple solution is to fix a large set of frequencies ω_k, and then regress only on the m_k. If the frequencies are dense compared to the largest sampling time, and the frequencies extend high enough to the largest excitations (or alternately, smallest timescales) we expect, then the reconstruction should be good. Fixing a minimum time scale $t_{\min}$ and latest sampling time $t_{\max}$, we take the set ω_k to range from $-1/t_{\min}$ to $+1/t_{\min}$, in equally spaced intervals of $1/t_{\max}$. Once we have restricted

[2] We note that the $m_k \leq 1$ constraint is, strictly speaking, redundant. In practice, many convex optimizers may benefit from having a restricted interval for each variable.

to this search space as our prior, this already imposes the smoothness constraints that we had in the Gaussian process prior.

With this set N_ω many frequencies fixed, the search space is reduced to m_k, which lie in an $(N_\omega - 1)$-dimensional simplex. The function $G(t)$ is a linear function of the coordinates, and the likelihood of a given observation is linear in $G(t)$. Unfortunately, integrating the full posterior over the simplex is intractable, so we are again led to approximation such as the maximum likelihood estimator. The MLE asks for the m_k that minimize the log-likelihood

$$L = \sum_{x_i \in \text{obs}} \log\left(\frac{1 \pm \mathcal{P}(G(x_i))}{2}\right) = \sum_{x_i \in \text{obs}} \log\left(\frac{1 \pm \sum_k m_k \text{TRIG}(i\omega_k x_i)}{2}\right),$$

where the sign in $\pm$ is chosen depending on the sign of the observation, the function $\mathcal{P}$ is either $\Re$ or $\Im$ part depending on the type of the observation x_i, and the TRIG function is correspondingly either sin or cos. This resulting function is nonlinear, but is convex, and can be optimized quite efficiently. Newton iteration performs poorly due to the linear inequality constraints, but general convex optimizers converge well. This optimization problem finds the maximum-likelihood estimator.

5.6.1 Uncertainty

A next-order improvement over the MLE is, again, incorporating a likelihood-weighted average over possible models for a better mean estimate. This gives

us two benenfits: one, we have a better estimator for the underlying model – including covariance in estimated points, which may be useful to carry forward to the final estimated $\tilde{G}(i\omega)$. Second, by knowing which points t have the greatest uncertainty $G(t)$, we can perform adaptive sampling. Due to the very nonlinear interaction between different samples, ideal sampling distributions may different significantly from uniform.

Once we have our MLE, it is straightforward to compute a Hessian of the likelihood function, $H = \nabla^2 L$. The $\sum m_k = 1$ constrained is easily dealt with by projecting this Hessian onto a subspace of one less dimension. The $0 \leq m_k$ constraints cause much greater difficulty however. Typically the vast majority of $m_{k,\mathrm{MLE}}$ variables are exactly equal to 0 at the MLE, and estimating the variation in these variables depends not only on the Hessian but on the gradient as well. Even among the active variables (those with $m_{k,\mathrm{MLE}} \neq 0$), the Hessian greatly overestimates the uncertainty: it will typically have eigenvalues $\lambda_i \ll 1$, suggesting directions in which the m_k can be adjusted by quantities $\lambda_i^{-1} \gg 1$ without appreciably reducing the likelihood. This is incorrect, as motion in this direction will 'hit' the $m_k \geq 0$ constraint much earlier. This problem does eventually disappear in the very high sampling limit, where the active variables are conclusively determined and the uncertainty is sufficiently small compared to the MLE values, $\|H^{-1}\|_2 < \min_{k:[m_k>0]}(m_k)$. In reality, this only occurs fast past the accuracy

thresholds that we care about, where $G(t)$ has been determined to good accuracy. This is because H is a very poorly conditioned matrix.

Using the Hessian $\nabla^2 L$ and gradient ∇L to build an approximation for the likelihood function is not a problem in itself. The approximation to the likelihood function by a Gaussian is generally good. The difficulty arises from the interaction between the Gaussian and the simplex region we allow. As we described, the constraints of the simplex cannot be neglected. The problem of integrating Gaussians (or equivalently, normal distributions) over simplices has been well studied [84, 55, 3], however these approaches largely focus on spaces of small dimension ([3, 55] only examine $N \leq 10$), and give prohibitively long runtimes for the cases of interest to us, $N \approx 1000$. At the same time, we do not need a high accuracy in our estimates: any sampling beyond the MLE will give us some modest improvements to our estimates. For this reason, the most viable algorithm seems to be Metropolis-Hastings based algorithms, which provide a simple and efficient sampling approach for log concave distributions. Metropolis-Adjusted Langevin Algorithm (MALA) [21] performed poorly, apparently due to the sharp constraints and discontinuous probability density, which is not reflected in the gradient-based Langevin dynamics. The Robust Adaptive Metropolis algorithm of [124] (implementation [10]) performed favorably.

5.7 Adaptive Sampling

Besides a better estimate of $G(t)$, one benefit an improved reconstruction algorithm is that after some initial sampling, we can choose more valuable points to sample subsequently – if we have information on the sensitivity of the reconstruction to additional samples.

5.7.1 Per-Point Adaptivity

For the simplest per-point reconstruction methods, adaptive reconstruction becomes very simple. For the simple frequentist estimate $\hat{G}(t) = \frac{n_+ - n_-}{n_{tot}}$, the uncertainty at a given point t should scale as $1/\sqrt{n_{tot}}$, so that we are led to sample each point equally. With a Bayes approach (or, alternately, a frequentist estimate of sample variance), there is less need to sample points where $|G(t)|$ is close to 1, as the observed variance is lower. This means we can focus more of our samples at the points with $G(t)$ closer to zero.

In practice, the benefit here is minimal. For a point t with k samples and a proportion x of positive samples (resp. $1 - x$ negative samples), then expected decrease in uncertainty with one additional sample scales as $k^{-3/2}\sqrt{x(1-x)}$, or that the number of samples should at a given point should scale as $\sqrt[3]{x(1-x)}$. This comes out to 25% more samples on a balanced point $G(t) \approx 0$, $x \approx 0.5$ than

an extreme point $G(t) \approx 0.7$, $x \approx 0.85$ – or a 10% reduction in overall samples needed by the algorithm, for only a very modest speedup.

The Gaussian Process setting offers more information than a strict per-point reconstruction. For one, we already have estimates of the uncertainty at any given point we choose to query, including those we have not yet sampled. As the Gaussian process is an inherently local process, however, it suffers from the same issue of only giving slightly more samples to a particular region before the uncertainty equilibrates. The extra sample efficiency of adaptive sampling was minimal, less than 30%, and we did not use it in our final evaluations.

5.7.2 Fourier Adaptivity

In the Fourier basis, adaptive measurements seemed more useful, as all observations interact nonlocally, and some positions could be considerably more or less useful than others. For a greedy adaptive measurement scheme, we can use mutual information between the distribution of m_k and a given sample as a figure of merit for picking the next point(s). To compute the mutual information for a real-value sampling at t, first we produce a series of S samples $\mathsf{m}_s \in \mathbb{R}^{N_k}$ from the Markov chain modeling the posterior distribution, and take their mean $\hat{m}$. Then

the mutual information I is calculated as,

$$H(p) = p\log(p) + (1-p)\log(p)$$

$$G(m;t) = \sum_{k}^{N_k} m_k \exp(i\omega_k t)$$

$$p_1(m;t) = \frac{1+\Re G(m;t)}{2}, \quad p_0(m;t) = \frac{1-\Re G(m;t)}{2}$$

$$\hat{p}_1(t) = p_1(\hat{m};t), \quad \hat{p}_0(t) = p_0(\hat{m};t)$$

$$I(t) = \frac{1}{S}\sum_{s} H(p_1(\mathsf{m}_s;t)) - p_0(\mathsf{m}_s;t))\log(\hat{p}_0(t)) - p_1(\mathsf{m}_s;t))\log(\hat{p}_1(t)) \quad (5.3)$$

For imaginary-part measurements, p_1 and p_0 use the imaginary part of G instead.

As sampling the Markov chain requires $\sim N_k^2$ samples, and I needs to be computed separately for each possible place to sample, this approach should not be used after each sample. The rough reconstruction for mutual information can use a much smaller (sparser) set of frequencies k. Once the mutual information is computed at many points t, this can be used for a large number of subsequent samples. The final reconstruction can use the full range of frequencies k.

