Learning Theoretic Foundations for Understanding Quantum Systems

by

Allen Liu

B.S., Massachusetts Institute of Technology (2020)
S.M., Massachusetts Institute of Technology (2022)

Submitted to the Department of Electrical Engineering and Computer Science in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2025

Authored By: Allen Liu
Department of Electrical Engineering and Computer Science
May 16, 2025

Certified By: Ankur Moitra
Norbert Wiener Professor of Mathematics
Thesis Supervisor

Accepted By: Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee for Graduate Students

Learning Theoretic Foundations for Understanding Quantum Systems

by

Allen Liu

Submitted to the Department of Electrical Engineering and Computer Science
on May 16, 2025, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Understanding and harnessing the power of quantum systems has the potential to transform many domains in science and technology. However, before we can achieve these aspirations, we must first build a better understanding of how quantum systems fundamentally behave.

In this thesis, we approach this question through the lens of learning theory to develop new paradigms for learning about quantum systems and understanding their structural properties. We deliver several surprising results, upending previous beliefs about even fundamental laws and giving provably efficient algorithms for learning about quantum systems in settings previously conjectured to be intractable.

Typically in quantum many-body systems, the particles in the system interact locally with respect to some geometry as described by a local Hamiltonian. Two key questions are first, understanding equilibrium properties of a system with a given Hamiltonian and second, recovering the Hamiltonian from measurements of the properties of the system. For the first, we prove a universal law that there is a sudden death of entanglement, at a critical temperature depending only on the geometry but not on the system size. For the second, we give the first efficient algorithm for recovering the Hamiltonian at any temperature, breaking a conjectured barrier at low temperatures.

Beyond systems with local interactions, we also consider learning and testing properties of general quantum states, focusing on the interplay between statistical complexity and near-term quantum device constraints, only allowing for entangled measurements over a limited number of copies of the state. We characterize the optimal rates for learning and testing with single-copy measurements and for multi-copy measurements in many relevant near-term regimes.

Thesis Supervisor: Ankur Moitra
Title: Norbert Wiener Professor of Mathematics

Acknowledgments

First, I would like to thank my advisor, Ankur Moitra, who has been an incredible mentor to me ever since I approached him back when I was an undergrad. He has been invaluable in helping me both directly in our research projects and in broader, long-term research and career perspective aspects. I also thank Guy Bresler, Samuel Hopkins, and Jonathan Kelner for serving on my thesis committee.

I would also like to thank a few other mentors who have helped me immensely throughout my journey. I thank Jerry Li, who hosted me for two summers at Microsoft Research and exposed me to many new topics, including quantum information, the focus of this thesis, and helped me start many immensely successful collaborations. I also thank the mentors from my earlier years, Alex Iosevich who gave me my first exposure to research back in high school, and Zuming Feng and Po-Shen Loh for their guidance during my math Olympiad days.

I would also like to thank all of the amazing coauthors whom I have had the fortune to work with during my time in graduate school: Anders Aamand, Sara Ahmadian, Ainesh Bakshi, John Bostanci, Clément Canonne, Justin Chen, Sitan Chen, Zeev Dvir, Guru Guruganesh, Samuel Hopkins, Brice Huang, Jonathan Kelner, William Kretschmer, Zeph Landau, Shyam Narayanan, Ryan O'Donnell, Renato Paes Leme, Martin Pal, Binghui Peng, Jon Schneider, Mark Sellke, Aaron Sidford, Sandeep Silwal, Balasubramanian Sivan, Pattara Sukprasert, Ewin Tang, Kevin Tian, Ali Vakilian, Joshua Wang, Morris Yau, Morteza Zadimoghaddam, Fred Zhang. Through these collaborations, I have learned and grown immensely as a researcher and made new friendships that extend far beyond research.

I would also like to thank the many other friends I have made along the way. The MIT Theory Group and broader MIT community has been an incredibly vibrant and welcoming community and has been an essential part of my journey for the past five years. Shout-out to my office mates Justin Chen, Maxwell Fishelson, Noah Golowich, Vardis Kandiros, Lisa Yang, and Wei Zhang for being a constant, welcoming presence.

I also thank Meghal Gupta, Brice Huang, Yang Liu, Mehtaab Sawhney, and Mark Sellke for the conversations and games that greatly helped my morale for my first year during COVID times.

I would like to thank the NSF Graduate Research Fellowship Program, Hertz Foundation Fellowship, and Citadel GQS Fellowship for generously supporting my graduate studies.

Finally, and most importantly, I would like to thank my parents Beilei and Chuheng for their unwavering support.

Contents

List of Figures

Chapter 1

Introduction

Understanding and harnessing the power of quantum systems has the potential to transform many domains in science and technology, from enabling breakthroughs in computational problems such as factoring large numbers, to guiding the design of materials with extraordinary properties like superconductivity and superfluidity. However, before we can achieve these aspirations, we must first build a better understanding of how quantum systems fundamentally behave. While quantum mechanics provides a foundation for describing the behavior of atomic-scale particles, relating these atomic-scale interactions to emergent physical properties remains largely a mystery. Despite growing interest and years of research efforts, quintessential phenomena such as quantum entanglement are still poorly understood.

In this thesis, we will focus on the following guiding questions which are central to understanding the properties of quantum systems. First:

Guiding Question 1. Given a description of the interactions between particles in a quantum system, what can we say about its physical properties?

However, for systems that we observe in a laboratory or in nature, we often do not know a description of the underlying interactions. Thus, it is also crucial to understand when we can recover information about these interactions from observing the system.

Usually, the way that an observer can interact with a quantum system is through measurement *. This leads to our second question, which is a natural counterpart to the first.

Guiding Question 2. Given measurements of the physical properties of a quantum system, can we recover a description of the system?

We approach these questions theoretically, from a statistical and computational perspective. In general, quantum mechanics gives us mathematical models of quantum systems and basic laws about their behavior. Our focus will be on deriving from these laws rigorous predictions of their properties and designing algorithms with provable guarantees for recovering their parameters. While theoretical in nature, the answers to these questions will inform our understanding of the limits of what we can hope to know about quantum systems and what we can use them for.

Much of the early work on these questions came from the perspective of quantum physics and quantum information. However, these questions have recently garnered increased interest from other communities, especially within computer science. In this thesis, we will show how a novel approach, inspired by theoretical computer science and especially computational learning theory, enables us to develop new ways of understanding complex quantum phenomena such as entanglement. We deliver several surprising results that uncover fundamental laws and inform the tractability and limitations of potential applications of quantum technologies.

Remark 1.0.1 (On Measurements)**.** At a high-level, measuring a system results in a (random) classical outcome that reveals a small amount of information about the underlying system, analogous to sampling from a classical distribution. Quantum mechanics stipulates that measurement destroys the system. Thus, in all settings where we study recovery from measurements, we will assume access to multiple copies of the system in question. This is generally obtained by, say, repeating the experimental setup many times.

*A formal definition of measurement is deferred to Definition 1.4.3

1.1 Local Hamiltonians

Local Hamiltonians are the canonical, ubiquitous model for quantum many-body systems across condensed-matter physics, quantum chemistry, and quantum information [155, 136, 168, 82]. We give a very brief, high-level introduction here and defer a more formal treatment to Section 1.4.2.

Typically, in quantum many-body systems, the particles interact locally with respect to some geometry, meaning that particles interact more strongly with those nearby. Their interactions are described by a local Hamiltonian, which is a sum of the local interactions among different subsets of particles.

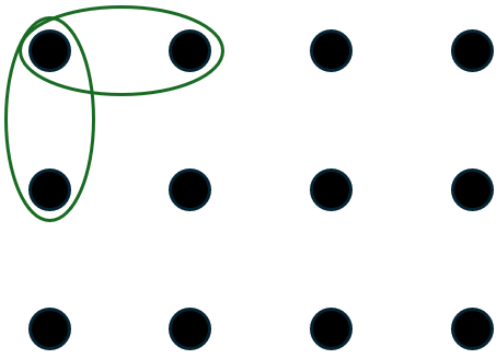

Figure 1-1: Local Hamiltonian where the interaction terms E_a range over edges between adjacent points in the grid. The black dots represent particles and the green edges represent interactions between pairs of particles (only two are shown for simplicity).

Throughout this thesis, we use $\mathfrak{n}$ to denote the number of particles. For each subset of interacting particles, there is a term E_a that describes their interaction. The local Hamiltonian is then given by

$$H = \sum_{a \in \mathcal{A}} E_a$$

where the sum ranges over all subsets of interacting particles. H is represented as a $2^{\mathfrak{n}} \times 2^{\mathfrak{n}}$ matrix. For a given Hamiltonian, the Gibbs state describes the behavior of the system at thermal equilibrium, at a specified inverse temperature $\beta > 0$ [†]. The

[†] β is proportional to $1/T$ where T is the actual temperature.

Gibbs state of the system is given by the matrix [‡]

$$\rho_\beta = \frac{e^{-\beta H}}{\mathrm{tr}(e^{-\beta H})}\,. \tag{1.1}$$

With this model, we can restate our main questions in the context of quantum many-body systems:

1. Given $H = \sum_{a\in\mathcal{A}} E_a$, for various β, what can we say about the properties of ρ_β?

2. Given measurements of (copies of) ρ_β, what can we recover about the E_a?

1.1.1 From Microscopic to Macroscopic

The above questions fall under the broad paradigm of understanding how microscopic interactions (the local Hamiltonian) drive emergent macroscopic properties (the Gibbs state) in complex systems. This has been a driving theme in statistical physics, a rich area that has seen many notable successes throughout decades of research, including Parisi's work that was awarded the Nobel Prize in Physics in 2021 [202]. Of course, the same guiding principles are also useful in many other domains ranging from machine learning to biology [4, 161, 188]. Statistical physics has both been a grounds for developing rich and beautiful mathematics and played an important role in helping us understand physical phenomena such as magnetization and surface tension. For example, understanding phase transitions in the Ising model has helped us understand how interactions between atomic-scale magnetic dipoles drive spontaneous magnetization across a material when it is cooled below a critical temperature [172].

[‡]Note that this is a valid quantum state because it is a positive semidefinite matrix with trace 1.

Low temperature (large β) High temperature (small β)

Figure 1-2: The Ising model is in many ways the foundational model in statistical physics. In this example studied in the seminal work of Kramers and Wannier [143] and later Onsager [172], we imagine $\mathfrak{n}$ magnets arranged in a two-dimensional grid. The magnets can either point up or down, represented by the corresponding variable being $+1$ or -1. The Ising model defines a distribution over $\{-1, 1\}^{\mathfrak{n}}$ where the probability of a configuration is proportional to $\exp(-\beta \sum_{i,j \text{ adjacent}} -x_i x_j)$ where β is the inverse temperature. Onsager's work demonstrates a phase transition from a disordered phase, where a typical configuration looks random, to an ordered phase, where a typical configuration has a large majority of magnets pointing in the same direction, as the temperature decreases below a critical threshold.

However, many modern challenges in physics and technology such as superconductivity and superfluidity revolve around understanding quantum phenomena. Compared to classical systems, we have a much more limited understanding in the quantum realm, where additional complexities such as quantum entanglement and non-commutativity pose new challenges not present in classical settings. Nevertheless, we hope that the same guiding principles from statistical physics may help us. In particular, can we use statistical models to understand how quantum phenomena arise in complex systems of particles?

One of the main themes of this thesis will be developing new tools for taming the complexities of quantum systems. The work presented here draws upon our understanding of phase transitions and other properties of classical statistical physics models — the definition in (1.1) generalizes classical Markov random fields which have been extensively studied in statistics, machine learning, and many other domains [186, 49, 71]. However, to deal with the new challenges in the quantum setting, we take inspiration from unexpected sources — modern techniques from a variety of domains

in theoretical computer science, such as counting and sampling, approximation theory, and optimization, will play a key role.

1.1.2 Important Quantum Properties

Returning to Guiding Question 1, we first specify the key properties that we want to understand. Entanglement is the quintessential quantum property [190, 43, 46]. As Schrödinger argues in [190],

> I would not call that [entanglement] *one*, but rather *the* characteristic trait of quantum mechanics.

Entanglement is what distinguishes quantum systems from classical ones and is the essence of what makes quantum systems interesting for scientific and engineering applications. Yet, entanglement remains a mysterious concept, and we generally lack a sound theoretical understanding of entanglement in quantum many-body systems. In fact, despite extensive study, we even lack clean, well-accepted definitions of entanglement in many settings [199, 180]. A crucial step towards understanding the properties of quantum systems will be understanding when they can exhibit entanglement.

The Role of Temperature In classical statistical physics, the (inverse) temperature parameter, given by β in (1.1), controls phase transitions and the emergence of important structural properties. Adjusting temperature shifts the balance among microscopic configurations — favoring disorder at high temperatures and order at low temperatures. Phase transitions occur at a critical threshold temperature, when the system undergoes an abrupt, collective transformation in its macroscopic properties (see Figure 1-2). This phase transition in statistical models underlies real-world phenomena such as the boiling of water and the spontaneous magnetization of ferromagnets [178, 217].

Naturally, this motivates us to study the role that temperature plays in the emergence

of quantum properties, namely entanglement. Specifically, we ask:

Question 1. How does temperature affect entanglement in quantum many-body systems. Specifically, at what values of β can the Gibbs state ρ_β be entangled?

The study of the relationship between temperature and entanglement dates back to the start of modern quantum information [179, 22, 173]. However, the wealth of ideas generated since then has yielded surprisingly little insight into this basic question. Here, we find new insight in a departure from previous approaches. We will take inspiration from classical algorithms for approximate counting and sampling to deliver a striking new law.

1.1.3 Algorithmic Challenges for Learning

The previous thread broadly concerns Guiding Question 1, but the counterpart, Guiding Question 2, is equally important. Namely, can we take measurements of the macroscopic properties of a quantum system and recover information about the microscopic interactions driving it? This question falls under the umbrella of *Hamiltonian learning*, which has become a cornerstone in the design, engineering, and benchmarking of quantum devices.

For example, in an attempt to understand phenomena like topological order and superconductivity that are important in condensed matter physics, experimentalists carefully design systems which exhibit these exotic behaviors. To achieve this, analog quantum simulators are tuned to obey poorly-understood Hamiltonians like the Fermi–Hubbard model for experimental exploration [95, 78, 115]. For these experiments, a natural goal is to learn the interactions which give rise to various phenomena [212, 140]. Intractability of computation is a major barrier to resolving open problems like finding the phase diagram of various canonical families of Hamiltonians, so having better algorithmic tools is of key importance in this domain [149]. Hamiltonian learning also arises when engineering quantum systems: a major challenge in building near-term quantum devices is being able to validate them—certify that they implement the

desired Hamiltonian—and understand sources of error [73, 193]. Quantum devices with 100 or more qubits are challenging to simulate classically, but quantum Hamiltonian learning has emerged as an alternative strategy for benchmarking devices by combining quantum resources and classical learning techniques [57, 97].