5.8 Results

We tested our methods on spin-1 Heisenberg chain, with Hamiltonian $H = \sum_i \vec{S}_i \cdot \vec{S}_{i+1}$. Our system has length $L = 6$ and periodic boundary conditions. Our

state of interest is the unique ground state $|\psi_0\rangle$ of H, with energy $E = -8.617$ and a gap of $\Delta = E_2 - E = 0.721$. Our perturbation was an X flip at site 1, so that

$$G(t) = \langle\psi_0| \exp(iHt) X \exp(-iHT) X |\psi_0\rangle \, .$$

This function has 30 different frequency components, from a lowest frequency component of $f_{\min} = 2.156$ to a highest-frequency component of $f_{\max} = 11.865$. The 0.721 component does not appear as the $\psi_0 \to \psi_1$ transition is forbidden under a single X perturbation. This frequency information was not provided to any algorithms, and the Fourier algorithm used a frequency range of $\omega \in [0, 10]$ with steps of 0.2. This was chosen to provide a realistic guess for what a physicist who does *not* know the full spectrum may hope to search, and despite losing a small part (2×10^{-5}) of the spectrum, the Fourier reconstruction is able to perform well.

We evaluated each method by their ability to reconstruct a given interval of G from a given number of samples, with accuracy graded as L_2 error over the interval. For simple methods such as a linear interpolation, this includes the L_2 error at unsampled points, otherwise they could game the metric by only reconstructing their sampled points. As the Fourier method has the fact that $G(0) = 1$ fixed in, essentially giving it free samples, we focused on intervals that did *not* include $t = 0$, to give it a fair starting point. We initially tested each method at logarithmically

spaced sample counts from $S = 30$ up to 1800, reconstructing the interval $t \in [2, 42]$. The Gaussian Process method became unreasonably slow at $S = 1800$, and it was clear that the Fourier method was significantly outperforming, so we then continued comparing only the interpolation method and Fourier method up to $S = 51200$ points.

For the traditional method, we chose the theoretically optimal relation $N_{pts} = S^{1/5}$ to choose how many distinct points to reconstruct within our sample budget S. We observed a scaling of $\text{Err} = 3.11 \times S^{-0.351}$, with a confidence interval $[-0.361, -0.342]$ on the exponent.

For the Fourier method, we observed a scaling of $\text{Err} = 2.49 \times S^{-0.405}$, with a confidence interval $[-0.412, -0.397]$ on the exponent. For an average L_2 accuracy on this interval of 10^{-4}, this translates to a difference between 6×10^{12} samples with a linear interpolation, or 7×10^{10} samples with the Fourier method, nearly two orders of magnitude in sample complexity.

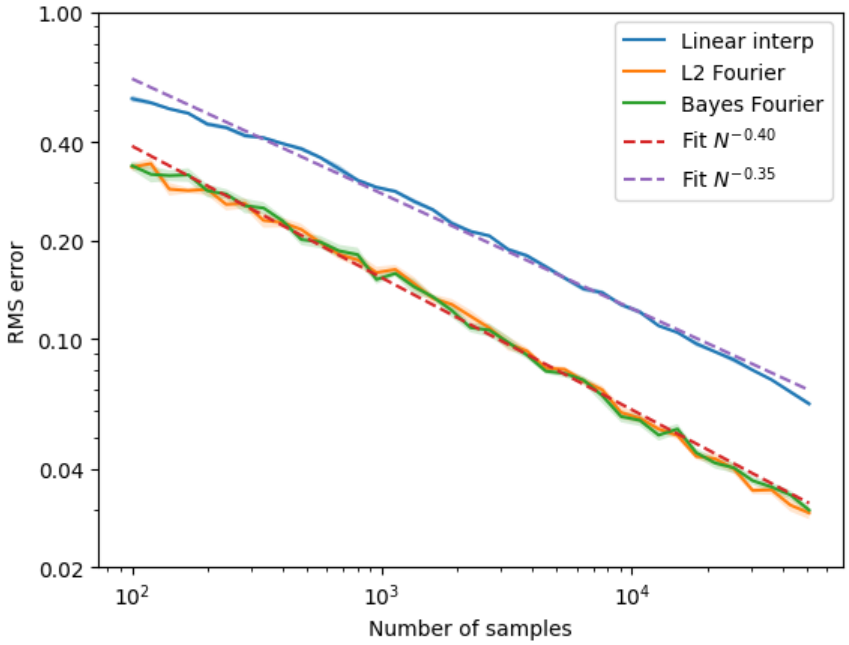

Figure 5.1: Comparison of three methods discussed in text, and best-fit polynomial scaling laws.

Chapter 6

Conclusion

This thesis reviewed four separate questions related to quantum information and optimization theory. Before we close the work out, it will be good to review once what work awaits.

6.1 Frustration-Free Hamiltonians

From the perspective of structural complexity theory, we examined the problem of whether local Hamiltonian terms frustrate one another. We were able to grow the list of candidate complexities from 4 to 7:

1. P: Classical, no proof, deterministic checks.

2. coRP: Classical, no proof, probabilistic checks.

3. NP: Classical, classical proof, deterministic checks.

4. MA: Classical, classical proof, probabilistic checks.

5. BQP_1: Quantum, no proof, probabilistic checks.

6. QCMA: Quantum, classical proof, probabilistic checks.

7. QMA_1: Quantum, quantum proof, probabilistic checks.

The most interesting question is whether this list is complete, or if we can put any sort of bounds on what other classes could occur. Given structural operations such as pairwise intersection, which are very hard to reason with without knowing the relationships between basic classes, it's likely that any proofs will require a long list of hypotheses, such as that $\mathsf{P}{=}\mathsf{coRP}{\neq}\mathsf{NP}{=}\mathsf{MA}{\neq}\mathsf{BQP}_1{\neq}\ \mathsf{QCMA}{=}\mathsf{QMA}$, a reasonably popular view of the world. A reasonable shorter goal could be classifying Hamiltonians on qubits of locality at most k, or arbitrary qudits of locality at most 2. As we proved in the chapter, merely limiting to qubits is not sufficient itself, as dimensionality can be reduced to 2 at the price of locality.

Another enticing work would be cleaner constructions of the cases shown in this work. A purely classical description of coRP could be useful in giving new complete problems for the class.

6.2 Permanents and Tomography

The main theorem was that PSD permanents, or quantum states, are hard to compute, even approximately. In particular, they are NP-hard within any subexponential factor. For a fixed rank d (or, a fixed dimension Hilbert space $\mathbb{C}^d$), they are slicewise polynomial - XP - meaning they can be computed in time $O(n^d)$. In a forthcoming work I will prove that they are in fact W[1]-hard to estimate, a type of hardness from parameterized complexity suggesting that $O(f(d)n^C)$ is impossible. This leaves open a question of whether $O(n^{\sqrt{d}})$ or similar should be possible.

The interplay between rank (or more generally, eigenvalues) of PSD matrices and they permanents is promising, with much work. In all likelihood, perturbing the difficult matrices constructed this work will also give lower bounds on difficulty for matrices of full rank but high spectral radius.

Regarding quantum state tomography, there are a variety of questions, such as average-case hardness, or what subsets of observations admit an efficient reconstruction algorithm. Separately there are questions about how this fits together with shadow tomography, and how much accuracy can improve with time-intensive postprocessing of sampled data.

6.3 Quasi-1D generalized Hartree-Fock

The generalized Hartree-Fock method, as we showed, can be efficiently applied to Gaussian Matrix Product States. Besides merely finding ground states, there could be room for exploring excitations, spectra, or thermal states in the same systems. By focusing on quasi-1D Hubbard models, we had a toy model of superconductivity in which we were able to reason and test our method's convergence – and we showed that finite-size effects stay relevant out to thousands of sites, especially in certain tunings of parameters. Testing and quantifying the finite-size effects in other classical systems could capture other interesting physics, as the consequences of finite-size effects are often poorly captured in theory with numerics as an important tool.

6.4 Green's Function Estimation

We showed that higher statistical reconstruction methods, such as Gaussian processes or sparse signal reconstruction, can improve the sample efficiency, both by a large constant factor and improving the asymptotic scaling. We also determined that adaptive sampling is of little use when using Gaussian Processes, or any other local reconstruction method, as the differences in uncertainty are quickly dominated by sampling density. Although the Gaussian process helped, it

was frustratingly slow due to the large covariance matrices involved. One major question is whether the Gaussian process could be accelerated to give a practical algorithm for large systems. This would likely involve specializing the linear algebra routines for banded matrices, or using the Toeplitz structure of the covariance matrices when points are uniformly spaced.

A satisfactory adaptive sampling algorithm was never possible for the Fourier reconstruction, because estimating the relevant marginals became costly. This is a question of estimating the marginals of a logconvex distribution, a problem which is receiveing steady interest and work today. It is likely that a better choice of Langevin damping, hierarchical sampling techniques, or some analytic methods could lead to usefully fast approaches for adaptive sampling. A useful measure of uncertainty and information is necessary as well, and at the moment it is not clear what the correct expression should be. Finally, it would be nice to have a theoretical analysis on the performance of the Fourier reconstruction: at the moment our results are entirely empirical, and analytical predictions of the convergence rate (perhaps as a function of the spectrum sparsity, frequency range, and maximum sampled time) would provide useful insight as to improving the algorithm further.

Bibliography

[1] Aaronson, S. [2007], 'The learnability of quantum states', *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **463**(2088), 3089–3114.
URL: *https://royalsocietypublishing.org/doi/abs/10.1098/rspa.2007.0113*

[2] Aaronson, S. [2018], Shadow tomography of quantum states, *in* 'Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing', STOC 2018, Association for Computing Machinery, New York, NY, USA, p. 325–338.
URL: *https://doi.org/10.1145/3188745.3188802*

[3] Adams, M. P. [2022], 'Integral, mean and covariance of the simplex-truncated multivariate normal distribution'.

[4] Affleck, I., Kennedy, T., Lieb, E. H. and Tasaki, H. [1987], 'Rigorous results on valence-bond ground states in antiferromagnets', *Phys. Rev. Lett.* **59**, 799–802.
URL: *https://link.aps.org/doi/10.1103/PhysRevLett.59.799*

[5] Aharonov, D. and Bredariol Grilo, A. [2019], Stoquastic pcp vs. randomness, *in* '2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)', pp. 1000–1023.
URL: *https://ieeexplore.ieee.org/abstract/document/8948612*

[6] Aharonov, D. and Naveh, T. [2002], 'Quantum np - a survey'.
URL: *https://arxiv.org/pdf/quant-ph/0210077.pdf*

[7] Allender, E., Bauland, M., Immerman, N., Schnoor, H. and Vollmer, H. [2009], 'The complexity of satisfiability problems: Refining schaefer's theorem', *Journal of Computer and System Sciences* **75**(4), 245 – 254.
URL: *http://www.sciencedirect.com/science/article/pii/S0022000008001141*

[8] Anari, N., Gurvits, L., Gharan, S. O. and Saberi, A. [2017], Simply exponential approximation of the permanent of positive semidefinite matrices, *in* '2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)', pp. 914–925.

[9] Andrei, E. Y. and MacDonald, A. H. [2020], 'Graphene bilayers with a twist', *Nature materials* **19**(12), 1265–1275.

[10] Anthoff, D. [2021], 'Robustadaptivemetropolissampler.jl v1.1.0', `https://github.com/anthofflab/RobustAdaptiveMetropolisSampler.jl/releases/tag/v1.1.0`.

[11] Bach, V., Lieb, E. H. and Solovej, J. P. [1994], 'Generalized hartree-fock theory and the hubbard model', *Journal of statistical physics* **76**(1), 3–89.

[12] Bardeen, J., Cooper, L. N. and Schrieffer, J. R. [1957*a*], 'Microscopic theory of superconductivity', *Physical Review* **106**(1), 162.

[13] Bardeen, J., Cooper, L. N. and Schrieffer, J. R. [1957*b*], 'Theory of superconductivity', *Physical review* **108**(5), 1175.

[14] Barto, L., Krokhin, A. and Willard, R. [2017], Polymorphisms, and How to Use Them, *in* A. Krokhin and S. Zivny, eds, 'The Constraint Satisfaction Problem: Complexity and Approximability', Vol. 7 of *Dagstuhl Follow-Ups*, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, pp. 1–44.
URL: *http://drops.dagstuhl.de/opus/volltexte/2017/6959*

[15] Barvinok, A. [2017], *Combinatorics and Complexity of Partition Functions*, 1st edn, Springer Publishing Company, Incorporated.

[16] Barvinok, A. [2020], 'A remark on approximating permanents of positive definite matrices'.

[17] Bauer, B., Wecker, D., Millis, A. J., Hastings, M. B. and Troyer, M. [2015], 'Hybrid quantum-classical approach to correlated materials'.

[18] Ben-Dor, A. and Halevi, S. [1993], 'Zero-one permanent is #p-complete, a simpler proof', *Proceedings of the 2nd Israel Symposium on the Theory and Computing Systems* .
URL: *http://people.csail.mit.edu/shaih/pubs/01perm.pdf*

[19] Bera, D. [2018], Amplitude amplification for operator identification and randomized classes, *in* L. Wang and D. Zhu, eds, 'Computing and Combinatorics', Springer International Publishing, Cham, pp. 579–591.

[20] Berthiaume, A. and Brassard, G. [1994], 'Oracle quantum computing', *Journal of Modern Optics* **41**(12), 2521–2535.
URL: *https://doi.org/10.1080/09500349414552351*

[21] Besag, J. [2004], An introduction to markov chain monte carlo methods, *in* M. Johnson, S. P. Khudanpur, M. Ostendorf and R. Rosenfeld, eds, 'Mathematical Foundations of Speech and Language Processing', Springer New York, New York, NY, pp. 247–270.

[22] Blume-Kohout, R. [2010], 'Optimal, reliable estimation of quantum states', **12**(4), 043034.
URL: *https://doi.org/10.1088/1367-2630/12/4/043034*

[23] Bogoliubov, N. N. [1958], 'New method in the theory of superconductivity. i', *Sov. Phys. - JETP (Engl. Transl.); (United States)* .
URL: *https://www.osti.gov/biblio/7340966*

[24] Botero, A. and Reznik, B. [2004], 'Bcs-like modewise entanglement of fermion gaussian states', *Physics Letters A* **331**(1), 39–44.
URL: *https://www.sciencedirect.com/science/article/pii/S0375960104011764*

[25] Bouland, A., Fitzsimons, J. F. and Koh, D. E. [2018], Complexity classification of conjugated clifford circuits, *in* 'Proceedings of the 33rd Computational Complexity Conference', CCC '18, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, DEU.

[26] Bravyi, S. [2004], 'Lagrangian representation for fermionic linear optics', *Preprint* .

[27] Bravyi, S. [2006], 'Efficient algorithm for a quantum analogue of 2-SAT', *arXiv e-prints* pp. quant–ph/0602108.

[28] Bravyi, S. and Gosset, D. [2017], 'Complexity of quantum impurity problems', *Communications in Mathematical Physics* **356**(2), 451–500.
URL: *https://doi.org/10.1007/s00220-017-2976-9*

[29] Bravyi, S., Gosset, D., König, R. and Temme, K. [2019], 'Approximation algorithms for quantum many-body problems', *Journal of Mathematical Physics* **60**(3), 032203.
URL: *https://doi.org/10.1063/1.5085428*

[30] Bravyi, S. and Terhal, B. [2010], 'Complexity of stoquastic frustration-free hamiltonians', *SIAM Journal on Computing* **39**(4), 1462–1485.
URL: *https://doi.org/10.1137/08072689X*

[31] Bridgeman, J. C. and Chubb, C. T. [2017], 'Hand-waving and interpretive dance: An introductory course on tensor networks', *J. Phys. A: Math. Theor.* **50**, 223001.