The Low Temperature Regime For Hamiltonian learning, the temperature parameter β will again play a crucial role. Despite its importance, previous algorithms for Hamiltonian learning all had crucial limitations. For general Hamiltonians, previous algorithms were efficient, scaling polynomially with $\mathfrak{n}$, only at high temperature i.e. for $\beta < c$ for some small constant c. For larger β, these algorithms required running time exponential in $\mathfrak{n}$, making them impractical beyond small systems. In fact, due to the lack of success and even viable approaches, Hamiltonian learning was conjectured to be intractable beyond this high temperature regime. For instance, in a recent survey on the complexity of learning quantum states [13], the authors ask

Conjecture ([13])**.** *Could low temperature Gibbs states be pseudorandom, which would explain the difficulty in finding a time efficient algorithm?*

This leaves us at an impasse. Quantum phenomena are most prominent at zero or near-zero temperature [9], precisely where approaches such as high-temperature series expansions fail [149]. In fact, many of these approaches work precisely because complex quantum phenomena like long-range entanglement do not arise at high temperatures. Given the seeming intractability at low temperatures, one might wonder whether quantum Hamiltonian learning, as a framework for benchmarking quantum devices or learning about the underlying physics of quantum systems, is inherently a dead end.

Nevertheless, we return to the central question, which has been posed as an open question throughout the community [15, 103, 7, 13], and examine it in a new light.

Question 2. Can we give efficient algorithms for Hamiltonian learning from Gibbs states at low temperatures?

We will show a surprising resolution to this question, upending prevailing beliefs. We

rely on modern tools in learning and optimization such as polynomial approximations and semidefinite programming hierarchies to tame the complexities of low-temperature systems and give an efficient algorithm, solving a problem previously conjectured to be impossible.

1.2 General Quantum States

Local Hamiltonians model many natural systems where the particles in the system interact locally. Still, there are many settings where we need to work with general quantum systems where the particles may exhibit arbitrary, non-local interactions. Of course, general systems, by definition, can exhibit arbitrary behavior, so here we will focus on the recovery problem — recovering a description of the system from measurements.

We model the state of the system as described by an unknown density matrix $\rho \in \mathbb{C}^{d \times d}$. Here, for a general system of $\mathfrak{n}$ particles, the dimension of the state is $d = 2^{\mathfrak{n}}$. We have copies of ρ which we can measure, and our goal is to recover a description of ρ. This is exactly the task of *quantum state tomography*. Tomography is, in many ways, the most general and basic quantum learning task.

While compared to local Hamiltonians, the exponential scaling in the number of particles seems to limit the paradigm of state tomography to relatively small systems, it nonetheless plays a key role in verification of current quantum technologies [34], protocols such as entanglement detection [118, 98], and various other applications. For some concrete examples, in [158], tomography was used to verify that a 2×2 state was correctly teleported over 143 km in a teleportation experiment and in [104], tomography was used to verify that an experimental device they designed was generating the correct 8-particle (i.e. 256×256) W-state, an important resource in quantum communication.

In most applications of tomography, each copy of ρ is a resource that may be expensive to obtain, so the main figure of merit is the copy complexity i.e. the number of copies

of ρ that we need to measure. This is a quantum generalization of the usual notion of sample complexity from statistics.

1.2.1 Near-Term Hardware Constraints

Recently, the copy complexity of quantum state tomography was settled by [170, 102] who showed that $n = \Theta(d^2/\varepsilon^2)$ copies are necessary and sufficient to learn ρ to ε accuracy in trace distance. However, a major downside of these protocols is that they require heavily entangled measurements over the joint state $\rho^{\otimes n}$. This poses a number of challenges to porting these algorithms into practical settings. First, the descriptions of the measurements are quite large (as the overall joint state is of size $d^n \times d^n$), and cannot be implemented on current (or near-term) quantum devices. Second, the measurements require that all n copies of ρ are simultaneously present. In many realistic settings, such as streaming settings where one copy of ρ is given to the algorithm at a time, this would require that the quantum device be able to store all of these copies in persistent quantum memory. Such a task is also well out of reach for current or near-term quantum devices, in essentially any non-trivial regime of the parameters.

An appealing class of algorithms which avoids both of these issues, and which can be implemented on even current devices, are algorithms which only rely on single-copy (i.e. unentangled) measurements. In contrast to general protocols which may manipulate the joint state over all n copies $\rho^{\otimes n}$, these algorithms only apply measurements to one copy of ρ at a time. Consequently, these measurements are performed on much smaller states, and moreover, can be performed without any quantum memory.

This single-copy setting, being a natural and practical formulation, has seen growing interest for both quantum state tomography [72, 86, 144, 156] and related tasks such as testing and certification [54, 64, 66, 124, 5]. Still, before the work presented here, even the most basic question remained open. Namely,

Question 3. What is the optimal copy complexity for single-copy quantum state

tomography?

This question will be our starting point for quantum state tomography. The best known upper bound for this question was $n = O(d^3/\varepsilon^2)$ [144]. However, the best known lower bound was only $\Omega(d^2/\varepsilon^2)$, the same rate as in the unconstrained multi-copy setting. The protocol in [144] essentially works by measuring all of the copies of ρ randomly. However, an important feature in single-copy tomography is that the measurements may be adaptively chosen based on the outcomes of the previous measurements. This leaves the possibility that we can design better algorithms that leverage adaptively chosen measurements.

Understanding the power of adaptivity will be a crucial component in answering the above question. The *adaptivity* in the quantum setting is a significant contrast compared to classical distribution learning settings, where the samples the learner receives are usually drawn independently. This added wrinkle of adaptivity is perhaps the main reason why even some of the most basic questions about learning with single-copy measurements remained open until very recently. To resolve these questions, we will need to develop new frameworks for reasoning about the power of adaptive learning algorithms — this is especially notable from a lower bounds perspective, as existing techniques often heavily rely on the learner's observations being independent.

1.2.2 Gracefully Scaling Up

While the single-copy setting is a useful formalism for designing algorithms for current quantum hardware, as the computational power of our quantum devices continues to grow, we may soon go beyond the limitations posited in the single-copy formalism. It is timely not only to understand the fundamental limitations of current devices, but also to map out avenues for gracefully scaling up our near-term algorithms as the platforms on which we run these algorithms mature.

As a motivating example, an experimenter could simultaneously prepare t copies of the state for some integer $t > 1$ by simply replicating the experimental setup

t times. Moreover, some existing devices can already store multiple (albeit few) copies of a quantum state at a time [23, 126, 153]. For some tasks, this is already very powerful. For instance, recent experimental demonstrations of exponential advantage for estimating Pauli observables using Bell measurements naturally operate on two copies of the input state at a time [125], and cannot be done with single-copy measurements. While the single-copy setting precludes even basic quantum learning primitives like the SWAP test, the t-copy setting encompasses a much broader class of possibilities. This begs the natural question:

Question 4. What is the copy complexity of quantum state tomography with t-copy measurements?

The essence of the above question is understanding whether we can interpolate between the single-copy and unconstrained multi-copy settings for quantum state tomography. Prior to the work discussed here, there has been very little understanding of this intermediate regime. Remarkably, it was not even known if we could achieve the optimal $\Theta(d^2/\varepsilon^2)$ copy complexity in the unconstrained setting just using measurements of $t = 2$ copies at a time! The t-copy setting poses an important conceptual and algorithmic challenge — the approaches in the single-copy and unconstrained multi-copy extremes are very different, and seem fundamentally incompatible. To address this, we will need to develop a new algorithmic framework to synthesize the two techniques.

1.3 Overview of Contributions

Here, we outline the results and contributions in this thesis, addressing all of the questions posed in the previous section.

1.3.1 Sudden Death of Entanglement

Our first main result concerns Question 1. We prove a new universal law about entanglement and temperature. The results discussed in this section are from [31] and

will be explained in more detail in Chapter 2.

Theorem 1.3.1 (Sudden Death of Entanglement (Informal, see Theorem 2.1.1))**.** *For any local Hamiltonian on* $\mathfrak{n}$ *particles, there is a critical (inverse) temperature* β_c *depending only on the geometry of the Hamiltonian and not on the system size* $\mathfrak{n}$ *such that for any* $\beta < \beta_c$*, the Gibbs state* ρ_β *has no entanglement.*

For a formal definition of what it means for a state to have no entanglement, see the definition of *separability* in Definition 1.4.20. With this result, we witness a *"sudden death of thermal entanglement"* — there is a constant critical temperature above which the Gibbs state has no entanglement. In other words, above this temperature, correlations are purely classical.

Despite the extensive body of work on understanding the relationship between entanglement and temperature, previous work yields little insight into when the Gibbs state can or cannot be entangled. For instance, computational investigations on specific models [21, 99, 18], show that various proxy entanglement measures like concurrence and negativity go to zero above a constant temperature, but these proxy measures can evaluate to zero even for entangled states. As for rigorous results, a simple argument about the geometry of entanglement shows that Gibbs states are unentangled, i.e. separable, at some temperature that is exponentially large in the system size $\mathfrak{n}$, but this dependence on system size makes it unphysical for many-body systems [100].

Over the past two decades, an exciting new suite of techniques has been developed to prove results about correlations in Gibbs states [7]. However, these results, including thermal area laws [213, 146], bounds on conditional mutual information [133], local indistinguishability [137], efficient state preparation [50, 40], and efficient learning algorithms [16], only bound entanglement through proxy correlation measures that combine both classical and quantum correlations.[§] Consequently, these results can only

[§]In the referenced work, such measures include mathematical quantities like two-point correlation functions, mutual information and conditional mutual information, and locality of effective Hamiltonians, along with operational and algorithmic notions like rapid thermalization (or low-depth state preparation more broadly) and efficient learning from local observables.

prove strong bounds on entanglement at *long range*, when classical correlations are also small. This remains true even for results which assume the Gibbs state is above a critical temperature [50, 108, 103]. The existing literature reflects a gap in our set of tools: we only have exiguous methods for controlling entanglement independently of classical correlations. Taken together, this body of work might even lead one to believe that quantum correlations, much like their classical counterparts, can exist at any temperature that is less than exponential in the system size.

In stark contrast, our result shows that above some constant temperature, the Gibbs state of any local Hamiltonian exhibits *zero* entanglement. In particular, our result separates quantum effects from classical correlations a feature not captured by any of the aforementioned vast literature on correlations in Gibbs states.

Algorithmic Inspiration The key idea behind our main result comes from a surprising source: an *algorithmic* technique, inspired by classical approximate counting and sampling. At a high level, the sampling-to-counting reduction of [129] reduces sampling from a high-dimensional distribution, say $\mathcal{D}$ over the hypercube $\{-1, 1\}^{\mathfrak{n}}$, to a sequence of "counting" problems where we compute the marginal distribution on the first coordinate, sample it, and then recurse on the posterior distribution over the remaining $\mathfrak{n} - 1$ coordinates. We develop a quantum analog of this sampling-to-counting reduction, where we decompose the Gibbs state by iteratively "sampling" and factorizing out one particle at a time — showing that this procedure succeeds suffices to prove that the original state is separable.

Gibbs Sampling This algorithmic technique has further implications for the computational task of preparing quantum Gibbs states. This task, known as quantum Gibbs sampling, has been studied extensively [60, Table 1], dating back to the work of [201]. However, remarkably little is known about provably efficient methods. Despite a wealth of approaches and proposals, efficient algorithms have only been rigorously established for highly restricted classes of Hamiltonians, such as Hamiltonians with constant operator norm [92], commuting Hamiltonians [132], 1D Hamiltonians [50, 40],

or under strong assumptions, such as the eigenstate thermalization hypothesis [59].

On the other hand, preparing classical Gibbs states at low temperatures is known to be NP-hard in general [196, 197, 90]. Thus, at least for general Hamiltonians, there is a natural target: Are all high-temperature Gibbs states efficiently preparable? We resolve this question:

Theorem 1.3.2 (Gibbs State Preparation (Informal, see Theorem 2.1.2)**.** *For any local Hamiltonian on* $\mathfrak{n}$ *particles, we can prepare its Gibbs state in time polynomial in* $\mathfrak{n}$ *at inverse temperatures* $\beta < \beta_c$ *where the critical (inverse) temperature* β_c *depends only on the geometry of the Hamiltonian and not on* $\mathfrak{n}$*.*

The task of preparing a Gibbs state has been widely proposed as a place to look for quantum speedups [201, 70, 60]. However, since our algorithm is almost entirely classical, we can conclude that this task offers no super-polynomial quantum speedup above the critical temperature. On the other hand, at temperatures below a (different) critical temperature, preparing a Gibbs state is NP-hard, so this task is also unsuitable for finding quantum advantage, at least for general Hamiltonians. This leaves a limited regime between our critical temperature and the computational hardness threshold. Looking forwards, obtaining a more fine-grained understanding of these thresholds will be crucial toward understanding Gibbs sampling as a candidate for quantum advantage.

1.3.2 Hamiltonian Learning

Our next main result delivers a surprising resolution to Question 2. We show that efficient Hamiltonian learning is possible at any constant temperature. The results discussed in this section are from [32] and will be explained in more detail in Chapter 3.

Theorem 1.3.3 (Efficient Hamiltonian Learning at Any Temperature (Informal, see Theorem 3.7.1))**.** *For any constant inverse temperature* $\beta > 0$ *and a local Hamiltonian* H *on* $\mathfrak{n}$ *particles with known geometry but unknown terms, we can recover the terms* E_a *of* H *to* ε *accuracy from measurements of (copies of) its Gibbs state* ρ_β *in time*

polynomial in $\mathfrak{n}, 1/\varepsilon$.

Whereas previous works [15, 103] gave polynomial time algorithms only when β was smaller than a fixed constant, our result gives a polynomial time algorithm for *any constant* β. This result is a fortunate development since, if learning were truly computationally hard in the low temperature regime, then, for example, we could not hope to understand the behavior of analog quantum simulators in precisely the regimes where they outperform classical ones [181, Section 6.10].

Remark 1.3.4 (Knowing the Geometry of the Hamiltonian)**.** Our result relies on the geometry of the Hamiltonian being known. This is typical in many applications when the particles are on a two or three-dimensional lattice. The variation of the problem when the geometry of the Hamiltonian is also unknown is called *structure-learning*. Designing efficient algorithms for structure learning from Gibbs states remains an interesting open question. There has been work on structure learning assuming stronger access models to the Hamiltonian, such as controlled time evolutions [223, 33, 157].

Previous Work For the problem of Hamiltonian learning from Gibbs states, Anshu, Arunachalam, Kuwahara, and Soleimanifar gave the first polynomial sample complexity bounds for this task in 2020 [15], attaining coefficient estimates using

$$\frac{2^{\mathrm{poly}(\beta)}\mathfrak{n}^2 \log \mathfrak{n}}{\beta^c \varepsilon^2} \qquad [15]$$

copies of the Gibbs state [103, Remark 4.5]. However, their work comes with a serious drawback: it is computationally inefficient. In particular, they give a stochastic gradient descent algorithm and show that it converges to the true parameters in a small number of iterations, but actually computing an iterate involves evaluating a log-partition function, which is well-known to be computationally hard even for classical systems [163].