[32] Buntrock, G., Damm, C., Hertrampf, U. and Meinel, C. [1992], 'Structure and importance of logspace-mod class', *Mathematical systems theory* pp. 223–237.
URL: *https://doi.org/10.1007/BF01374526*

[33] Cancès, E. and Le Bris, C. [2000], 'Can we outperform the diis approach for electronic structure calculations?', *International Journal of Quantum Chemistry* **79**(2), 82–90.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/1097-461X%282000%2979%3A2%3C82%3A%3AAID-QUA3%3E3.0.CO%3B2-I*

[34] Chakhmakhchyan, L., Cerf, N. J. and Garcia-Patron, R. [2017], 'Quantum-inspired algorithm for estimating the permanent of positive semidefinite matrices', *Phys. Rev. A* **96**, 022329.
URL: *https://link.aps.org/doi/10.1103/PhysRevA.96.022329*

[35] Chen, X., Zeng, B., Gu, Z.-C., Yoshida, B. and Chuang, I. L. [2009], 'Gapped two-body hamiltonian whose unique ground state is universal for one-way quantum computation', *Phys. Rev. Lett.* **102**, 220501.
URL: *https://link.aps.org/doi/10.1103/PhysRevLett.102.220501*

[36] Clark, R. J., Lin, T., Brown, K. R. and Chuang, I. L. [2009], 'A two-dimensional lattice ion trap for quantum simulation', *Journal of Applied Physics* **105**(1), 013114.
URL: *https://doi.org/10.1063/1.3056227*

[37] Coffman, V., Kundu, J. and Wootters, W. K. [2000], 'Distributed entanglement', *Phys. Rev. A* **61**, 052306.
URL: *https://link.aps.org/doi/10.1103/PhysRevA.61.052306*

[38] Cohen, D. A., Cooper, M. C., Jeavons, P. G. and Živný, S. [2019], 'Binary constraint satisfaction problems defined by excluded topological minors', *Information and Computation* **264**, 12 – 31.
URL: *http://www.sciencedirect.com/science/article/pii/S0890540118301482*

[39] Cubitt, T. and Montanaro, A. [2016], 'Complexity classification of local hamiltonian problems', *SIAM Journal on Computing* **45**(2), 268–316.
URL: *https://doi.org/10.1137/140998287*

[40] D'Auria, V., Fornaro, S., Porzio, A., Solimeno, S., Olivares, S. and Paris, M. G. A. [2009], 'Full characterization of gaussian bipartite entangled states by a single homodyne detector', *Phys. Rev. Lett.* **102**, 020502.
URL: *https://link.aps.org/doi/10.1103/PhysRevLett.102.020502*

[41] De Gennes, P. G. and Pincus, P. A. [1999], *Superconductivity of Metals and Alloys*, CRC Press.

[42] Debnath, S., Linke, N. M., Wang, S.-T., Figgatt, C., Landsman, K. A., Duan, L.-M. and Monroe, C. [2018], 'Observation of hopping and blockade of bosons in a trapped ion spin chain', *Phys. Rev. Lett.* **120**, 073001.
URL: *https://link.aps.org/doi/10.1103/PhysRevLett.120.073001*

[43] Dichtel, K., Jelitto, R. J. and Koppe, H. [1971], 'The ground state of the neutral hubbard model', *Zeitschrift für Physik A Hadrons and nuclei* **246**(3), 248–260.
URL: *https://doi.org/10.1007/BF01395363*

[44] Dorit Aharonov, A. B. G. [2020], 'A combinatorial ma-complete problem'.
URL: *https://arxiv.org/abs/2003.13065*

[45] Dubail, J. and Read, N. [2015], 'Tensor network trial states for chiral topological phases in two dimensions and a no-go theorem in any dimension', *Phys. Rev. B* **92**, 205307.
URL: *https://link.aps.org/doi/10.1103/PhysRevB.92.205307*

[46] Eisert, J., Cramer, M. and Plenio, M. [2010], 'Area laws for the entanglement entropy - a review', *Reviews of Modern Physics* **82**, 277–306.

[47] Eldar, L. and Regev, O. [2008], Quantum sat for a qutrit-cinquit pair is qma1-complete, *in* 'Proceedings of the 35th International Colloquium on Automata, Languages and Programming - Volume Part I', ICALP '08, Springer-Verlag, Berlin, Heidelberg, p. 881–892.
URL: *https://doi.org/10.1007/978-3-540-70575-8*

[48] Essler, F. H. L., Frahm, H., Göhmann, F., Klümper, A. and Korepin, V. E. [2005], *The One-Dimensional Hubbard Model*, Cambridge University Press.

[49] Evenbly, G. and Vidal, G. [2010], 'Entanglement renormalization in noninteracting fermionic systems', *Physical Review B* **81**(23), 235102.

[50] Evenbly, G. and White, S. R. [2016], 'Entanglement renormalization and wavelets', *Phys. Rev. Lett.* **116**, 140403.
URL: *https://link.aps.org/doi/10.1103/PhysRevLett.116.140403*

[51] Fannes, M., Nachtergaele, B. and Werner, R. F. [1992], 'Finitely correlated states on quantum spin chains', *Communications in Mathematical Physics* **144**(3), 443–490.
URL: *https://doi.org/10.1007/BF02099178*

[52] Fishman, M. T. and White, S. R. [2015], 'Compression of correlation matrices and an efficient method for forming matrix product states of fermionic gaussian states', *Phys. Rev. B* **92**, 075132.
URL: *https://link.aps.org/doi/10.1103/PhysRevB.92.075132*

[53] Folland, G. B. [2001], 'How to integrate a polynomial over a sphere', *The American Mathematical Monthly* **108**(5), 446–448.
URL: *https://doi.org/10.1080/00029890.2001.11919774*

[54] Fujii, K., Kobayashi, H., Morimae, T., Nishimura, H., Tamate, S. and Tani, S. [2018], 'Impossibility of classically simulating one-clean-qubit model with multiplicative error', *Phys. Rev. Lett.* **120**, 200502.
URL: *https://link.aps.org/doi/10.1103/PhysRevLett.120.200502*

[55] Gessner, A., Kanjilal, O. and Hennig, P. [2019], 'Integrals over gaussians under linear domain constraints'.

[56] Gosset, D. and Nagaj, D. [2013], Quantum 3-sat is qma1-complete, *in* 'Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science', FOCS '13, IEEE Computer Society, USA, p. 756–765.
URL: *https://doi.org/10.1109/FOCS.2013.86*

[57] Grier, D. and Schaeffer, L. [2018], New hardness results for the permanent using linear optics, *in* 'Proceedings of the 33rd Computational Complexity Conference', CCC '18, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, DEU.

[58] Gurvits and Samorodnitsky [2002], 'A deterministic algorithm for approximating the mixed discriminant and mixed volume, and a combinatorial corollary', *Discrete & Computational Geometry* **27**(4), 531–550.
URL: *https://doi.org/10.1007/s00454-001-0083-2*

[59] Haegeman, J., Osborne, T. J., Verschelde, H. and Verstraete, F. [2013], 'Entanglement renormalization for quantum fields in real space', *Phys. Rev. Lett.* **110**, 100402.
URL: *https://link.aps.org/doi/10.1103/PhysRevLett.110.100402*

[60] Haegeman, J., Swingle, B., Walter, M., Cotler, J., Evenbly, G. and Scholz, V. B. [2018], 'Rigorous free-fermion entanglement renormalization from wavelet theory', *Phys. Rev. X* **8**, 011003.
URL: *https://link.aps.org/doi/10.1103/PhysRevX.8.011003*

[61] Harrow, A. W., Mehraban, S. and Soleimanifar, M. [2020], Classical algorithms, correlation decay, and complex zeros of partition functions of quantum many-body systems, *in* 'Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing', STOC 2020, Association for Computing Machinery, New York, NY, USA, p. 378–386.
URL: *https://doi.org/10.1145/3357713.3384322*

[62] Hartree, D. R. and Hartree, W. [1935], 'Self-consistent field, with exchange, for beryllium', *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences* **150**(869), 9–33.

[63] Hejazi, K., Chen, X. and Balents, L. [2021], 'Hybrid wannier chern bands in magic angle twisted bilayer graphene and the quantized anomalous hall effect', *Phys. Rev. Research* **3**, 013242.
URL: *https://link.aps.org/doi/10.1103/PhysRevResearch.3.013242*

[64] Hohenberg, P. and Kohn, W. [1964], 'Inhomogeneous electron gas', *Phys. Rev.* **136**, B864–B871.
URL: *https://link.aps.org/doi/10.1103/PhysRev.136.B864*

[65] Holzhey, C., Larsen, F. and Wilczek, F. [1994], 'Geometric and renormalized entropy in conformal field theory', *Nuclear Physics B* **424**(3), 443 – 467.

[66] Huang, H.-Y., Kueng, R. and Preskill, J. [2020], 'Predicting many properties of a quantum system from very few measurements', *Nature Physics* **16**(10), 1050–1057.
URL: *https://doi.org/10.1038/s41567-020-0932-7*

[67] Huang, Y. [2015], 'Computing energy density in one dimension', *arXiv preprint arXiv:1505.00772* .

[68] Huszár, F. and Houlsby, N. M. T. [2012], 'Adaptive bayesian quantum tomography', *Phys. Rev. A* **85**, 052120.
URL: *https://link.aps.org/doi/10.1103/PhysRevA.85.052120*
Řeháček et al.

[69] Řeháček, J., Hradil, Z. and Ježek, M. [2001], 'Iterative algorithm for reconstruction of entangled states', *Phys. Rev. A* **63**, 040303.
URL: *https://link.aps.org/doi/10.1103/PhysRevA.63.040303*

[70] Impagliazzo, R. and Paturi, R. [1999], Complexity of k-sat, *in* 'Proceedings. Fourteenth Annual IEEE Conference on Computational Complexity (Formerly: Structure in Complexity Theory Conference) (Cat.No.99CB36317)', pp. 237–240.

[71] Jahn, A., Gluza, M., Pastawski, F. and Eisert, J. [2019], 'Holography and criticality in matchgate tensor networks', *Science Advances* **5**(8), eaaw0092.

[72] Jerrum, M., Sinclair, A. and Vigoda, E. [2004], 'A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries', *J. ACM* **51**(4), 671–697.
URL: *https://doi.org/10.1145/1008731.1008738*

[73] Jones, R. O. [2015], 'Density functional theory: Its origins, rise to prominence, and future', *Rev. Mod. Phys.* **87**, 897–923.
URL: *https://link.aps.org/doi/10.1103/RevModPhys.87.897*

[74] Jordan, J., Orús, R., Vidal, G., Verstraete, F. and Cirac, J. I. [2008], 'Classical Simulation of Infinite-Size Quantum Lattice Systems in Two Spatial Dimensions', *Phys. Rev. Lett.* **101**(25), 250602.

[75] Jordan, S. P., Kobayashi, H., Nagaj, D. and Nishimura, H. [2012], 'Achieving perfect completeness in classical-witness quantum merlin-arthur proof systems', *Quantum Info. Comput.* **12**(5–6), 461–471.

[76] Kass, R. E. and Wasserman, L. [1996], 'The selection of prior distributions by formal rules', *Journal of the American Statistical Association* **91**(435), 1343–1370.
URL: *https://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10477003*

[77] Kempe, J., Kitaev, A. and Regev, O. [2006], 'The complexity of the local hamiltonian problem', *SIAM Journal on Computing* **35**(5), 1070–1097.
URL: *https://doi.org/10.1137/S0097539704445226*

[78] Kempe, J. and Regev, O. [2003], '3-local hamitonian is qma-complete', *Quantum Information and Computation* **3**(3), 258–264.
URL: *https://dl.acm.org/doi/10.5555/2011534.2011541*

[79] Kim, Y., Hong, K.-H., Kim, Y.-H. and Huh, J. [2020], 'Connection between bosonsampling with quantum and classical input states', *Opt. Express* **28**(5), 6929–6936.
URL: *http://www.osapublishing.org/oe/abstract.cfm?URI=oe-28-5-6929*

[80] Kirkpatrick, S., Györgyi, G., Tishby, N. and Troyansky, L. [1993], The statistical mechanics of k-satisfaction, *in* J. Cowan, G. Tesauro and J. Alspector, eds, 'Advances in Neural Information Processing Systems', Vol. 6, Morgan-Kaufmann.
URL: *https://papers.nips.cc/paper/1993/hash/a5cdd4aa0048b187f7182f1b9ce7a6a7-Abstract.html*

[81] Kitaev, A. Y., Shen, A. H. and Vyalyi, M. N. [2002], *Classical and quantum computation*, AMS.

[82] Knill, E. [2002], 'Quantum gates using linear optics and postselection', *Phys. Rev. A* **66**, 052306.
URL: *https://link.aps.org/doi/10.1103/PhysRevA.66.052306*

[83] Kohn, W. and Sham, L. J. [1965], 'Self-consistent equations including exchange and correlation effects', *Physical review* **140**(4A), A1133.

[84] Koyama, T. [2015], 'Holonomic gradient method for the probability content of a simplex region with a multivariate normal distribution'.

[85] Kraus, C. V. and Cirac, J. I. [2010], 'Generalized hartree–fock theory for interacting fermions in lattices: numerical methods', *New Journal of Physics* **12**(11), 113004.
URL: *https://doi.org/10.1088/1367-2630/12/11/113004*

[86] Kraus, C. V. and Osborne, T. J. [2013], 'Generalized hartree-fock theory for the dispersion relations of interacting fermionic lattice systems', *Phys. Rev. B* **88**, 195126.
URL: *https://link.aps.org/doi/10.1103/PhysRevB.88.195126*

[87] Kraus, C. V., Schuch, N., Verstraete, F. and Cirac, J. I. [2010], 'Fermionic projected entangled pair states', *Phys. Rev. A* **81**, 052338.
URL: *https://link.aps.org/doi/10.1103/PhysRevA.81.052338*

[88] Landau, Z., Vazirani, U. and Vidick, T. [2014], An efficient algorithm for finding the ground state of 1d gapped local hamiltonians, *in* 'Proceedings of the 5th conference on Innovations in theoretical computer science', pp. 301–302.

[89] Lieb, E. H. and Wu, F. Y. [1968], 'Absence of mott transition in an exact solution of the short-range, one-band model in one dimension', *Phys. Rev. Lett.* **20**, 1445–1448.
URL: *https://link.aps.org/doi/10.1103/PhysRevLett.20.1445*

[90] Lin, Lin and Wu, Xiaojie [2021], 'Numerical solution of large scale hartree-fock-bogoliubov equations', *ESAIM: M2AN* **55**(3), 763–787.
URL: *https://doi.org/10.1051/m2an/2020074*

[91] Liu, J., Sinclair, A. and Srivastava, P. [2019], 'The ising partition function: Zeros and deterministic approximation', *Journal of Statistical Physics* **174**(2), 287–315.
URL: *https://doi.org/10.1007/s10955-018-2199-2*

[92] Lvovsky, A. I. [2004], 'Iterative maximum-likelihood reconstruction in quantum homodyne tomography', **6**(6), S556–S559.
URL: *https://doi.org/10.1088/1464-4266/6/6/014*

[93] Maksimova, N. A., Garrison, L. H., Eisenstein, D. J., Hadzhiyska, B., Bose, S. and Satterthwaite, T. P. [2021], 'AbacusSummit: a massive set of high-accuracy, high-resolution N-body simulations', *Monthly Notices of the Royal Astronomical Society* **508**(3), 4017–4037.
URL: *https://doi.org/10.1093/mnras/stab2484*

[94] Marcus, M. [1963], 'The permanent analogue of the Hadamard determinant theorem', *Bulletin of the American Mathematical Society* **69**(4), 494 – 496.
URL: *http://doi.org/10.1090/s0002-9904-1963-10975-1*

[95] Motwani, R. and Raghavan, P. [1995], *Randomized Algorithms*, Cambridge University Press.

[96] Nishino, T. and Okunishi, K. [1998], 'A Density Matrix Algorithm for 3D Classical Models', *Journal of the Physical Society of Japan* **67**(9), 3066–3072.

[97] Nishino, T., Okunushi, K., Hieida, Y., Maeshima, N. and Akutsu, Y. [2000], 'Self-consistent tensor product variational approximation for 3D classical models', *Nucl. Phys. B* **575**, 504–512.

[98] Nishio, Y., Maeshima, N., Gendiar, A. and Nishino, T. [2004], 'Tensor Product Variational Formulation for Quantum Systems', *Preprint* .

[99] Orús, R. [2014*a*], 'Advances on tensor network theory: symmetries, fermions, entanglement, and holography', *The European Physical Journal B* **87**(11), 280.
URL: *https://doi.org/10.1140/epjb/e2014-50502-9*

[100] Orús, R. [2014*b*], 'A practical introduction to tensor networks: Matrix product states and projected entangled pair states', *Annals of Physics* **349**, 117 – 158.
URL: *http://www.sciencedirect.com/science/article/pii/S0003491614001596*

[101] Östlund, S. and Rommer, S. [1995], 'Thermodynamic Limit of Density Matrix Renormalization', *Phys. Rev. Lett.* **75**, 3537.