Prior work has obtained fast algorithms for Hamiltonian learning in limited regimes. A follow-up paper of Anshu, Arunachalam, Kuwahara and Soleimanifar [14] shows that

when the terms of H commute, then a direct generalization of the classical algorithm learns the parameters efficiently. Further, [15] notes that their suggested algorithm can be performed in polynomial time for sufficiently high temperature (small β), since in this regime the log-partition function can be evaluated, using that its multivariate Taylor series expansion converges rapidly. Haah, Kothari, and Tang [103] later gave an improved algorithm that achieves the sample and time complexity

$$\frac{e^{\mathcal{O}(\beta)} \log \mathfrak{n}}{\beta^2 \varepsilon^2} \quad \text{and} \quad \frac{\mathfrak{n} e^{\mathcal{O}(\beta)} \log \mathfrak{n}}{\beta^2 \varepsilon^2}, \qquad \text{[103]}$$

respectively, which they prove is tight up to the constant factor in the exponential, even in the classical case.

Barriers at Low Temperature However, no strategies had been suggested for solving Hamiltonian learning at low temperature. In fact, the situation is even more dire: all approaches to Hamiltonian learning used in prior settings fail catastrophically at low temperatures, since reduction to sufficient statistics [15], efficient computation of the partition function [15], the approximate Markov property [145], and cluster expansion [103] all provably fail for sufficiently large constant β. This state of the literature reflects a broader scarcity of algorithmic tools known for understanding Hamiltonians outside of special settings like high temperature or one dimension. So, a negative resolution to the general, low temperature question seemed plausible, or even likely.

The Learning Theory Toolkit Previous approaches all formulate the learning problem as a system of equations in essentially the same way and thus run into the same fundamental barrier at low temperature. At a high level, they use measurements to estimate $\operatorname{tr}(X\rho_\beta)$ where X ranges over a basis for the allowable interaction terms of the Hamiltonian e.g. if the Hamiltonian involves only 2-body interactions between adjacent particles, then X will range over 2-body correlations between adjacent particles. They then solve a system of equations for a Hamiltonian H', such that the Gibbs state for H', say ρ'_β matches those measurement values i.e. $\operatorname{tr}(X\rho_\beta) = \operatorname{tr}(X\rho'_\beta)$.

This is, in some sense, the most natural system of equations to set up and [15] shows that statistically, solving this system suffices to recover the true Hamiltonian. However, the challenge is to give an efficient algorithm for solving this system of equations. Unfortunately, this is already a dead end because for classical Hamiltonians, this system of equations exactly reduces to "matching sufficient statistics of a graphical model", which is known to be computationally intractable even for simple Ising models [163].

We find new light from an important paradigm in classical learning theory — we can formulate an alternative system, by measuring a richer family of statistics, that may be more tractable. This is the starting point of our approach. In the classical setting, there are simple ways to measure a richer family of statistics and reformulate the system, relying on the Markov property, the defining characteristic of graphical models. However, the Markov property is known to fail for quantum systems. To overcome this, we rely on another key workhorse in learning theory, polynomial approximations. We develop a new polynomial approximation to the matrix exponential that lets us understand the propagation of information in spite of the failure of the Markov property. Using this approximation, we can formulate the learning problem as a polynomial system.

To solve the polynomial system, we rely on the sum-of-squares (SoS) method, which gives a hierarchy of semidefinite programming relaxations that are efficiently solvable. The SoS method, first introduced in the context of control theory and polynomial optimization [177, 147], has been at the heart of recent advances across a wide range of domains including complexity theory [37, 36], combinatorial optimization [183, 35], and learning theory [38, 116, 142, 139, 76], and here we show that this powerful tool also leads to a new algorithm for learning about quantum systems.

1.3.3 Quantum State Tomography

Now we discuss our contributions to learning general quantum states. Our result stated below settles the complexity of single-copy quantum state tomography, answering

Question 3. This, and other results discussed in this section are from [62] and will be explained in more detail in Chapter 4.

Theorem 1.3.5 (Informal, see Theorem 4.2.11)**.** *For an unknown state $\rho \in \mathbb{C}^{d\times d}$, $n = \Omega(d^3/\varepsilon^2)$ copies of ρ are necessary for single-copy quantum state tomography of ρ to within trace distance ε.*

The lower bound shown here matches the upper bound obtained by the random measurement scheme in [144] showing that $\Theta(d^3/\varepsilon^2)$ is the optimal copy complexity for single-copy quantum state tomography.

While this is not the most satisfying resolution, that, up to constant factors, we might as well measure all of the copies randomly, we show that the situation is different if we change the way that we measure accuracy. In particular, if instead of trace distance, we measure accuracy in terms of infidelity (see Definition 1.4.7) between our estimate $\widehat{\rho}$ and ρ, then adaptivity does make a difference. We show:

Theorem 1.3.6 (Informal, see Theorem 4.5.1)**.** *For an unknown state $\rho \in \mathbb{C}^{d\times d}$, $n = \widetilde{\Theta}(d^3/\gamma)$ copies of ρ are necessary and sufficient for single-copy quantum state tomography of ρ to within infidelity γ.*

This result contrasts the lower bound in [102] showing that $\Omega(d^3/\gamma^2)$ copies are necessary with nonadaptive measurements. Thus, for learning in fidelity, adaptive measurements are asymptotically more powerful than nonadaptive measurements. This provides arguably the first natural instance of a separation between the power of adaptive and nonadaptive measurements.

Adaptive Lower Bounds Focusing on Theorem 1.3.5, the main novelty is a framework for proving a lower bound against a learner that makes adaptive measurements. In general, existing lower-bound approaches in the literature mostly followed techniques from classical learning theory. Thus they rely heavily on the observations being independent and these approaches can only work for nonadaptive measurements.

We introduce a new, general framework for proving lower bounds against adaptive learners. Our framework involves a bare-handed approach where we directly bound the information gained by the learner after each measurement. The unknown state ρ is sampled from a prior over the space of all possible states and then we directly track how the posterior distribution evolves after each measurement outcome. We prove that no matter what the learner does, the posterior remains mostly away from the true state ρ and thus the learner cannot recover ρ accurately with high probability.

1.3.4 State Certification

Beyond quantum state tomography, we also study the closely related problem of state certification. In state certification, rather than trying to learn the unknown state ρ, we are given an explicit description of a state $\sigma \in \mathbb{C}^{d \times d}$ and our goal is simply to test whether $\rho = \sigma$ or ρ is far from σ in trace distance. From a practical point of view, state certification is also a key subroutine in verification of the outcomes of quantum experiments. For instance, if an experimentalist wishes to check that a quantum algorithm with quantum output is running correctly and outputting the right state, then this is exactly the problem of state certification.

State certification also has important mathematical connections to classical property testing. While state tomography is the quantum analogue of learning a classical distribution from samples, state certification is the natural quantum analogue of identity testing, one of the most well-studied problems in classical distribution testing.

Prior to the work discussed here, the situation for state certification was similar to that for quantum state tomography. Recent works [169, 27] characterize the optimal copy complexity of state certification in the unconstrained multi-copy setting, but in the single-copy setting, the question remained open. The best upper and lower bounds were $n = O(d^{3/2}/\varepsilon^2)$ and $n = \Omega(d^{4/3}/\varepsilon^2)$ respectively [54].

Our result for state certification is a tight lower bound that closes this gap. This result is from [67] and will be explained in more detail in Chapter 5.

Theorem 1.3.7 (Informal, see Theorem 5.1.1)**.** *State certification for a state $\rho \in \mathbb{C}^{d\times d}$ using single-copy measurements requires $n = \Omega(d^{3/2}/\varepsilon^2)$ copies (for worst case over σ).*

While the above settles the complexity of state certification against a worst-case target state σ, we can obtain improved bounds that depend on σ — in particular, in Chapter 5, we will also explain how to extend our result to obtain *instance optimal* bounds for state certification in terms of σ.

Following the theme for state tomography in the single-copy setting, the power of adaptivity is again an important challenge for certification. The proof of the lower bound in Theorem 1.3.7 builds on the framework introduced in our lower bound for tomography (Theorem 1.3.5), carefully tracking the information gained by the learner after each measurement, but requires a more refined analysis.

1.3.5 Quantum Memory-Sample Tradeoffs

Finally, we return to quantum state tomography and present our results for interpolating between the single-copy and unconstrained multi-copy settings. We give a smooth and optimal (up to logarithmic factors) interpolation in many parameter regimes, addressing Question 4. This result is from [68] and will be explained in more detail in Chapter 6.

Theorem 1.3.8 (Informal, see Theorem 6.4.8 and Theorem 6.6.13)**.** *For an unknown state $\rho \in \mathbb{C}^{d\times d}$ and $t \leqslant d^2$, $n = \widetilde{\Theta}(d^3/(\sqrt{t}\varepsilon^2))$ are necessary and sufficient for t-copy quantum state tomography of ρ to within trace distance ε .*

Observe that the above result matches both the fully entangled bound of $\Theta(d^2/\varepsilon^2)$ (by taking $t = d^2$) and the single-copy bound of $\Theta(d^3/\varepsilon^2)$ (when $t = 1$). For formal statements of the precise regimes of d, t, ε where our results hold, see Chapter 6. Worth noting is that this gives the first smooth tradeoff between entanglement (between copies) and overall copy complexity for a natural quantum learning task.

Resolving the Incompatibility Focusing on the upper bound, the main challenge is to resolve the incompatibility between previous approaches in the fully entangled and single-copy settings as we explain below. The fully entangled algorithms of [102] and [170] make a single measurement on the joint state $\rho^{\otimes n}$ over all of the copies we are given. This measurement is based on a technique called weak Schur sampling [134], which is grounded in the representation theory of the symmetric and general linear groups.

However, this approach seems inherently tied to running a "one-shot" algorithm. This is because the estimator for ρ derived from Schur sampling (e.g. in [170]) is actually a *biased* estimator — its expectation is not equal to ρ. The bias decreases with the number of copies we measure at once, but is only sufficiently small when this number is $\Omega(d^2/\varepsilon^2)$. For small t, it is unclear if weak Schur sampling on t copies gives any useful information, let alone if one can somehow combine the outcome of several different trials of weak Schur sampling in a meaningful fashion. This is in contrast to algorithms for single-copy measurements, which rely on many unbiased (but less informative) measurements of the underlying state, which can then be averaged to obtain a good overall estimate. Unfortunately, this unbiasedness seems specific to single-copy measurements. Ultimately, neither approach by itself seems capable of yielding a nontrivial result for general t-copy measurements.

This calls for a new algorithmic framework that can synthesize the two techniques. We introduce a two-stage approach. We first analyze the case when ρ is close to maximally mixed, which we call the balanced case. We show that in this case, we can use a linear approximation to a power series to upper bound the bias in the estimators derived from Schur sampling in [170]. We then give a general reduction, that we term *quantum splitting*. This allows us to reduce the general case to the balanced one by obtaining a rough estimate of ρ and then simulating the learning algorithm run on a (reversibly) transformed version of ρ that is close to maximally mixed.

1.4 Notation and Preliminaries

We begin by defining some notation that will be used throughout this thesis. For any natural number k, we use the notation $[k] = \{1, 2, \ldots, k\}$. We take $\mathrm{i} = \sqrt{-1}$ to be the imaginary unit. For vectors in $\mathbb{C}^d$, we will use e_i to denote the ith standard basis vector, i.e. the vector that has a 1 in its ith coordinate and is 0 everywhere else.

For a matrix A, we use $A^\dagger$ to denote its conjugate transpose, $\|A\|_{\mathrm{op}}$ to denote its operator norm, $\|A\|_F$ to denote its Frobenius norm, and $\|A\|_1$ to denote its trace norm; for a vector v, we use $\|v\|$ to denote its Euclidean norm. For operators A, B, $A \preceq B$ denotes the Loewner order: $A \preceq B$ when $B - A$ is PSD (positive semi-definite), and similarly for $\succeq$. We use I_d to denote the $d \times d$ identity matrix.

Throughout this thesis, we will use the usual asymptotic notation $\mathcal{O}(f)$, $\Omega(f)$, and $\Theta(f)$. To suppress logarithmic factors, we will use $\widetilde{\mathcal{O}}(f) = \mathcal{O}(f \operatorname{polylog}(f))$ and $\widetilde{\Omega}(f) = \Omega(f/\operatorname{polylog}(f))$. We say $f = \widetilde{\Theta}(g)$ if $f = \widetilde{\mathcal{O}}(g)$ and $f = \widetilde{\Omega}(g)$. We use the notation $f \lesssim g$ to mean $f = \mathcal{O}(g)$. For a parameter t, $\mathcal{O}_t$ denotes big O notation where t is treated as a constant; the same holds for the notation for polynomial scaling, $\operatorname{poly}_t(\cdot)$.

Beyond the usual operations like addition and multiplication, the following operations on matrices will be important later on.

Definition 1.4.1 (Hadamard product)**.** For $A, B \in \mathbb{C}^{N \times N}$, their *Hadamard product*, denoted $A \circ B$, satisfies $[A \circ B]_{ij} = A_{ij} B_{ij}$.

Definition 1.4.2 (Commutator)**.** Given operators $A, B \in \mathbb{C}^{N \times N}$, the *commutator* of A and B is defined as $[A, B] = AB - BA$. The *nested commutator* of order ℓ is defined recursively as $[A, B]_k = [A, [A, B]_{k-1}]$, with $[A, B]_1 = [A, B]$.

1.4.1 Quantum States and Measurements

Here we introduce the basic formalisms for quantum states and measurements. A quantum state is represented by a matrix $\rho \in \mathbb{C}^{d\times d}$ which is PSD and satisfies $\mathrm{tr}(\rho) = 1$. Such a matrix ρ is called a density matrix. We say the state ρ is pure if it is rank-1.

Measurements. We now define the standard measurement formalism, which is the way a learner can interact with a quantum state ρ.

Definition 1.4.3 (Positive operator valued measurement (POVM), see e.g. [168])**.** A positive operator valued measurement $\mathcal{M}$ is a finite collection of psd matrices $\mathcal{M} = \{M_z\}_{z\in\mathcal{Z}}$ satisfying $\sum_z M_z = I_d$. When a state ρ is measured using $\mathcal{M}$, we get a draw from a classical distribution over $\mathcal{Z}$, where we observe z with probability $\mathrm{tr}(\rho M_z)$. Afterwards, the quantum state is destroyed.

Remark 1.4.4. Note that the outcome distribution is indeed a valid probability distribution — the probabilities are nonnegative because ρ and M_z are PSD and sum to 1 since $\sum_z \mathrm{tr}(\rho M_z) = \mathrm{tr}(\rho) = 1$.

Quantum measurements are a generalization of getting samples from a classical probability distribution. Concretely, we can view a classical (discrete) probability distribution $\mathcal{D}$ on d elements, with density $(p_1, \ldots, p_d)$, as a $d \times d$ quantum state with entries $p_1, \ldots, p_d$ on the diagonal. Measuring this quantum state in the computational basis, i.e. with the POVM given by $\{e_1e_1^\dagger, \ldots, e_de_d^\dagger\}$, is exactly equivalent to getting a classical sample from $\mathcal{D}$. Note that in this analogy, a pure state corresponds to a trivial distribution on a single point.