[102] Qi, B., Hou, Z., Li, L., Dong, D., Xiang, G. and Guo, G. [2013], 'Quantum state tomography via linear regression estimation', *Scientific Reports* **3**(1), 3496.
URL: *https://doi.org/10.1038/srep03496*

[103] Quek, Y., Fort, S. and Ng, H. K. [2021], 'Adaptive quantum state tomography with neural networks', *npj Quantum Information* **7**(1), 105.
URL: *https://doi.org/10.1038/s41534-021-00436-9*

[104] Rahimi-Keshari, S., Lund, A. P. and Ralph, T. C. [2015], 'What can quantum optics say about computational complexity theory?', *Phys. Rev. Lett.* **114**, 060501.
URL: *https://link.aps.org/doi/10.1103/PhysRevLett.114.060501*

[105] Rasmussen, C. E. and Williams, C. K. I. [2005], *Gaussian processes for machine learning*, Adaptive Computation and Machine Learning series, MIT Press, London, England.

[106] Ryser, H. J. [1963], *Combinatorial Mathematics*, Vol. 14, 1 edn, Mathematical Association of America.
URL: *http://www.jstor.org/stable/10.4169/j.ctt5hh8v6*

[107] Schaefer, T. J. [1978], The complexity of satisfiability problems, *in* 'Proceedings of the Tenth Annual ACM Symposium on Theory of Computing', STOC '78, Association for Computing Machinery, New York, NY, USA, p. 216–226.
URL: *https://doi.org/10.1145/800133.804350*

[108] Schuch, N. and Bauer, B. [2019], 'Matrix product state algorithms for gaussian fermionic states', *Phys. Rev. B* **100**, 245121.
URL: *https://link.aps.org/doi/10.1103/PhysRevB.100.245121*

[109] Schuch, N. and Verstraete, F. [2007], 'Computational complexity of interacting electrons and fundamental limitations of density functional theory', *arXiv preprint arXiv:0712.0483* .

[110] Schuch, N., Wolf, M. M. and Cirac, J. I. [2012], 'Gaussian matrix product states', *Preprint* .

[111] Sierra, G. and Martin-Delgado, M. A. [1998], 'The Density Matrix Renormalization Group, Quantum Groups and Conformal Field Theory', *Preprint* .

[112] Smith, A. W. R., Gray, J. and Kim, M. S. [2021], 'Efficient quantum state sample tomography with basis-dependent neural networks', *PRX Quantum* **2**, 020348.
URL: *https://link.aps.org/doi/10.1103/PRXQuantum.2.020348*

[113] Sterling, R. C., Rattanasonti, H., Weidt, S., Lake, K., Srinivasan, P., Webster, S. C., Kraft, M. and Hensinger, W. K. [2014], 'Fabrication and operation of a two-dimensional ion-trap lattice on a high-voltage microchip', *Nature Communications* **5**(1), 3637.
URL: *https://doi.org/10.1038/ncomms4637*

[114] Stockmeyer, L. [1983], The complexity of approximate counting, *in* 'Proceedings of the Fifteenth Annual ACM Symposium on Theory of Computing', STOC '83, Association for Computing Machinery, New York, NY, USA, p. 118–126.
URL: *https://dl.acm.org/doi/10.1145/800061.808740*

[115] Susskind, L. [2020], *The Power of One Clean Qubit*, Springer International Publishing, Cham, pp. 89–92.
URL: *https://doi.org/10.1007/978-3-030-45109-7*

[116] Tamma, V. and Laibacher, S. [2014], 'Multiboson correlation interferometry with multimode thermal sources', *Phys. Rev. A* **90**, 063836.
URL: *https://link.aps.org/doi/10.1103/PhysRevA.90.063836*

[117] Torlai, G., Mazzola, G., Carrasquilla, J., Troyer, M., Melko, R. and Carleo, G. [2018], 'Neural-network quantum state tomography', *Nature Physics*

14(5), 447–450.
URL: *https://doi.org/10.1038/s41567-018-0048-5*

[118] Valiant, L. [1979], 'The complexity of computing the permanent', *Theoretical Computer Science* **8**(2), 189–201.
URL: *https://www.sciencedirect.com/science/article/pii/0304397579900446*

[119] Van Acoleyen, K., Mariën, M. and Verstraete, F. [2013], 'Entanglement rates and area laws', *Phys. Rev. Lett.* **111**, 170501.
URL: *https://link.aps.org/doi/10.1103/PhysRevLett.111.170501*

[120] Verstraete, F. and Cirac, J. I. [2004], 'Renormalization algorithms for Quantum-Many Body Systems in two and higher dimensions', *Preprint* .

[121] Vidal, G. [2007], 'Entanglement Renormalization', *Phys. Rev. Lett.* **99**, 220405.

[122] Vidal, G. [2008], 'A class of Quantum Many-Body States That Can Be Efficiently Simulated', *Phys. Rev. Lett.* **101**(11), 110501.

[123] Vidal, G., Latorre, J. I., Rico, E. and Kitaev, A. [2003], 'Entanglement in quantum critical phenomena', *Phys. Rev. Lett.* **90**, 227902.
URL: *https://link.aps.org/doi/10.1103/PhysRevLett.90.227902*

[124] Vihola, M. [2012], 'Robust adaptive metropolis algorithm with coerced acceptance rate', *Statistics and Computing* **22**(5), 997–1008.
URL: *https://doi.org/10.1007/s11222-011-9269-5*

[125] White, S. R. [1992], 'Density matrix formulation for quantum renormalization groups', *Phys. Rev. Lett.* **69**, 2863.

[126] White, S. R. [1993], 'Density-matrix algorithms for quantum renormalization groups', *Phys. Rev. B* **48**, 10345–10356.
URL: *https://link.aps.org/doi/10.1103/PhysRevB.48.10345*

[127] White, S. R. [2005], 'Density matrix renormalization group algorithms with a single center site', *Phys. Rev. B* **72**, 180403.
URL: *https://link.aps.org/doi/10.1103/PhysRevB.72.180403*

[128] White, S. R. and Noack, R. M. [1992], 'Real-space quantum renormalization groups', *Phys. Rev. Lett.* **68**(24), 3487–3490.

[129] Wilson, D. B. [2002], 'On the critical exponents of random k-sat', *Random Structures & Algorithms* **21**(2), 182–195.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/rsa.10050*

[130] Yuan, C. and Parrilo, P. A. [2021], 'Maximizing products of linear forms, and the permanent of positive semidefinite matrices', *Mathematical Programming* .
URL: *https://doi.org/10.1007/s10107-021-01616-3*

[131] Zhang, W. [2004], 'Phase transitions and backbones of the asymmetric traveling salesman problem', *Journal of Artificial Intelligence Research* **21**, 471–497.

[132] Zhuk, D. [2017], 'A proof of csp dichotomy conjecture'.
URL: *http://ieee-focs.org/FOCS-2017-Papers/3464a331.pdf*

[133] Zvonkin, A. [1997], 'Matrix integrals and map enumeration: An accessible introduction', *Mathematical and Computer Modelling* **26**(8), 281–304.
URL: *https://www.sciencedirect.com/science/article/pii/S0895717797002100*

Appendix A

Reducibility of QCSPs to Qubits

Consider the following Hamiltonian on 4 qubits:

$$H_{4\to 2} = 1 - |\psi_1\rangle\langle\psi_1| - |\psi_2\rangle\langle\psi_2| - |\psi_3\rangle\langle\psi_3| - |\psi_4\rangle\langle\psi_4| \tag{A.1}$$

$$|\psi_1\rangle = \frac{1}{2}\left(\frac{3}{5}|0000\rangle - \frac{4}{5}|0001\rangle + |0100\rangle + |1010\rangle + \frac{8}{17}|1100\rangle + \frac{15}{17}|1111\rangle\right)$$

$$|\psi_2\rangle = \frac{1}{2}\left(\frac{4}{5}|0000\rangle + \frac{3}{5}|0001\rangle - |0110\rangle + |1001\rangle + \frac{20}{29}|1101\rangle + \frac{21}{29}|1110\rangle\right)$$

$$|\psi_3\rangle = \frac{1}{2}\left(\frac{5}{13}|0010\rangle + \frac{12}{13}|0011\rangle - |0111\rangle + |1000\rangle - \frac{21}{29}|1101\rangle + \frac{20}{29}|1110\rangle\right)$$

$$|\psi_4\rangle = \frac{1}{2}\left(\frac{-12}{13}|0010\rangle + \frac{5}{13}|0011\rangle - |0101\rangle + |1011\rangle - \frac{15}{17}|1100\rangle + \frac{8}{17}|1111\rangle\right)$$

Each $|\psi_i\rangle$ is orthonormal, so $H_{4\to 2}$ has a nullspace of dimension four. By inspecting the 840 distinct ways to apply two copies of $H_{4\to 2}$ to seven qubits, it can be checked that each sum will have a ground state above zero, except for the case where they are applied in the same way. This is a kind of "uniqueness" property that could be very loosely interpreted as monogamy for whole subspaces, instead of just one state.

By counting dimensions, one can check that this property is generic: it would hold almost always for any four random vectors. Unfortunately, for any simple and clean expressions one would write down, it would lack this property by one symmetry or another. This is why the simplest construction readily available, given above, is actually quite ugly.

Given a problem on 4-qudits, we can replace each 4-qudit with a collection of four qubits. Each clause is modified to act on the $|\psi_i\rangle$ basis of qubits instead

of $|i\rangle$ basis of the 4-qudits. Then, for each 4-qudit that the clause acted on, we add a copy of $H_{4\to 2}$. The above uniqueness property ensures that no other clause can act on the same qubits in any other order, or mix them with any other set of qubits.

Since each other clause acts on the same 4-dimensional induced subspace, this lets us rewrite any QCSP on $d = 4$ qudits into one on qubits. Together with the result in the text that 4-qudits suffice, we can Theorem 2.

Appendix B

Review of Gaussian States and Hartree-Fock

B.1 Majorana form of Hamiltonians

In standard creation and annihilation operators, a quartic Hamiltonian is written

$$\hat{\mathcal{H}} = \sum t_{ij}\hat{a}_i\hat{a}_j^\dagger + \sum u_{ijk\ell}\hat{a}_i\hat{a}_j\hat{a}_k^\dagger\hat{a}_\ell^\dagger + h.c. \tag{B.1}$$

This can be expanded in terms of Majorana operators,

$$\hat{a}_i \to \frac{1}{2}(\hat{c}_{2i-1} + i\hat{c}_{2i}) \tag{B.2}$$

$$\hat{a}_i^\dagger \to \frac{1}{2}(\hat{c}_{2i-1} - i\hat{c}_{2i}) \tag{B.3}$$

Then any term from t becomes

$$t_{ij}\hat{a}_i\hat{a}_j^\dagger = \frac{1}{4}t_{ij}(\hat{c}_{2i-1}\hat{c}_{2j-1} + i\hat{c}_{2i}\hat{c}_{2j-1} - i\hat{c}_{2i-1}\hat{c}_{2j} + \hat{c}_{2i}\hat{c}_{2j}) \tag{B.4}$$

Its Hermitian conjugate is

$$t_{ij}^*\hat{a}_j\hat{a}_i^\dagger = \frac{1}{4}t_{ij}^*(\hat{c}_{2j-1}\hat{c}_{2i-1} + i\hat{c}_{2j}\hat{c}_{2i-1} - i\hat{c}_{2j-1}\hat{c}_{2i} + \hat{c}_{2j}\hat{c}_{2i})$$

$$= \frac{1}{4}t_{ij}^*(-\hat{c}_{2i-1}\hat{c}_{2j-1} + i\hat{c}_{2i}\hat{c}_{2j-1} - i\hat{c}_{2i-1}\hat{c}_{2j} - \hat{c}_{2i}\hat{c}_{2j} + 4\delta_{i,j}$$

where $\{\hat{c}_i, \hat{c}_j\} = 2\delta_{ij}$ was used. The δ_{ij}'s that arise can be discarded, as they only add some constant shift to the Hamiltonian. Adding (B.4) to its Hermitian conjugate yields

$$\frac{t_{ij} + t_{ij}^*}{4}(i\hat{c}_{2i}\hat{c}_{2j-1} - i\hat{c}_{2i-1}\hat{c}_{2j}) + \frac{t_{ij} - t_{ij}^*}{4}(\hat{c}_{2i-1}\hat{c}_{2j-1} + \hat{c}_{2i}\hat{c}_{2j}) \tag{B.5}$$

$$= \frac{i\Re[t_{ij}]}{2}(\hat{c}_{2i}\hat{c}_{2j-1} - \hat{c}_{2i-1}\hat{c}_{2j}) + \frac{i\Im[t_{ij}]}{2}(\hat{c}_{2i-1}\hat{c}_{2j-1} + \hat{c}_{2i}\hat{c}_{2j}) \tag{B.6}$$

This shows that the resulting expansion has only imaginary coefficients on the quadratic terms. The anticommutation of the Majorana operators lets us rewrite any $\hat{c}_i\hat{c}_j \to \delta_{ij} + \frac{1}{2}(\hat{c}_i\hat{c}_j - \hat{c}_j\hat{c}_i)$. And then, again, we can neglect the δ_{ij} terms that lead to a constant offset in the Hamiltonian. In this way, the quadratic terms $t_{ij}\hat{a}_i\hat{a}_j^\dagger$ can always be transformed into

$$C + \sum_{ij} iT_{ij}\hat{c}_i\hat{c}_j \tag{B.7}$$

with T antisymmetric and real, and some constant shift C. If we apply the same substitution and expansion to the quartic terms

$$\sum_{ijk\ell} u_{ijk\ell}\hat{a}_i\hat{a}_j\hat{a}_k^\dagger\hat{a}_\ell^\dagger,$$

we will obtain additional constant terms (from e.g. $\hat{c}_1\hat{c}_2\hat{c}_1\hat{c}_2 = -1$), quadratic terms ($\hat{c}_1\hat{c}_2\hat{c}_3\hat{c}_2 = -\hat{c}_1\hat{c}_3$), and new quartic terms ($\hat{c}_i\hat{c}_j\hat{c}_k\hat{c}_\ell$, all indices distinct). When this is combined with its Hermitian conjugate, the terms combine and antisymmetrize as before, and the constant and quadratic terms are again of the form

$$C + \sum_{ij} iT'_{ij}\hat{c}_i\hat{c}_j$$

which can be absorbed into our other earlier quadratic term. The quartic terms remain where all indices are distinct, and the sum with the Hermitian conjugate is

$$U'_{ijk\ell}\hat{c}_i\hat{c}_j\hat{c}_k\hat{c}_\ell + U'^*_{ijk\ell}\hat{c}_\ell\hat{c}_k\hat{c}_j\hat{c}_i \tag{B.8}$$

$$= U'_{ijk\ell}\hat{c}_i\hat{c}_j\hat{c}_k\hat{c}_\ell + (-1)^6 U'^*_{ijk\ell}\hat{c}_i\hat{c}_j\hat{c}_k\hat{c}_\ell \tag{B.9}$$

$$= (U'_{ijk\ell} + U'^*_{ijk\ell})\hat{c}_i\hat{c}_j\hat{c}_k\hat{c}_\ell = U_{ijk\ell}\hat{c}_i\hat{c}_j\hat{c}_k\hat{c}_\ell \tag{B.10}$$

The $(-1)^6$ arises from the 6 swaps necessary to reorder the $\hat{c}$'s, and the resulting U is completely real. Because all four $\hat{c}$'s anticommute, this U can be taken completely antisymmetric.

In general, a term with k-fermion interactions can be written with a totally antisymmetric rank-k tensor. Since it will introduce k-choose-2 swaps when taking the Hermitian conjugate, it will be real when k is a multiple of 4, and completely imaginary otherwise.

B.2 Fermionic Parity

A Majorana operator $\hat{c}_i$ has a spectrum of ± 1. The $+1$ (resp -1) eigenspace corresponds to the 0 (resp 1) eigenspace of $\hat{a}_i^\dagger \hat{a}_i$, so

$$\prod_i \hat{c}_i = \prod_i \left(e^{i\pi \hat{n}_i}\right) \tag{B.11}$$

Then, since the number operators $\hat{n}_i$ all pairwise commute, they behave as c-numbers and

$$\prod_i \left(e^{i\pi \hat{n}_i}\right) = e^{i\pi \sum_i \hat{n}_i} = e^{i\pi \sum_i \hat{a}_i^\dagger \hat{a}_i} \tag{B.12}$$

which is the common form of the fermionic parity operator.