Next, we define a few notions of distance between quantum states. These distances generalize standard classical notions of distance between probability distributions.

Definition 1.4.5. [Trace Distance] For two quantum states $\rho, \sigma \in \mathbb{C}^{d\times d}$, the trace distance between them is $\|\rho - \sigma\|_1$.

Remark 1.4.6. Note that this is the natural generalization of total variation distance

between classical probability distributions (up to a factor of 2).

Definition 1.4.7. [Infidelity] For two quantum states $\rho, \sigma \in \mathbb{C}^{d\times d}$, the fidelity between ρ, σ is $F(\rho,\sigma) = \mathrm{tr}(\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}})^2$, and the infidelity is $1 - F(\rho,\sigma)$.

Remark 1.4.8. Fidelity is the quantum analogue of the *Bhattacharyya coefficient*, which for two classical probability distributions p, q over $[d]$, is defined to be $BC(p,q) = \sum_{i=1}^{d} \sqrt{p_i}\sqrt{q_i}$. Note that if ρ and σ commute, then $F(\rho,\sigma) = BC(p,q)^2$, where p, q are the classical distributions given by the eigenvalues of ρ and σ respectively.

1.4.2 Primer on Many-Body Systems

Now we introduce the basic foundational concepts for describing quantum many-body systems. Throughout this thesis, we will use $\mathfrak{n}$ to denote the number of particles in the system. We assume that the particles are represented by qubits, so each particle is associated to a local Hilbert space $\mathbb{C}^2$. The state of each particle is described by a density matrix in $\mathbb{C}^{2\times 2}$. The state of the full system is represented by a density matrix in $\underbrace{\mathbb{C}^{2\times 2} \otimes \cdots \otimes \mathbb{C}^{2\times 2}}_{\mathfrak{n}}$; we will often denote this space simply as $\mathbb{C}^{2^{\mathfrak{n}}\times 2^{\mathfrak{n}}}$.

The following notion, which identifies the particles upon which an operator acts nontrivially, will be crucial.

Definition 1.4.9 (Support of an operator)**.** For an operator $P \in \mathbb{C}^{2^{\mathfrak{n}}\times 2^{\mathfrak{n}}}$ on a system of $\mathfrak{n}$ qubits, its *support*, $\mathrm{supp}(P) \subseteq [\mathfrak{n}]$ is the subset of qubits that P acts non-trivially on. That is, $\mathrm{supp}(P)$ is the minimal set S of qubits such that P can be written as $P = O_S \otimes I_{[\mathfrak{n}]\setminus S}$ where $O \in \mathbb{C}^{2^{|S|\times|S|}}$ is an operator on the subsystem of particles in S and $I_{[\mathfrak{n}]\setminus S} = I_{2^{\mathfrak{n}}-|S|}$ is the identity matrix (of appropriate dimension).

In other words, an operator P is supported on a subset of qubits S if P can be written as the tensor product of some operator on the subsystem S and the identity matrix on $[\mathfrak{n}] \setminus S$. Generally, when we say that particles interact locally, we mean that their interactions can be described by operators that are supported only on adjacent or small neighborhoods of particles. This will be made more formal below.

A basic property that we will use repeatedly later on is that the matrix exponential does not increase support size, which follows from its series expansion.

Fact 1.4.10. *For a square matrix B, $e^{I\otimes B} = I \otimes e^B$.*

Later on, it will be useful to work with the following canonical orthonormal basis for the space of 2×2 Hermitian matrices, called the Pauli basis.

Definition 1.4.11 (Pauli matrices)**.** The Pauli matrices are the following 2×2 Hermitian matrices.

$$\sigma_I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \sigma_y = \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix}, \qquad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

These matrices are unitary and (consequently) involutory. Further, $\sigma_x\sigma_y = \mathrm{i}\sigma_z$, $\sigma_y\sigma_z = \mathrm{i}\sigma_x$, and $\sigma_z\sigma_x = \mathrm{i}\sigma_y$, so the product of Pauli matrices is a Pauli matrix, possibly up to a factor of $\{\mathrm{i}, -1, -\mathrm{i}\}$. The non-identity Pauli matrices are traceless. We also consider tensor products of Pauli matrices, $P_1 \otimes \cdots \otimes P_\mathfrak{n}$ where $P_i \in \{\sigma_I, \sigma_x, \sigma_y, \sigma_z\}$ for all $i \in [\mathfrak{n}]$. The set of such products of Pauli matrices, which we denote $\mathcal{P}$, form an orthogonal basis for the vector space of $2^\mathfrak{n} \times 2^\mathfrak{n}$ (complex) Hermitian matrices under the trace inner product. The product of two elements of $\mathcal{P}$ is an element of $\mathcal{P}$, possibly up to a factor of $\{\mathrm{i}, -1, -\mathrm{i}\}$.

Example 1.4.12. The support of a tensor product of Paulis, $P_1 \otimes \cdots \otimes P_\mathfrak{n}$ is the set of $i \in [\mathfrak{n}]$ such that $P_i \neq \sigma_I$.

Local Hamiltonians

Now we can define notions of locality and local Hamiltonians, which describe the interactions between particles. Specifically, we consider a class of Hamiltonians which are local with respect to an underlying graph, which we call *low-intersection* Hamiltonians. This means that the Hamiltonian is a sum of individual interaction terms, each involving only a small number of particles, and such that any term has nonempty intersection with only a small number of other terms.

Definition 1.4.13 (Hamiltonian)**.** A *Hamiltonian* on $\mathfrak{n}$ qubits is an operator $H \in \mathbb{C}^{2^{\mathfrak{n}} \times 2^{\mathfrak{n}}}$ that we consider as a sum of m local *terms* H_a, with $H = \sum_{a=1}^{m} H_a$. We also refer to these qubits as *sites*. For normalization, we assume that the terms have bounded operator norm, $\|H_a\|_{\text{op}} \leqslant 1$. We say this Hamiltonian is $\mathfrak{K}$-*local* if every term H_a is supported on at most $\mathfrak{K}$ qubits: $|\text{supp}(H_a)| \leqslant \mathfrak{K}$.

We say the Hamiltonian has Pauli terms if, for every $a \in [m]$, $H_a = \lambda_a E_a$ for $E_a \in \mathcal{P} = \{I, \sigma_x, \sigma_y, \sigma_z\}^{\otimes \mathfrak{n}}$ a tensor product of Pauli matrices (see Definition 1.4.11) and $-1 \leqslant \lambda_a \leqslant 1$.

Definition 1.4.14 (Low-intersection Hamiltonian)**.** For an $\mathfrak{n}$-qubit Hamiltonian $H = \sum_{a=1}^{m} H_a$, we define its underlying *dual interaction graph* $\mathfrak{G}$ to have vertices labeled by $\{1, 2, \ldots, m\}$ and an edge between a and b if and only if $\text{supp}(H_a) \cap \text{supp}(H_b) \neq \varnothing$. We say H has degree $\mathfrak{d}$ if the degree of every vertex in $\mathfrak{G}$ is at most $\mathfrak{d}$.[¶] We call a Hamiltonian H a $(\mathfrak{d}, \mathfrak{K})$-*low-intersection Hamiltonian* if H has locality $\mathfrak{K}$ and degree $\mathfrak{d}$.

We will mostly work with $(\mathfrak{d}, \mathfrak{K})$-low-intersection Hamiltonians with distinct, traceless Pauli terms. We think of n as very large, going to infinity, but $\mathfrak{d}, \mathfrak{K}$ as fixed constants. Note that the parameters $\mathfrak{d}, \mathfrak{K}$ determine the "geometry" of the Hamiltonian and thus we can talk about the system size as growing with a fixed geometry.

Assuming that the terms are Pauli can be done without loss of generality: an arbitrary H_a supported on $\mathfrak{K}$ qubits can be expanded into the Pauli basis, inflating the number of terms by at most a factor of $4^{\mathfrak{K}}$ (which is just a constant). The assumption that the terms are traceless is also without loss of generality, since adding a multiple of the identity to the Hamiltonian does not affect its corresponding Gibbs state.

Example 1.4.15. Geometrically local Hamiltonians on low-dimensional lattices are Hamiltonians where the particles are arranged on some low-dimensional lattice and the Hamiltonian is a sum of terms that are supported on the different connected

[¶]In the classical literature, it is more conventional to define the degree of a spin system to be the degree of the *interaction graph*, the most number of terms which act on the same site. To be consistent with the quantum information literature, we use dual interaction graph degree, and note that this degree is at most $\mathfrak{K}$ times the degree of the interaction graph.

components (of the lattice's adjacency graph) of size at most $\mathfrak{K}$. Geometrically local Hamiltonians are a type of Hamiltonian often considered in physically motivated settings. Low-intersection Hamiltonians generalize geometrically local Hamiltonians, at least in constant dimensional spaces (which is by far the most common setting). For example, a $\mathfrak{K}$-local Hamiltonian which is geometrically local with respect to a D-dimensional lattice is a $(\mathfrak{K}(2D)^{\mathfrak{K}-1}, \mathfrak{K})$-low intersection Hamiltonian.

Efficiency For Hamiltonian problems, even though the dimension of the underlying Hilbert space is $2^{\mathfrak{n}}$, we will be interested in algorithms that run in time polynomial in $\mathfrak{n}$. Note as discussed below, any low-intersection Hamiltonian has description length $O(\mathfrak{n})$ and thus, we can at least in principle, work only with succinct representations that have description length $\text{poly}(\mathfrak{n})$.

Definition 1.4.16 (Hamiltonian input model)**.** Recall the definition of a $\mathfrak{K}$-local Hamiltonian (Definition 1.4.13). When a Hamiltonian $H = \sum_{a=1}^{m} H_a$ is given as input, we assume it is given as a list of the terms restricted to their support, along with their corresponding supports, $(\text{supp}(H_1), (H_1)_{\text{supp}(H_1)}), \ldots, (\text{supp}(H_m), (H_m)_{\text{supp}(H_m)})$. This description has size linear in $\mathfrak{n}$ and m. If this Hamiltonian also has degree $\mathfrak{d}$, then the total number of terms must satisfy $m \leqslant \mathfrak{d}\mathfrak{n}$.

We will commonly need to consider Hamiltonians restricted to various subsets of terms, which we notate in the following way.

Definition 1.4.17 (Restricted Hamiltonian)**.** For a Hamiltonian $H = \sum_{a=1}^{m} H_a$ on n sites and a subset of sites $S \subseteq [\mathfrak{n}]$, we define the restricted Hamiltonian and its corresponding set of terms to be

$$H^{(S)} = \sum_{a \in \mathcal{E}^{(S)}} H_a \quad \text{where} \quad \mathcal{E}^{(S)} = \{a \in [m] \mid \text{supp}(H_a) \subseteq S\}$$

and the localized Hamiltonian to be

$$H_{(S)} = \sum_{a \in \mathcal{E}_{(S)}} H_a \quad \text{where} \quad \mathcal{E}_{(S)} = \{a \in [m] \mid \text{supp}(H_a) \cap S \neq \varnothing\}\,.$$

We will also consider restricting a Hamiltonian to a set of terms, $H^{(\mathcal{Q})} = \sum_{a \in \mathcal{Q}} H_a$ when $\mathcal{Q} \subseteq [m]$.

The Gibbs State

For particles whose interactions are given by a Hamiltonian H, the Gibbs state is a $2^{\mathfrak{n}} \times 2^{\mathfrak{n}}$ density matrix that describes the state of the system at thermal equilibrium at a specified temperature. Formally,

Definition 1.4.18 (Gibbs State)**.** For a Hamiltonian H on $\mathfrak{n}$ qubits, the Gibbs state at an inverse temperature $\beta > 0$ is given by

$$\rho_\beta = \frac{e^{-\beta H}}{\text{tr}(e^{-\beta H})}\,.$$

Note that the state ρ_β defined above is indeed a valid density matrix — it is PSD because of the matrix exponential and H is Hermitian and it has trace 1 due to the normalization in the denominator.

The Gibbs state generalizes the classical Gibbs distribution from statistical mechanics. We give an illustrative example below.

Example 1.4.19 (Classical Ising Model)**.** A classical Ising model is a distribution over $\{-1, 1\}^{\mathfrak{n}}$ where $p(x_1, \ldots, x_{\mathfrak{n}}) \propto \exp(-\beta \sum_{(i,j) \in E} c_{ij} x_i x_j)$ for some graph on $[\mathfrak{n}]$ with edge set E and some real coefficients c_{ij} with $|c_{ij}| \leqslant 1$ (for sake of normalization).

To see how this is a special case of a local Hamiltonian, for each edge $(i,j) \in E$, define

$$H_{ij} = \underbrace{\begin{bmatrix} c_{ij} & & & \\ & -c_{ij} & & \\ & & -c_{ij} & \\ & & & c_{ij} \end{bmatrix}}_{\text{on } i,j} \otimes \underbrace{I_{2^{\mathfrak{n}-2}}}_{\text{on } [\mathfrak{n}]\setminus\{i,j\}}.$$

Observe that $H = \sum_{(i,j)\in E} H_{ij}$ is a $2^{\mathfrak{n}} \times 2^{\mathfrak{n}}$ diagonal matrix. If we index the diagonal entries with the possible assignments to $(x_1, \ldots, x_{\mathfrak{n}}) \in \{-1, 1\}^{\mathfrak{n}}$ in the natural way, then the diagonal entries of H are exactly the values of $\sum_{(i,j)\in E} c_{ij} x_i x_j$ for the possible assignments to $(x_1, \ldots, x_{\mathfrak{n}})$. Thus, the Gibbs state $\rho_\beta = e^{-\beta H}/\operatorname{tr}(e^{-\beta H})$ exactly captures the classical Ising model distribution — measuring it in the computational basis is exactly equivalent to sampling from the Ising model (also recall the discussion after Definition 1.4.3).

1.4.3 Quantum Entanglement

Here we provide a basic discussion of quantum entanglement. There is a vast literature on different measures and definitions of quantum entanglement [47, 214, 208, 10, 119] and we refer the reader to [168] and [180] for a more in depth discussion. But for the results presented in this thesis, we will only need the most basic definitions.

We start with a very simple example. For a two-qubit system, a state $\rho \in \mathbb{C}^{4\times 4}$ is a product state if it can be written as a tensor product $\rho = \rho_1 \otimes \rho_2$ for two states $\rho_1, \rho_2 \in \mathbb{C}^{2\times 2}$. More generally, for a $\mathfrak{n}$-qubit system, a state $\rho \in \mathbb{C}^{2^{\mathfrak{n}}\times 2^{\mathfrak{n}}}$ is a product state if it can be written as a tensor product $\rho = \rho_1 \otimes \rho_2 \otimes \cdots \otimes \rho_{\mathfrak{n}}$ for states $\rho_1, \ldots, \rho_{\mathfrak{n}} \in \mathbb{C}^{2\times 2}$. In product states, the $\mathfrak{n}$ qubits essentially behave independently, and thus they are defined to be unentangled.

For pure states, this definition suffices — a pure state has no entanglement if and only if it is a product state. However, for mixed states, we require a broader definition. As

defined below, any convex combination of product states is said to be separable.

Definition 1.4.20 (Separable operator)**.** For an operator $A \in \mathbb{C}^{2^{\mathfrak{n}} \times 2^{\mathfrak{n}}}$ on $\mathfrak{n}$ qubits, we say that A is a *product operator* if it can be expressed as a tensor product of operators over subsystems, $A = A^{(1)} \otimes \cdots \otimes A^{(\mathfrak{n})}$. A is *separable* if it can be written as a sum of product operators $A = \sum_i (A_i^{(1)} \otimes \cdots \otimes A_i^{(\mathfrak{n})})$ where every $A_i^{(j)}$ is positive semi-definite. We say a state is separable if its density matrix ρ is separable.

Separable states are defined to be unentangled, whereas any state that is not separable is said to be entangled.

Part I

Understanding Locally Interacting Systems

Chapter 2

On the Sudden Death of Entanglement

Entanglement is, in many ways, the defining property of quantum systems [190, 43, 46]. In this chapter, we focus on understanding the relationship between entanglement and temperature in quantum many-body systems — specifically at what temperatures these systems can exhibit entanglement.

This question, about the relationship between temperature and entanglement, dates back to the early days of modern quantum information [179, 22, 173]. Yet, the vast body of work since then has been unable to rigorously address this question. We present a striking new universal law about entanglement and temperature. We prove that for any local Hamiltonian, its Gibbs state becomes unentangled above some constant critical temperature, depending only on the geometry of the Hamiltonian, but independent of the system size. Our proof of this law takes inspiration from an unexpected source — classical approximate counting and sampling algorithms. Thus, our technique also gives important algorithmic consequences for efficient, almost-classical preparation of Gibbs states.

2.1 Results

The main result of this chapter, that Gibbs states are separable (i.e. unentangled) above a constant critical temperature, is stated formally below.

Theorem 2.1.1 (Separability of Gibbs States Above Critical Temperature)**.** *Let H be a $(\mathfrak{d}, \mathfrak{K})$-low-intersection Hamiltonian on $\mathfrak{n}$ qubits with Pauli terms. Then, for $\beta < 1/(100\mathfrak{d}\mathfrak{K})$, the corresponding Gibbs state $\rho = e^{-\beta H}/\operatorname{tr}(e^{-\beta H})$ is separable.*

In particular, ρ can be expressed as a distribution over stabilizer product states, i.e. $A_1 \otimes A_2 \otimes \ldots \otimes A_n$ such that for each $j \in [n]$, $A_j \in \left\{ \frac{1}{2}(I \pm \sigma_x), \frac{1}{2}(I \pm \sigma_y), \frac{1}{2}(I \pm \sigma_z) \right\}$.*

We will also show that we can efficiently sample from the separable distribution over product states and this gives us an almost fully classical algorithm for preparing the Gibbs state above the critical temperature.

Theorem 2.1.2 (Classical Preparation of Gibbs States Above Critical Temperature)**.** *Given $\varepsilon \in (0, 1)$, a $(\mathfrak{d}, \mathfrak{K})$-low-intersection Hamiltonian H on $\mathfrak{n}$ qubits with Pauli terms, and a $\beta < \beta_c = 1/(\gamma\mathfrak{d}\mathfrak{K})$, where γ is a fixed universal constant, consider the Gibbs state $\rho = e^{-\beta H}/\operatorname{tr}\left(e^{-\beta H}\right)$. Then, there exists a classical randomized algorithm which runs in time*

$$\widetilde{\mathcal{O}}\left(\mathfrak{n}^{6+\frac{\log(\mathfrak{d})}{\log(\beta_c/\beta)}} \cdot \log^3(1/\varepsilon) \cdot \operatorname{poly}(\mathfrak{K}, \mathfrak{d})\right).$$

The algorithm outputs a classical description of a product state $\widehat{\rho} = A_1 \otimes \cdots \otimes A_n$ where every A_j is an eigenvector of a Pauli matrix, $A_j \in \left\{ \frac{1}{2}(I \pm \sigma_x), \frac{1}{2}(I \pm \sigma_y), \frac{1}{2}(I \pm \sigma_z) \right\}$, such that the mixture over $\widehat{\rho}$ is close to ρ in trace distance,

$$\left\| \rho - \mathbb{E}[\widehat{\rho}] \right\|_1 \leqslant \varepsilon,$$

where the expectation is only over the randomness of the algorithm.

As long as $\beta \leqslant \beta_c/\mathfrak{d}$, the runtime above is fully polynomial in all parameters. The

*The single-qubit stabilizer states, written here as $\frac{1}{2}(I \pm \sigma_z)$, $\frac{1}{2}(I \pm \sigma_x)$, and $\frac{1}{2}(I \pm \sigma_y)$, may be more familiar as $\{|0\rangle\langle 0|, |1\rangle\langle 1|, |+\rangle\langle +|, |-\rangle\langle -|, |\mathrm{i}\rangle\langle \mathrm{i}|, |-\mathrm{i}\rangle\langle -\mathrm{i}|\}$.

output $\widehat{\rho}$ can be prepared on a quantum computer with one layer of single-qubit gates. With this algorithm, one can prepare a copy of ρ by running our randomized algorithm, taking the classical description of $\widehat{\rho}$, and preparing it with a depth-one quantum circuit. Note that preparing ρ in expectation is equivalent to preparing ρ, since one can just "forget the algorithm's steps" to get a copy of ρ without classical side correlations. Even stronger, if one performs the algorithm coherently, it outputs a purification of ρ.

On temperature. The bound on the critical temperature for the algorithm cannot be significantly improved. Work of Sly and Sun [197] shows that approximately sampling from the anti-ferromagnetic Ising model on a Δ-regular graph is NP-hard at the "uniqueness threshold" for the Δ-regular tree, which is at $\beta = \Theta(1/\Delta) = \Theta(1/\mathfrak{d})$ [195]. This hardness statement for classical Gibbs sampling implies hardness for the more general problem of quantum Gibbs sampling.

On quantum advantage. The results show that any reasonable downstream computational task derived from high-temperature Gibbs sampling can be performed efficiently on a classical computer, and so is not a good candidate for quantum advantage. In some sense, this strengthens prior work on algorithmic cluster expansion, which gives efficient classical algorithms for various tasks associated with high-temperature Gibbs states, including computing partition functions [160], estimating local observables $\mathrm{tr}(O\rho)$, and sampling from ρ measured in a product basis [218]. Indeed, our structural result shows that such states are almost entirely classical, giving a simple explanation for why these efficient classical algorithms are possible.

2.1.1 Additional Related Work

Locality in high-temperature Gibbs states. Several works in the quantum information literature focus on high-temperature Gibbs states. These bound correlation measures which do not distinguish quantum and classical, including covariance of observables [108] and local indistinguishability [137, 50]. Our work, which shows a lack

of quantum correlation even in the presence of classical correlations, is a significant departure from this prior work.

Sudden death of entanglement. Sudden death of entanglement is the phenomenon that two entangled qubits, when subject to environmental noise, does not exhibit exponentially decaying entanglement with time, as classical correlations do, but rather become entirely disentangled after a finite amount of time [222]. The body of literature on ESD (entanglement sudden death) focuses on analyzing two-qubit systems under various noise models [221, 83, 6] and experimental demonstrations of ESD [8]. Its study as a phenomenon of many-body systems is more limited, likely because even defining a measure of entanglement for mixed states is non-trivial [119], and separability is difficult to detect for large systems.

For many-body systems, prior work studies the sudden death of *entanglement negativity*, an entanglement monotone which can be computed efficiently [207], in specific spin systems, either through heuristic arguments or numerical calculations [21, 99, 18, 191, 109]. States with zero entanglement negativity need not be separable [117]. So, our work is vastly more general and formally proves separability of Gibbs states at high temperature, a much stronger result than lack of entanglement negativity.

Other kinds of quantum systems. There are natural quantum systems which exhibit entanglement at even infinite temperature: a recently studied example is systems with non-abelian symmetries [151, 162]. From the perspective of statistical physics, our setting of considering states $\rho \propto e^{-\beta H}$ is closest to the grand canonical ensemble, where there is no particle nor energy conservation; when conservation is enforced, entanglement appears able to persist to high temperature, suggesting a possible difference between these statistical ensembles.

Geometry of entanglement. Prior arguments for the separability of Gibbs states use properties about the geometry of entanglement [45], and in particular, the distance from the maximally mixed state to the closest non-separable state. This distance

is exponentially small in the number of subsystems $\mathfrak{n}$, as shown by Gurvits and Barnum [100], and subsequent work has further strengthened and generalized this result [24, 114, 211]. As previously observed, this suffices to show separability for Gibbs states for exponentially small β [56, 176]. Since our result shows separability at constant temperature, it implies that sudden death of thermal entanglement occurs at physically reasonable temperatures, even for many-body systems.

Classical Gibbs sampling. Classical Gibbs sampling is, comparatively, well-understood and researchers have characterized sharp phase transitions wherein there is some critical temperature, above which sampling the Gibbs state is computationally efficient and below which it is computationally hard [196, 197, 90]. The fact that there are such wide gaps in our understanding of quantum Gibbs sampling is especially surprising given the diverse range of techniques we have for classical spin systems, such as path coupling [55], canonical paths [128], correlation decay [210], abstract polymer models [141], zero-free regions [42], spectral independence [11], and stochastic localization [69]. Our results can be thought of as a sampling-to-counting reduction for quantum systems. Thus, we give an algorithmic alternative to directly working with Lindbladians of dissipative evolutions.

Concurrent work on high-temperature Gibbs sampling. In concurrent and independent work, Rouzé, França, and Alhambra [185] prove that the dissipative evolution studied by Chen, Kastoryano, and Gilyén [61] has a constant spectral gap at high temperature, showing that this evolution is an efficient Gibbs sampling algorithm at high temperature. The techniques are significantly different than ours, controlling the evolution by viewing it as a perturbation of an infinite-temperature dissipation. We do not analyze such an evolution.

Cluster expansion and abstract polymer models. Our approach is based on cluster expansion, which allows quantities like the log-partition function and marginals of high-temperature Gibbs states to be expressed as exponentially decaying Taylor series. This tool has been used to show a variety of efficient algorithms around Gibbs

states and real-time evolution, including the computation of partition functions [160], learning of high-temperature Gibbs states [103], and sampling from the measurement distribution of a high-temperature Gibbs state [218].

Among these, the latter sampling result of Yin and Lucas bears the most similarity to our result, using a sampling-to-counting reduction with cluster expansion to give a classical algorithm to sample from the measurement probabilities of a Gibbs state in, say, the computational basis. Our work also implies an efficient algorithm for this task, and achieves a stronger result by tackling the additional challenge of performing these arguments "without measuring".

2.2 Proof Overview

Before delving into the details, we give a high-level overview of the ingredients in the proofs of Theorems 2.1.1 and 2.1.2.

To prove separability, we construct a randomized algorithm which outputs an unnormalized state σ such that, over the randomness of the algorithm, $\mathbb{E}[\sigma] = e^{-\beta H}$. For small enough β, σ is separable, thereby proving that $e^{-\beta H}$ and ρ are separable. Our algorithm is inspired by the classical sampling-to-counting reduction from [129] and can be viewed as a quantum generalization of sampling-to-counting. We start with all $\mathfrak{n}$ sites "unpinned". Then, we sample a separable state on one site according to the marginal of $e^{-\beta H}$, and are left with a posterior-like state on the rest of the sites. We then iterate, sampling a neighboring site according to the marginal of the posterior state, and so on until we have "pinned" all of the sites, at which point we output the resulting state as σ. Notice that the idea of pinning site by site only makes sense if $e^{-\beta H}$ is indeed separable.

The main technical challenge is controlling the effect of the pinned sites on the unpinned sites: if these correlations grow too large, they can push us out of the hull of separable states, which in the algorithm corresponds to the probabilities in the distribution becoming negative. We show that this does not occur via a potential

argument. Pinned sites only affect nearby sites, so for a choice of β sufficiently small in terms of degree, over the course of the algorithm, sites stay close to maximally mixed, and therefore we can pin them without issue.

To turn the proof of separability into a state preparation algorithm, we need to account for normalization: for the separability argument, we were not concerned with mild, constant-factor errors in sampling probabilities, provided that these did not lead to a failure in separability. In particular, since we need to output a quantum state, $\sigma/\operatorname{tr}(\sigma)$, to get $\rho = e^{-\beta H}/\operatorname{tr}(e^{-\beta H})$ in expectation, the probability the algorithm outputs σ is off by a factor of $\operatorname{tr}(\sigma)/\operatorname{tr}(e^{-\beta H})$. This can be fixed by adjusting the sampling probabilities for pinning by an appropriate factor which corresponds to an expectation value of a local observable with respect to ρ; this is the "counting" of the sampling-to-counting reduction. Doing this directly achieves a sub-optimal running time, so we instead use an alternative reduction goes from sampling to weak approximate counting, inspired by a similar algorithm of Sinclair and Jerrum [194].

2.3 Separability of High-Temperature Gibbs States

In this section, we describe a method to write a Gibbs state ρ as a positive linear combination of simple, "configuration" states. For sufficiently small β, these configuration states are positive and separable, thereby proving the separability of the corresponding Gibbs state.

Recall that we work with an $\mathfrak{n}$-qubit low-intersection Hamiltonian $H = \sum_{a\in[m]} H_a$. Though the terms H_a refer to underlying matrices, for the results in this section we will typically be treating these as formal, non-commutating indeterminates.

2.3.1 Low-Degree Polynomial Approximation to a Propagator

We now consider the series expansion of $e^{-\beta H}e^{\beta(H-H^{(\Omega)})}$. Some refer to this operator as a propagator [7], because multiplying by this operator maps a Gibbs state on a restricted set of terms, $e^{-\beta(H-H^{(\Omega)})}$, to the full Gibbs state $e^{-\beta H}$.

This expansion will allow us to understand the marginal of the Gibbs state on one site and its posterior on the remaining sites. This power series will be the centerpiece of the iterative site-by-site pinning process for proving separability.