B.3 GFMPS techniques

The covariance matrix γ on the i'th site of a GFMPS is written as a block matrix

$$\gamma = \begin{bmatrix} \gamma_{pp} & \gamma_{p\ell} & \gamma_{pr} \\ \gamma_{\ell p} & \gamma_{\ell\ell} & \gamma_{\ell r} \\ \gamma_{rp} & \gamma_{r\ell} & \gamma_{rr} \end{bmatrix}, \tag{B.13}$$

where the subscript labels p, ℓ, r denote the physical, left auxiliary, and right auxiliary modes, respectively. The γ_{pp} block describes covariance with the p sites, while $\gamma_{p\ell}$ describes covariance between the p and ℓ sites. The contraction between two such tensors is most easily illustrated for two states where the modes are arranged in two groups each; the generalization to three or more groups is straightforward. Consider two covariance matrices given, in block-form, by

$$G = \begin{bmatrix} G_{aa} & G_{ac} \\ -G_{ca}^T & G_{cc} \end{bmatrix}, \quad H = \begin{bmatrix} H_{bb} & H_{bc'} \\ -H_{c'b}^T & H_{c'c'} \end{bmatrix} \tag{B.14}$$

with a common subsystem $c = c'$. They can be contracted into the ab covariance matrix

$$\begin{aligned} G \triangleright H = &\begin{bmatrix} G_{aa} & 0 \\ 0 & H_{bb} \end{bmatrix} + \\ &\begin{bmatrix} G_{ac} & 0 \\ 0 & H_{bc'} \end{bmatrix} \begin{bmatrix} G_{cc} & \mathbf{1} \\ -\mathbf{1} & H'_{c'c} \end{bmatrix}^{-1} \begin{bmatrix} G_{ac} & 0 \\ 0 & H_{bc} \end{bmatrix}^T \end{aligned} \tag{B.15}$$

The other crucial step in DMRG is the Schmidt decomposition, where a state is split into two blocks (two physical subsystems), as accurately as possible, given

the limited entanglement between the two. In a standard MPS, this is achieved by a singular value decomposition $|\psi\rangle = UDV^\dagger = \sum |\ell_k\rangle \lambda_k |r_k\rangle$; the smallest λ_k are discarded in order to meet the bond dimension limit. The analogous operation for fermionic Gaussian states was first derived by Botero and Reznik [24]. It proceeds by an SVD of the submatrix $\gamma_{\ell r}$, which contains the correlations between the two halves (but not their internal correlations):

$$O\gamma_{ab}Q^T = \bigoplus_k \begin{bmatrix} \mu_k & 0 \\ 0 & \mu_k \end{bmatrix} \tag{B.16}$$

This necessarily produces paired singular values μ_k and real orthogonal matrices O and Q. These are related to the symplectic eigenvalues λ_k of γ_{aa} by $\mu_k = \sqrt{1-\lambda_k^2}$. The Schmidt decomposition of γ is then given by

$$(O \oplus Q)\gamma(O \oplus Q)^T = \bigoplus_k \begin{bmatrix} 0 & \lambda_k & \mu_k & 0 \\ -\lambda_k & 0 & 0 & \mu_k \\ -\mu_k & 0 & 0 & -\lambda_k \\ 0 & -\mu_k & \lambda_k & 0 \end{bmatrix} \tag{B.17}$$

and can be split into

$$\gamma = L_{ac|ac} \triangleright R_{bc|bc} = \begin{bmatrix} 0 & O^T \\ -O & 0 \end{bmatrix} \triangleright \begin{bmatrix} 0 & Q \\ -Q^T & 0 \end{bmatrix} \tag{B.18}$$

The modes where $\lambda = 1$ are fully decoupled modes, and can be omitted from the c index to reduce the bond dimension without altering the underlying γ. Truncating the bond dimension is achieved by setting the largest several λ_k to 1 and discarding them, keeping only the modes with smaller λ_k (and thus higher entanglement).

The final ingredient to reconstruct most standard algorithms for MPS in the Gaussian context is a canonical form of the GFMPS. This can be constructed in an analogous fashion to standard MPS using the SVD decomposition described above; in the case of GFMPS, it turns out that the canonical form is essential in order to make the computation of the total energy as well as the local effective Hamiltonian in the DMRG iteration efficient. For technical details of this, we refer to Ref. [108]. With these tools in hand, one can perform the conventional single- and two-site DMRG algorithms to find the lowest-energy GFMPS for a given quadratic fermionic Hamiltonian.

Appendix C

GFMPS DMRG: Extended Pseudocode

Algorithm 2 Extracting blocks of Γ

```
function ExtractGamma(gfmps)
    Move gfmps to canonical form at block 0;
    Γ_result ← sparseZeros();
    carriedBlocks ← [];
    for i ← 1 to numBlocks do
        Add the Γ_ir to carriedBlocks;
        Move gfmps to canonical form at block i;
        for Γ_jr in carriedBlocks do
            if block Γ_ji is needed by T or U then
                Use Γ_jr to compute Γ_ji;
                Γ_result[j, i] ← Γ_ji;
                Move the 2nd index of Γ_jr from i − 1 to i;
            else
                Remove Γ_ji from carriedBlocks;
            end if
        end for
    end for
    Return Γ_result
end function
```

Algorithm 3 GFMPS DMRG

```
function GfmpsDmrg(F,gfmps)
    for s ← 1 to maxSweeps do
        Initialize effective potential H_0 to 0;
        for i ← 1 to numBlocks do
            Move gfmps to canonical form at block i;
            Update effective potential H_0 using F and block i of gfmps;
            Gaussian SVD to optimize block i;
        end for
        If |ΔE| < 10^−3, break;
    end for
end function
```

Appendix D

Quantum Phase Estimation

This explains how to estimate the real part of $G = \langle\psi|U|\psi\rangle$. The unitary U applied to $|\psi\rangle$ prepares a state $U|\psi\rangle = \alpha|\psi\rangle + \sqrt{1-|\alpha|^2}|\phi\rangle$, so that the desired value is α.

1. Prepare state $|\psi\rangle$
2. Prepare ancilla qubit $|0\rangle$ and apply Hadamard gate, $\frac{|0\rangle+|1\rangle}{\sqrt{2}}$
3. Apply controlled unitary U on $|\psi\rangle$, controlled by ancilla. The system is now in the state $\frac{1}{\sqrt{2}}\left(|\psi\rangle|0\rangle + \alpha|\psi\rangle|1\rangle + \sqrt{1-|\alpha|^2}|\phi\rangle|1\rangle\right)$
4. Apply Hadamard again on the ancilla and measure. We measure $|0\rangle$ with probability $|(1+\alpha)|^2/4 + (1-|\alpha|^2)/4 = \frac{1+\Re(\alpha)}{2}$.

This produces a Bernoulli measurement with the desired outcome. To measure the imaginary part of $\langle\psi|U|\psi\rangle$, simply measure the real part of $\langle\psi|iU|\psi\rangle$.

This procedure is in fact just the Quantum Phase Estimation algorithm applied to single-bit precision in phase.

Appendix E

Ordinary Kriging Equations

For a translationally invariant (or "stationary") prior, $C(x, y) = C(y, x) = C(|y - x|)$

Ordinary kriging allows the mean $\mu(x)$ to take a nonzero (Gaussian) prior. Then instead the weights are computed together with a Lagrange multiplier c, which enforces that $\sum w_i = 1$, so that that any prediction is a weighted mean of the samples. This contrasts with simple kriging, where a prediction far away from all the samples will regress to zero, even if all the samples are large and positive.

Given a kernel function $C(x, y)$, expressing the covariance between values $\Re G(x)$ and $\Re G(y)$, and a set of samples at points $x_1 \dots x_n$, the kernel matrix K is formed by

$$K = \begin{bmatrix} C(x_1, x_1) & \cdots & C(x_1, x_n) \\ \vdots & \ddots & \vdots \\ C(x_n, x_1) & \cdots & C(x_n, x_n) \end{bmatrix}.$$

For prediction at a new point x^*, the optimal weights $\mathbf{w}$ and constant c are given by

$$\begin{pmatrix} w_1 \\ \vdots \\ w_n \\ c \end{pmatrix} = \begin{bmatrix} K & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix}^{-1} \begin{pmatrix} C(x_1, x^*) \\ \vdots \\ C(x_n, x^*) \\ 1 \end{pmatrix}$$

for prediction

$$\tilde{G}(x^*) = c + \sum w_i z_i$$

where z_i are the previous samples.