Theorem 2.3.1 (Propagator series)**.** *Let $H = \sum_{a\in[m]} H_a$ be a Hamiltonian with dual interaction graph $\mathfrak{G}$ with degree $\mathfrak{d}$. Then, for a subset of terms $\mathcal{Q} \subseteq [m]$, we can write*

$$e^{-\beta H} \cdot e^{\beta(H-H^{(\mathcal{Q})})} = \sum_{t=0}^{\infty} \frac{\beta^t}{t!} f_t(H, H^{(\mathcal{Q})})$$

where f_t satisfies the recurrence $f_0(H, H^{(\mathcal{Q})}) = I$ and

$$f_{t+1}(H, H^{(\mathcal{Q})}) = -[H, f_t(H, H^{(\mathcal{Q})})] - f_t(H, H^{(\mathcal{Q})})H^{(\mathcal{Q})}\,. \tag{2.1}$$

Furthermore, let

$$\mathcal{Q}^{(t)} = \Big\{ b \in [m]^t \;\Big|\; \textit{every connected component of } \{b_1, \ldots, b_t\} \textit{ in } \mathfrak{G} \textit{ contains an element of } \mathcal{Q} \Big\}.$$

Then, for each $t > 0$, we can write

$$f_t(H, H^{(\mathcal{Q})}) = \sum_{b\in\mathcal{Q}^{(t)}} \mu_{(b_1,\ldots,b_t)} H_{b_1} \ldots H_{b_t} \textit{ where } \sum_{b\in\mathcal{Q}^{(t)}} |\mu_{(b_1,\ldots,b_t)}| \leqslant \prod_{s=1}^{t} (|\mathcal{Q}| + 2(\mathfrak{d}+1)s). \tag{2.2}$$

Proof. We can expand out the series and write

$$\begin{aligned} e^{-\beta H} e^{\beta(H-H^{(\mathcal{Q})})} &= \Big(\sum_{k=0}^{\infty} \frac{\beta^k (-H)^k}{k!}\Big)\Big(\sum_{\ell=0}^{\infty} \frac{\beta^\ell (H - H^{(\mathcal{Q})})^\ell}{\ell!}\Big) \\ &= \sum_{t=0}^{\infty} \frac{\beta^t}{t!} \sum_{k=0}^{t} \frac{(-H)^k (H - H^{(\mathcal{Q})})^{t-k} t!}{k!(t-k)!} \\ &= \sum_{t=0}^{\infty} \frac{\beta^t}{t!} \underbrace{\sum_{k=0}^{t} \binom{t}{k} (-H)^k (H - H^{(\mathcal{Q})})^{t-k}}_{f_t(H, H^{(\mathcal{Q})})}. \end{aligned} \tag{2.3}$$

Now observe that, because $\binom{t}{k} = \binom{t-1}{k} + \binom{t-1}{k-1}$, $f_0(H, H^{(\mathcal{Q})}) = I$ and $f_t(H, H^{(\mathcal{Q})})$ satisfies the recurrence

$$\begin{aligned} f_t(H, H^{(\mathcal{Q})}) &= -H f_{t-1}(H, H^{(\mathcal{Q})}) + f_{t-1}(H, H^{(\mathcal{Q})})(H - H^{(\mathcal{Q})}) \\ &= -[H, f_{t-1}(H, H^{(\mathcal{Q})})] - f_{t-1}(H, H^{(\mathcal{Q})}) H^{(\mathcal{Q})} \,. \end{aligned} \tag{2.4}$$

This is the recurrence in the statement (2.1). We will prove that f_t satisfies (2.2) by induction. First, we have that $f_1(H, H^{(\mathcal{Q})}) = H^{(\mathcal{Q})}$, which can be expressed as a sum of terms as specified in the theorem statement. For the inductive step, assume that we have proven the desired properties up to some t. Then we can write

$$f_t(H, H^{(\mathcal{Q})}) = \sum_{b \in \mathcal{Q}^{(t)}} \mu_{(b_1,\ldots,b_t)} H_{b_1} \ldots H_{b_t} \,.$$

The recurrence (2.1) implies

$$f_{t+1}(H, H^{(\mathcal{Q})}) = \sum_{b \in \mathcal{Q}^{(t)}} \mu_{(b_1,\ldots,b_t)} (-[H, H_{b_1} \ldots H_{b_t}] - H_{b_1} \ldots H_{b_t} H^{(\mathcal{Q})}) \,.$$

By definition of $\mathcal{Q}^{(t)}$, $\{b_1, \ldots, b_t\}$'s connected components in $\mathfrak{G}$ each contain an element of $\mathcal{Q}$. We now consider the above expression in two parts. First, we can write

$$\begin{aligned} [H, H_{b_1} \ldots H_{b_t}] &= \sum_{a \in [m]} (H_a H_{b_1} \ldots H_{b_t} - H_{b_1} \ldots H_{b_t} H_a) \\ &= \sum_{\substack{a \in [m] \\ (a, b_1, \ldots, b_t) \in \mathcal{Q}^{(t+1)}}} (H_a H_{b_1} \ldots H_{b_t}) - \sum_{\substack{a \in [m] \\ (b_1, \ldots, b_t, a) \in \mathcal{Q}^{(t+1)}}} (H_{b_1} \ldots H_{b_t} H_a) \,. \end{aligned} \tag{2.5}$$

The last step follows from noting that $[H_a, H_{b_1} \ldots H_{b_t}]$ is zero unless the support of H_a intersects that of $H_{b_1} \ldots H_{b_t}$, which can only occur if a is in the neighborhood of $\{b_1, \ldots, b_t\}$. So, $(a, b_1, \ldots, b_t) \in \mathcal{Q}^{(t+1)}$, since adding a does not create a new connected component on $\mathfrak{G}$, and $(a, b_1, \ldots, b_t) \in \mathcal{Q}^{(t+1)}$ if and only if $(b_1, \ldots, b_t, a) \in \mathcal{Q}^{(t+1)}$. Further, there are at most $(\mathfrak{d}+1)t$ such a's, since every b_i has at most $\mathfrak{d}+1$ neighbors,

so in (2.5), there are at most that many summands. A similar argument works for the other part:

$$H_{b_1} \dots H_{b_t} H^{(\mathcal{Q})} = \sum_{a \in \mathcal{Q}} H_{b_1} \dots H_{b_t} H_a, \tag{2.6}$$

and we can conclude $H_{b_1} \dots H_{b_t} H_a$ is an element of $\mathcal{Q}^{(t+1)}$ because $a \in \mathcal{Q}$. Further, there are $|\mathcal{Q}|$ summands. Combining the two parts (2.5) and (2.6), we can write

$$f_{t+1}(H, H^{(\mathcal{Q})}) = \sum_{b' \in \mathcal{Q}^{(t+1)}} \mu_{(b'_1, \dots, b'_{t+1})} H_{b'_1} \dots H_{b'_{t+1}},$$

where, by the above argument and the inductive hypothesis,

$$\sum_{b' \in \mathcal{Q}^{(t+1)}} |\mu_{(b'_1, \dots, b'_{t+1})}| \leqslant \sum_{b \in \mathcal{Q}^{(t)}} |\mu_{(b_1, \dots, b_t)}| (|\mathcal{Q}| + 2(\mathfrak{d}+1)t) \leqslant \prod_{s=1}^{t} (|\mathcal{Q}| + 2(\mathfrak{d}+1)s),$$

which completes the proof. □

In light of Theorem 2.3.1, we define the truncation of this series:

Definition 2.3.2 (Truncated propagator series)**.** For any integer $k \geqslant 0$ and parameter β, we define $T_{k,\beta}(H, H^{(\mathcal{Q})}) = \sum_{t=0}^{k} \frac{\beta^t}{t!} f_t(H, H^{(\mathcal{Q})})$ where f_t is as constructed in Theorem 2.3.1: $f_0(H, H^{(\mathcal{Q})}) = I$ and $f_{t+1}(H, H^{(\mathcal{Q})}) = -[H, f_t(H, H^{(\mathcal{Q})})] - f_t(H, H^{(\mathcal{Q})}) H^{(\mathcal{Q})}$.

We now give an algorithm for sampling a single monomial from one of the polynomials in Theorem 2.3.1 whose expectation is equal to the full polynomial f_t. This subroutine will be useful in our sampling algorithm later on, but our current goal is to prove separability, for which any algorithm suffices, regardless of its efficiency. For this setting, we can think about this algorithm as providing a constructive way to express these polynomials as (positive) linear combinations of simple monomials.

Algorithm 2.3.3 (Sampling a monomial of $f_k(H, H^{(\mathcal{Q})})$)**.**

Input: A Hamiltonian $H = \sum_{a \in [m]} H_a$ with dual interaction graph $\mathfrak{G}$ with degree

$\mathfrak{d}$; subset of terms $\mathcal{Q} \subseteq [m]$.

Input: Integer $k \geqslant 0$.

Output: Scalar $c \in \mathbb{R}$ and terms $(a_1, \ldots, a_k) \in [m]^k$ such that, for $E = H_{a_1} \ldots H_{a_k}$, $\mathbb{E}[cE] = f_k(H, H^{(\mathcal{Q})})$.

Operation:

Initialize $c_0 = 1$, and $b^{(0)} = (,) \in [m]^0$, the empty list;

▷ *The list $b^{(t)}$ implicitly represents the monomial $E_t = H_{b_1^{(t)}} \ldots H_{b_t^{(t)}}$.*

for $t = \{0, 1, \ldots, k-1\}$ **do**

  Flip a coin which is heads with probability $\frac{t}{t+1}$;

  if coin is heads **then**

    Let $\mathcal{R}_t = \{a \in [m] \mid a$ is in the neighborhood of $\{b_s^{(t)}\}_{s \in [t]}$ on $\mathfrak{G}\}$;

    Sample an element $a \in \mathcal{R}_t$ uniformly at random;

    Sample a bit $\xi \in \{0, 1\}$ uniformly at random;

    Set

$$c_{t+1} = \frac{(t+1)2|\mathcal{R}_t|}{t}(-1)^{\xi} c_t \ , \ b^{(t+1)} = \begin{cases} a \cup b^{(t)} & \text{if } \xi = 1 \\ b^{(t)} \cup a & \text{if } \xi = 0 \end{cases};$$

    ▷ *This is an unbiased estimator for $-\frac{t+1}{t}[H, c_t E_t]$ over the randomness of a and ξ; The notation $a \cup v$ and $v \cup a$ denotes prepending and appending to the list v.*

  if coin is tails **then**

    Sample an element $a \in \mathcal{Q}$ uniformly at random;

    Set

$$c_{t+1} = -(t+1)|\mathcal{Q}|c_t \ , \ b^{(t+1)} = b^{(t)} \cup a;$$

    ▷ *This is an unbiased estimator for $-(t+1)c_t E_t H^{(\mathcal{Q})}$ over the randomness of a.*

output $c_k, b^{(k)}$

Lemma 2.3.4 (Unbiased estimator of f_k)**.** *In the same setting as Theorem 2.3.1, for any integer $k \geqslant 0$, if we run Algorithm 2.3.3, then it outputs a $c \in \mathbb{R}$ with $|c| \leqslant (k!) \cdot (\max(2(\mathfrak{d}+1), |\mathcal{Q}|))^k$ and $b \in \mathcal{Q}^{(k)}$ such that, for $E = H_{b_1} \dots H_{b_k}$,*

$$\mathbb{E}[cE] = f_k(H, H^{(\mathcal{Q})}) .$$

Proof. We prove by induction that, for every t over the course of the algorithm, c_t and $b^{(t)}$ satisfy the desired properties of the output: for example, $\mathbb{E}[c_t E_t] = f_t(H, H^{(Q)})$, where $E_t = H_{b_1^{(t)}} \dots H_{b_t^{(t)}}$ is the monomial implicitly maintained by $b^{(t)}$. For the base case of $t = 0$, the algorithm sets $c_0 = 1$ and $E_0 = I$, giving $1 \cdot I = f_0(H, H_{(S)})$. Now we show how to go from t to $t+1$. By the inductive hypothesis, we can assume that after t iterations in the algorithm c_t and E_t satisfy $\mathbb{E}[c_t E_t] = f_t(H, H^{(\mathcal{Q})})$, $|c_t| \leqslant (t!) \cdot (\max(2(\mathfrak{d}+1), |\mathcal{Q}|))^t$, and $E_t \in \mathcal{Q}^{(t)}$. Recall from Theorem 2.3.1 that

$$f_{t+1}(H, H^{(\mathcal{Q})}) = -[H, f_t(H, H^{(\mathcal{Q})})] - f_t(H, H^{(\mathcal{Q})}) H^{(\mathcal{Q})} .$$

In iteration $t+1$ of Algorithm 2.3.3, a coin is flipped, splitting the algorithm into two cases, the output when the coin is heads and the output when the coin is tails.

When the coin lands heads, taking expectation over the randomness in iteration $t+1$,

$$\begin{aligned} \mathop{\mathbb{E}}_{a,\xi}[c_{t+1} E_{t+1} \mid \text{heads}] &= \frac{1}{|\mathcal{R}_t|} \sum_{a \in \mathcal{R}_t} \frac{1}{2} \Big(\frac{(t+1)2|\mathcal{R}_t|}{t} (-c_t H_a E_t + c_t E_t H_a) \Big) \\ &= \frac{t+1}{t} \sum_{a \in \mathcal{R}_t} -[H_a, c_t E_t] \qquad (2.7) \\ &= \frac{t+1}{t} (-[H, c_t E_t]), \end{aligned}$$

where the last step follows from the same argument as for (2.5): the only terms $a \in [m]$ for which $[H_a, E_t]$ is non-zero are the ones which are distance at most one from $b^{(t)}$ on the graph $\mathfrak{G}$. These are the terms in $\mathcal{R}_t$. By a similar argument, $b^{(t+1)} \in \mathcal{Q}^{(t+1)}$.

When the coin lands tails, taking expectation over the randomness in iteration $t+1$,

$$\begin{aligned}\mathop{\mathbb{E}}_{a}[c_{t+1}E_{t+1} \mid \text{tails}] &= \frac{1}{|\mathcal{Q}|}\sum_{a\in\mathcal{Q}} -(t+1)|\mathcal{Q}|c_tE_tH_a \\ &= -(t+1)c_tE_t\sum_{a\in\mathcal{Q}} H_a \\ &= -(t+1)c_tE_tH^{(\mathcal{Q})}\,.\end{aligned} \tag{2.8}$$

In this case, $b^{(t+1)}$ is $b^{(t)}$ with an element of $\mathcal{Q}$ added, and so is in $\mathcal{Q}^{(t+1)}$. The coin lands heads with probability $\frac{t}{t+1}$ and tails with probability $\frac{1}{t+1}$, so putting (2.7) and (2.8) together, we have

$$\mathop{\mathbb{E}}_{\text{iteration } t+1}[c_{t+1}E_{t+1}] = -[H, c_tE_t] - c_tE_tH^{(\mathcal{Q})}\,.$$

By the inductive hypothesis, c_tE_t was drawn from a distribution such that $\mathbb{E}[c_tE_t] = f_t(H, H_{(\mathcal{Q})})$, so $\mathbb{E}[c_{t+1}E_{t+1}] = f_{t+1}(H, H_{(\mathcal{Q})})$ as desired.

To get the bound on c_{t+1}, observe that in the heads case, $|c_{t+1}| = (t+1)2\frac{|\mathcal{R}_t|}{t}|c_t| \leqslant (t+1)2(\mathfrak{d}+1)|c_t|$, where the bound on $|\mathcal{R}_t|$ arises because the number of terms a neighboring some $b_s^{(t)}$ is bounded by $t(\mathfrak{d}+1)$. In the tails case, $|c_{t+1}| = (t+1)|\mathcal{Q}||c_t|$. So, in either case

$$|c_{t+1}| \leqslant (t+1)\max(2(\mathfrak{d}+1), |\mathcal{Q}|)|c_t| \leqslant (t+1)!\max(2(\mathfrak{d}+1), |\mathcal{Q}|)^{t+1},$$

where the last inequality uses the inductive hypothesis. This completes the proof. □

Now that we have a sampler for f_k, we can build a sampler for $e^{-\beta H}e^{\beta(H-H^{(\mathcal{Q})})}$ (and its truncated version $T_{k,\beta}(H, H^{(\mathcal{Q})})$).

Algorithm 2.3.5 (Sampling a monomial of $e^{-\beta H}e^{\beta(H-H^{(\mathcal{Q})})}$)**.**

Input: A Hamiltonian $H = \sum_a H_a$ with dual interaction graph $\mathfrak{G}$ with degree

$\mathfrak{d}$; temperature parameter $\beta \geqslant 0$; subset of terms $\mathcal{Q} \subseteq [m]$; threshold $t_{\max} \in \mathbb{N} \cup \{\infty\}$.

Output: Scalar $c \in \mathbb{R}$ and terms $b = (b_1, \ldots, b_t)$ such that their product $E = H_{b_1} \ldots H_{b_t}$ satisfies $\mathbb{E}[I + cE] = T_{t_{\max},\beta}(H, H_{(S)})$.

Operation:

1: Sample $t \sim \{0, 1, 2, \ldots, t_{\max}\}$ with probabilities $\{2^{-t_{\max}}, 2^{-1}, 2^{-2}, \ldots, 2^{-t_{\max}}\}$;
2: If $t = 0$, set $c' = 0$ and $b = (,)$ the empty list;
3: Otherwise, run Algorithm 2.3.3 on $H, \mathcal{Q}$ with parameter $k \leftarrow t$ to obtain c' and $b = (b_1, \ldots, b_t)$;
4: **output** $c = \frac{2^t \beta^t}{t!} c'$ and $b = (b_1, \ldots, b_t)$;

Lemma 2.3.6 (Unbiased estimator of $T_{t_{\max},\beta}$)**.** *The output of Algorithm 2.3.5, c and $(b_1, \ldots, b_t)$, satisfies for $E = H_{b_1} \ldots H_{b_t}$,*

$$\mathbb{E}[I + cE] = T_{t_{\max},\beta}(H, H^{(\mathcal{Q})}) .$$

Further, if $t = 0$, then $c = 0$, and otherwise, $|c| \leqslant (2\beta \max(2(\mathfrak{d} + 1), |\mathcal{Q}|))^t$.

Proof. By linearity and the guarantees of Lemma 2.3.4,

$$\mathbb{E}[I + cE] = I + \sum_{t=1}^{t_{\max}} \frac{\beta^t}{t!} f_t(H, H^{(\mathcal{Q})}) = T_{t_{\max},\beta}(H, H^{(\mathcal{Q})})$$

where the last step follows from Definition 2.3.2. Again using Lemma 2.3.4,

$$|c| = \frac{2^t \beta^t}{t!} |c'| \leqslant \frac{2^t \beta^t}{t!} (t!)(\max(2(\mathfrak{d} + 1), |\mathcal{Q}|))^t \leqslant (2\beta \cdot \max(2(\mathfrak{d} + 1), |\mathcal{Q}|))^t,$$

giving the coefficient bound. □

None of this argument so far has used a bound on β. However, we want our estimator $I + cE$ to always be a positive semi-definite matrix, which we can only guarantee if

$|c| \leqslant 1$ always. So, this algorithm will only be useful when $\beta < \frac{1}{2\max(2(\mathfrak{d}+1),|\mathcal{Q}|)}$.

2.3.2 Additional Structural Properties of Restricted Gibbs States

In this section, we prove additional lemmas which are necessary for the state preparation algorithm in Section 2.4. This can safely be skipped to get to the separability proof.

In order to convert the algorithm in Algorithm 2.3.13 to one that samples from the Gibbs state $e^{-\beta H}/\operatorname{tr} e^{-\beta H}$, we need to account for the normalization by the partition function $\operatorname{tr} e^{-\beta H}$, as well as the error from not computing the entire series (which arises in the choice of ε in Algorithm 2.3.11). We now prove some structural lemmas about Gibbs states which address these issues.

We first prove that the Gibbs state corresponding to $H - H_{(S)}$ has roughly the same spectrum as the Gibbs state corresponding to H, where S is a small subset of sites.

Lemma 2.3.7 (Spectrum of restricted Hamiltonians)**.** *Let $H = \sum_{a\in[m]} H_a$ be a Hamiltonian with dual interaction graph $\mathfrak{G}$ with degree $\mathfrak{d}$. Let $S = \operatorname{supp}(H_{a^*})$ for some $a^* \in [m]$. If $\beta < \frac{1}{2C(\mathfrak{d}+1)}$ for some constant $C > 5$ then*

$$\Big(1 - \frac{15}{C}\Big)e^{-\beta H} \preceq e^{-\beta(H-H_{(S)})} \preceq \Big(1 + \frac{15}{C}\Big)e^{-\beta H}.$$

Proof. By Theorem 2.3.1, we can write

$$e^{\frac{\beta}{2}H}e^{-\frac{\beta}{2}(H-H_{(S)})} = \sum_{t=0}^{\infty} \frac{(-\beta/2)^t}{t!} f_t(H, H_{(S)}),$$

where $\|f_t(H, H_{(S)})\|_{\mathrm{op}} \leqslant \prod_{s=1}^{t}((\mathfrak{d}+1)(2s+1))$, using that the number of terms in $H_{(S)}$ is at most $\mathfrak{d}+1$. So, there is a matrix A such that

$$e^{\frac{\beta}{2}H}e^{-\frac{\beta}{2}(H-H_{(S)})} = (I + A) \text{ with } \|A\|_{\mathrm{op}} \leqslant \sum_{t=1}^{\infty} (\beta(\mathfrak{d}+1))^t \frac{\prod_{s=1}^{t}(2s+1)}{2^t(t!)} \leqslant \frac{5}{C}.$$

Multiplying this by its conjugate, we have

$$e^{\frac{\beta}{2}H} \cdot e^{-\beta(H-H_{(S)})} \cdot e^{\frac{\beta}{2}H} = (I+A)(I+A)^\dagger$$

We are done upon concluding that, from the bound on $\|A\|_{\mathrm{op}}$ and $C > 5$, $\|(I+A)(I+A^\dagger) - I\|_{\mathrm{op}} \leqslant 2\|A\|_{\mathrm{op}} + \|A\|_{\mathrm{op}}^2 \leqslant 15/C$, so

$$\Big(1 - \frac{15}{C}\Big) I \preceq (I+A)(I+A)^\dagger \preceq \Big(1 + \frac{15}{C}\Big) I. \qquad \square$$

Next, we prove a sharper statement. Recalling the construction of $T_{t,\beta/2}(H, H^{(\mathcal{Q})})$ in Definition 2.3.2, we show that left and right multiplying a matrix by $e^{-\frac{\beta}{2}H}$ is close to left and right multiplying by $e^{-\frac{\beta}{2}(H-H^{(\mathcal{Q})})} T_{t,\beta/2}(H, H^{(\mathcal{Q})})^\dagger$ and its Hermitian conjugate.

Lemma 2.3.8 (Peeling the restricted Gibbs state)**.** *Let $H = \sum_{a\in[m]} H_a$ be a $\mathfrak{K}$-local Hamiltonian with dual interaction graph with degree $\mathfrak{d}$. Let $S = \mathrm{supp}(a^*)$ for some $a^* \in [m]$. Let P be a $2^n \times 2^n$ Hermitian matrix such that $0.5I \preceq P \preceq 2I$ and let $t \geqslant 0$ be an integer. Given any $\beta < \frac{1}{2C(\mathfrak{d}+1)}$ for some constant $C > 5$, we have*

$$\Big(1 - \frac{100}{C^t}\Big) e^{-\frac{\beta}{2}H} P e^{-\frac{\beta}{2}H} \preceq e^{-\frac{\beta}{2}(H-H_{(S)})} \cdot T_{t,\beta/2}(H, H_{(S)})^\dagger \cdot P \cdot T_{t,\beta/2}(H, H_{(S)}) \cdot e^{-\frac{\beta}{2}(H-H_{(S)})} \preceq \Big(1 + \frac{100}{C^t}\Big) e^{-\frac{\beta}{2}H} P e^{-\frac{\beta}{2}H},$$

where $T_{t,\beta/2}$ is the truncation defined in Definition 2.3.2.

Proof. It follows from Theorem 2.3.1 that

$$e^{-\frac{\beta}{2}H} \cdot e^{\frac{\beta}{2}(H-H_{(S)})} = T_{t,\beta/2}(H, H_{(S)}) + E$$

for some $E \in \mathbb{C}^{2^n \times 2^n}$ such that

$$\|E\|_{\mathrm{op}} \leqslant \sum_{s=t+1}^{\infty} (\beta(\mathfrak{d}+1))^s \frac{\prod_{r=1}^{s}(2r+1)}{2^s(s!)} \leqslant \frac{5}{C^{t+1}}.$$

We consider the expression

$$A = P - e^{\frac{\beta}{2}H} \cdot e^{-\frac{\beta}{2}(H-H_{(S)})} \cdot T_{t,\beta/2}(H, H_{(j)})^\dagger \cdot P \cdot T_{t,\beta/2}(H, H_{(j)}) \cdot e^{-\frac{\beta}{2}(H-H_{(S)})} \cdot e^{\frac{\beta}{2}H}$$

The lemma statement is equivalent to the statement that $-\frac{100}{C^t}P \preceq A \preceq \frac{100}{C^t}P$. We continue:

$$\begin{aligned} A &= P - (I - e^{\frac{\beta}{2}H} \cdot e^{-\frac{\beta}{2}(H-H_{(S)})} \cdot E^\dagger) \cdot P \cdot (I - E \cdot e^{-\frac{\beta}{2}(H-H_{(S)})} \cdot e^{\frac{\beta}{2}H}) \\ &= e^{\frac{\beta}{2}H} \cdot e^{-\frac{\beta}{2}(H-H_{(S)})} \cdot E^\dagger \cdot P + P \cdot E \cdot e^{-\frac{\beta}{2}(H-H_{(S)})} \cdot e^{\frac{\beta}{2}H} \\ &\quad - e^{\frac{\beta}{2}H} \cdot e^{-\frac{\beta}{2}(H-H_{(S)})} \cdot E^\dagger \cdot P \cdot E \cdot e^{-\frac{\beta}{2}(H-H_{(S)})} \cdot e^{\frac{\beta}{2}H} \end{aligned} \tag{2.9}$$

Lemma 2.3.7 implies that $\|e^{\frac{\beta}{2}H} e^{-\frac{\beta}{2}(H-H_{(S)})}\|_{\text{op}} \leqslant 4$. Using this, along with the bounds on $\|E\|_{\text{op}}$ and $\|P\|_{\text{op}}$, we have that

$$\begin{aligned} \|A\|_{\text{op}} &\leqslant 2\|e^{\frac{\beta}{2}H} e^{-\frac{\beta}{2}(H-H_{(S)})}\|_{\text{op}} \|E\|_{\text{op}} \|P\|_{\text{op}} + \|e^{\frac{\beta}{2}H} e^{-\frac{\beta}{2}(H-H_{(S)})}\|_{\text{op}}^2 \|E\|_{\text{op}}^2 \|P\|_{\text{op}} \\ &\leqslant \frac{80}{C^{t+1}} + \frac{800}{C^{2t+2}} \leqslant \frac{50}{C^t}. \end{aligned}$$

Thus,

$$-\frac{50}{C^t} I \preceq A \preceq \frac{50}{C^t} I.$$

We get the desired relations upon using that $\frac{1}{2}I \preceq P \preceq 2I$. □

2.3.3 Proof of Separability

We now use our sampling procedure to prove separability at high temperature. Our strategy is to show that $e^{-\beta H}$ can be written as a convex combination of simple matrices, which we call *configurations*, all of which are positive semi-definite and separable. Moreover, our proof is constructive, giving an algorithm for sampling configurations according to their weight in the combination. We now define the objects our algorithm works with, which we call Hermitian monomials.

Definition 2.3.9 (Hermitian monomial)**.** For a Hamiltonian $H = \sum_{a=1}^m H_a$, we call

an operator X a *Hermitian monomial* of (the terms of) H if we can form it by starting from $X = I$ and iteratively performing the following "multiply and symmetrize" operation: given $a_1, \ldots, a_s$ and $b_1, \ldots, b_t$, let

$$X \leftarrow \frac{1}{2}\Big(H_{a_1} \ldots H_{a_s} X H_{b_t} \ldots H_{b_1} + H_{b_1} \ldots H_{b_t} X H_{a_s} \ldots H_{a_1}\Big).$$

We can maintain a description of a Hermitian monomial as a list of operations used to construct it. A Hermitian monomial X can potentially be constructed in different ways, but we will abuse notation and use X to refer to the matrix X alongside an associated description. A description of a Hermitian monomial X has an associated multiset of terms $\{a_1, \ldots, a_t\}$, which are the H_a's which appear in the operations used to construct X. For example, the Hermitian monomial $\frac{1}{2}(H_1 H_1 H_2 + H_2 H_1 H_1)$ has the associated multiset $\{1, 1, 2\}$. We call t the *degree* of the monomial.[†]

A configuration is a tensor product of matrices $I + cX$, where X is a Hermitian monomial.

Definition 2.3.10 (Configuration)**.** For a Hamiltonian $H = \sum_{a=1}^{m} H_a$, a *configuration* is a collection $\mathcal{X} = \{(c_1, X_1), \ldots, (c_\ell, X_\ell)\}$ such that $c_i \in \mathbb{R}$, X_i is a Hermitian monomial for every $i \in [\ell]$, and the X_i's have disjoint supports. This configuration defines an associated operator $\sigma(\mathcal{X})$ which is $I + c_i X_i$ on the support of X_i and identity outside of the supports:

$$\sigma(\mathcal{X}) = (I)_{[\mathfrak{n}] \setminus (\mathrm{supp}(X_1) \cup \cdots \cup \mathrm{supp}(X_\ell))} \otimes \Big(\bigotimes_{i=1}^{\ell} (I + c_i X_i)_{\mathrm{supp}(X_i)}\Big).$$

[†]It is important to note that, for our algorithm, we will primarily be working with the *description* of the monomial, and not the actual matrix X. Again, a matrix can have multiple different descriptions: for example, if $H_1^2 = I$, then I can be described with a multiset of $\{1, 1\}$, and with that description it would have a degree of 2. This ambiguity will not cause issues, though; throughout, we will treat the H_a's as essentially formal non-commutative indeterminates without any relations, and only at the end of the proof interpret these indeterminates as matrices.

Algorithm 2.3.11 (Pinning a single term)**.**

Input: Hamiltonian $H = \sum_a H_a$ with locality $\mathfrak{K}$ and dual interaction graph $\mathfrak{G}$ with degree $\mathfrak{d}$; inverse temperature parameter β such that $\beta \leqslant \beta_c = 1/(100\mathfrak{d}\mathfrak{K})$; accuracy parameter $\varepsilon \geqslant 0$.

Input: Set of sites $S \subseteq [\mathfrak{n}]$, configuration $\mathcal{X} = \{(c_1, X_1), \ldots, (c_\ell, X_\ell)\}$ satisfying (2.10).

Output: Set of sites $\widehat{S} \subseteq [\mathfrak{n}]$, configuration $\widehat{\mathcal{X}} = \{(\widehat{c}_1, \widehat{X}_1), \ldots, (\widehat{c}_{\widehat{\ell}}, \widehat{X}_{\widehat{\ell}})\}$ satisfying (2.10).

Operation:

1: Let $t_{\max} = 10 \log(\mathfrak{n}/\varepsilon)$, or ∞ if $\varepsilon = 0$;

2: Let $\widehat{\ell} = \ell$:

3: Let $\{a_1, \ldots, a_t\}$ be the terms associated to (c_ℓ, X_ℓ), and choose an $a^* \in \mathcal{E}^{(S)}$ neighboring $\{a_1, \ldots, a_t\}$ in $\mathfrak{G}$; ▷ *For simplicity, these choices should be deterministic.*

4: **if** no such a^* exists **then**

5: Add $(c_{\ell+1}, X_{\ell+1}) = (0, I)$ to the configuration $\mathcal{X}$, and choose any $a^* \in \mathcal{E}^{(S)}$;

6: Set $\widehat{\ell} \leftarrow \ell + 1$;

7: Run Algorithm 2.3.5 twice (independently) with inputs $H \leftarrow H^{(S)}$, $\beta \leftarrow \beta/2$, $\mathcal{Q} \leftarrow \mathcal{E}_{(a^*)} \cap \mathcal{E}^{(S)}$, and $t_{\max}$ to obtain b_1, E_1, t_1 and b_2, E_2, t_2;

8: Let $\gamma = \frac{3}{5\mathfrak{K}}$;

9: Sample $\xi \in \{0, 1, 2, 3, 4, 5, 6\}$ with probabilities $\{1 - \gamma, \frac{\gamma}{6}, \frac{\gamma}{6}, \frac{\gamma}{6}, \frac{\gamma}{6}, \frac{\gamma}{6}, \frac{\gamma}{6}\}$.

10: Set $c_{\widehat{\ell}} \leftarrow \widehat{c}$ and $X_{\widehat{\ell}} \leftarrow \widehat{X}$, where

- If $\xi = 0$, then $\widehat{c} = c_{\widehat{\ell}}/(1-\gamma)$ and $\widehat{X} = X_{\widehat{\ell}}$;
- If $\xi = 1$, then $\widehat{c} = 6b_1/\gamma$ and $\widehat{X} = (E_1^\dagger + E_1)/2$;
- If $\xi = 2$, then $\widehat{c} = 6b_2/\gamma$ and $\widehat{X} = (E_2^\dagger + E_2)/2$;
- If $\xi = 3$, then $\widehat{c} = 6c_{\widehat{\ell}} b_1/\gamma$ and $\widehat{X} = (E_1^\dagger X_{\widehat{\ell}} + X_{\widehat{\ell}} E_1)/2$;
- If $\xi = 4$, then $\widehat{c} = 6c_{\widehat{\ell}} b_2/\gamma$ and $\widehat{X} = (E_2^\dagger X_{\widehat{\ell}} + X_{\widehat{\ell}} E_2)/2$;
- If $\xi = 5$, then $\widehat{c} = 6b_1 b_2/\gamma$ and $\widehat{X} = (E_2^\dagger E_1 + E_1^\dagger E_2)/2$;
- If $\xi = 6$, then $\widehat{c} = 6c_{\widehat{\ell}} b_1 b_2/\gamma$ and $\widehat{X} = (E_2^\dagger X_{\widehat{\ell}} E_1 + E_1^\dagger X_{\widehat{\ell}} E_2)/2$;

▷ *This is chosen so that* $\mathbb{E}_\xi[I + \widehat{c}\widehat{X}]$ *takes a particular form* (2.11).

11: Set $S \leftarrow S \setminus \mathrm{supp}(H_{a^*})$;

12: **output** $S, \mathcal{X}$;

To analyze Algorithm 2.3.11, we will need the following arithmetic fact. On the left-hand side are the operators which appear in our analysis, and on the right-hand side are Hermitian monomials.

Fact 2.3.12. *For matrices* X*,* Y_1*, and* Y_2*, we can write*

$$\begin{aligned}(I+Y_1)^\dagger(I+X)(I+Y_2) + (I+Y_2)^\dagger(I+X)(I+Y_1) \\ = 2I + 2X + (Y_1^\dagger + Y_1) + (Y_2^\dagger + Y_2) + (Y_1^\dagger X + XY_1) + (Y_2^\dagger X + XY_2) \\ + (Y_2^\dagger Y_1 + Y_1^\dagger Y_2) + (Y_2^\dagger X Y_1 + Y_1^\dagger X Y_2)\,.\end{aligned}$$

Algorithm 2.3.13 (Sampling a monomial of $e^{-\beta H}$)**.**

Input: Hamiltonian $H = \sum_a H_a$ with dual interaction graph $\mathfrak{G}$ with degree $\mathfrak{d}$; inverse temperature parameter β such that $\beta \leqslant \beta_c = 1/(100\mathfrak{d}\mathfrak{K})$.

Output: Configuration $\mathcal{X}$ such that $\mathbb{E}[\sigma(\mathcal{X})] = e^{-\beta H}$.

Operation:

1: Set $S = [\mathfrak{n}]$;

2: Set $\mathcal{X} = \varnothing$;

3: ▷ *We maintain the invariant that* $\mathbb{E}[e^{-\frac{\beta}{2}H^{(S)}}\sigma(\mathcal{X})e^{-\frac{\beta}{2}H^{(S)}}] = e^{-\beta H}$. ◁

4: **while** $\mathcal{E}^{(S)} \neq \varnothing$ **do**

5: Run Algorithm 2.3.11 with S and $\mathcal{X}$ and $\varepsilon = 0$;

6: Take the output $\widehat{S}$ and $\widehat{\mathcal{X}}$ to be the new S and $\mathcal{X}$;

7: **output** $\mathcal{X}$;

The set S corresponds to "unpinned sites", the set of sites which have not yet been sampled. For the algorithm to work, S must only interact with at most one of the monomials in the configuration, in the manner described in the following lemma.

Lemma 2.3.14 (Algorithm 2.3.11 is well-defined)**.** *Algorithm 2.3.11 is a well-defined procedure provided* $\mathcal{E}^{(S)} \neq \varnothing$*. Further, it maintains the following invariant: if the input* S *and* $\mathcal{X}$ *satisfies that*

$$\text{for all } i \in [\ell - 1], \text{ no term in } X_i \text{ is a neighbor of } \mathcal{E}^{(S)}, \tag{2.10}$$

then the output $\widehat{S}$ *and* $\widehat{\mathcal{X}}$ *satisfies the same condition. Finally, the output* $\widehat{\mathcal{X}}$ *is a valid configuration.*

Proof. The algorithm successfully completes provided there is an element of $\mathcal{E}^{(S)}$ so that a valid a^* can be chosen.

For the invariant, we need to show that, for all $i \in [\widehat{\ell} - 1]$, no term in $\widehat{X}_i$ is a neighbor of $\mathcal{E}^{(\widehat{S})}$. First, observe that $\widehat{\mathcal{X}}$ only differs from $\mathcal{X}$ in the last-indexed monomial in the configuration, which is either $i = \ell$ or $i = \ell + 1$. Further, $\widehat{S} \subset S$, so $\mathcal{E}^{(\widehat{S})} \subseteq \mathcal{E}^{(S)}$, and so the invariant is maintained for $i < \ell$. This proves the statement when $\widehat{\ell} = \ell$, and the remaining case is when $\widehat{\ell} = \ell + 1$, where we need to show that no term in $\widehat{X}_\ell$ is a neighbor of $\mathcal{E}^{(\widehat{S})}$. This case only occurs when there are no such neighbors, as seen in Line 3, proving the invariant.

The output $\mathcal{X}$ is a valid configuration because the input is a configuration, and $\widehat{\mathcal{X}}$ only modifies $\mathcal{X}$ by adding terms from $\mathcal{E}^{(S)}$ to monomials ℓ or $\ell+1$. From the invariant, these terms have disjoint support from all other terms. Since E_1 and E_2 are products of H_a's, $\widehat{X}$ is a Hermitian monomial when X is, by definition (Definition 2.3.9). $\square$

Lemma 2.3.15. *Algorithm 2.3.11 outputs a $\widehat{S} \subseteq [\mathfrak{n}]$ and configuration $\widehat{\mathcal{X}}$ such that*

$$\mathbb{E}[\sigma(\widehat{\mathcal{X}})] = T_{t_{\max},\beta/2}(H^{(S)}, H^{(S)}_{(S\setminus\widehat{S})})^\dagger \sigma(\mathcal{X}) T_{t_{\max},\beta/2}(H^{(S)}, H^{(S)}_{(S\setminus\widehat{S})})\,.$$

In particular, when $\varepsilon = 0$,

$$e^{-\frac{\beta}{2}H^{(\widehat{S})}}\,\mathbb{E}[\sigma(\widehat{\mathcal{X}})]e^{-\frac{\beta}{2}H^{(\widehat{S})}} = e^{-\frac{\beta}{2}H^{(S)}}\sigma(\mathcal{X})e^{-\frac{\beta}{2}H^{(S)}}\,.$$

Proof. Using linearity of expectation and Fact 2.3.12, we can conclude that, with $\widehat{c}$ and $\widehat{X}$ as defined in Line 10 of the algorithm,

$$\begin{aligned}\mathop{\mathbb{E}}_{\xi}[I+\widehat{c}\widehat{X}] &= I + c_{\widehat{\ell}}X_{\widehat{\ell}} + b_1(E_1^\dagger + E_1)/2 + b_2(E_2^\dagger+E_2)/2 + c_{\widehat{\ell}}b_1(E_1^\dagger X_{\widehat{\ell}} + X_{\widehat{\ell}}E_1)/2 \\ &\quad + c_{\widehat{\ell}}b_2(E_2^\dagger X_{\widehat{\ell}} + X_{\widehat{\ell}}E_2)/2 + b_1b_2(E_2^\dagger E_1 + E_1^\dagger E_2)/2 + c_{\widehat{\ell}}b_1b_2(E_2^\dagger X_{\widehat{\ell}}E_1 + E_1^\dagger X_{\widehat{\ell}}E_2)/2 \\ &= \frac{1}{2}\Big((I+b_1E_1)^\dagger(I+c_{\widehat{\ell}}X_{\widehat{\ell}})(I+b_2E_2) + (I+b_2E_2)^\dagger(I+c_{\widehat{\ell}}X_{\widehat{\ell}})(I+b_1E_1)\Big)\,.\end{aligned} \tag{2.11}$$

Lemma 2.3.6 implies that the outputs of Algorithm 2.3.5 in Line 7 satisfy

$$\mathbb{E}[I+b_1E_1] = \mathbb{E}[I+b_2E_2] = T_{t_{\max},\beta/2}(H^{(S)}, H^{(S)}_{(a^*)})\,,$$

where $H^{(S)}_{(a^*)}$ is the sum of all terms H_a where $\mathrm{supp}(H_a) \subseteq S$ and $\mathrm{supp}(H_a) \cap \mathrm{supp}(H_{a^*})$ is non-empty. So, taking the expectation over the entire algorithm and using that $I + b_1E_1$ and $I + b_2E_2$ are independent,

$$\begin{aligned}\mathbb{E}[I+\widehat{c}\widehat{X}] &= \mathbb{E}\Big[\frac{1}{2}\Big((I+b_1E_1)^\dagger(I+c_{\widehat{\ell}}X_{\widehat{\ell}})(I+b_2E_2) + (I+b_2E_2)^\dagger(I+c_{\widehat{\ell}}X_{\widehat{\ell}})(I+b_1E_1)\Big)\Big] \\ &= T_{t_{\max},\beta/2}(H^{(S)}, H^{(S)}_{(a^*)})^\dagger (I+c_{\widehat{\ell}}X_{\widehat{\ell}}) T_{t_{\max},\beta/2}(H^{(S)}, H^{(S)}_{(a^*)}).\end{aligned}$$

By Lemma 2.3.14, the support of $I + \widehat{c}\widehat{X}$ is always disjoint from all of the other monomials in the configuration $\widehat{\mathcal{X}}$, so

$$\mathbb{E}[\sigma(\widehat{\mathcal{X}})] = T_{t_{\max},\beta/2}(H^{(S)}, H^{(S)}_{(a^*)})^\dagger \sigma(\mathcal{X}) T_{t_{\max},\beta/2}(H^{(S)}, H^{(S)}_{(a^*)}).$$

When $\varepsilon = 0$, this can be written as

$$\begin{aligned}\mathbb{E}[\sigma(\widehat{\mathcal{X}})] &= e^{\frac{\beta}{2}(H^{(S)} - H^{(S)}_{(a^*)})} e^{-\frac{\beta}{2}H^{(S)}} \sigma(\mathcal{X}) e^{-\frac{\beta}{2}H^{(S)}} e^{\frac{\beta}{2}(H^{(S)} - H^{(S)}_{(a^*)})} \\ &= e^{\frac{\beta}{2}H^{(\widehat{S})}} e^{-\frac{\beta}{2}H^{(S)}} \sigma(\mathcal{X}) e^{-\frac{\beta}{2}H^{(S)}} e^{\frac{\beta}{2}H^{(\widehat{S})}},\end{aligned}$$

since $\widehat{S} = S \setminus \operatorname{supp}(H_{a^*})$. □

What we have so far is enough to show that Algorithm 2.3.13 indeed outputs a configuration $\mathcal{X}$ with $\mathbb{E}[\sigma(\mathcal{X})] = e^{-\beta H}$. However, for separability, we need more: namely, we need that $\sigma(\mathcal{X})$ is always positive semi-definite and separable. We handle the former with the following lemma. This is the key lemma where the critical temperature is used. It is nontrivial to show, but we are able to prove it via a carefully chosen potential function.

Lemma 2.3.16 (Coefficients remain small)**.** *Let $H = \sum_a H_a$ be a $\mathfrak{K}$-local Hamiltonian and let the temperature parameter β satisfy $\beta \leqslant \beta_c = 1/(50(\mathfrak{d}+1)\mathfrak{K})$, where $\mathfrak{d}$ is the degree of the dual interaction graph of H. Consider running Algorithm 2.3.11 on the input S and $\mathcal{X}$, and receiving $\widehat{S}$ and $\widehat{\mathcal{X}}$ as output. Then $|\widehat{S}| < |S|$. Further, consider the following invariant on the input:*

$$\begin{aligned}&\text{for all } (c, X) \in \mathcal{X},\ |c| \leqslant (1-\gamma)^{|S \cap (\operatorname{supp}(H_{a_1}) \cup \cdots \cup \operatorname{supp}(H_{a_t}))|} \Big(\frac{\beta}{\beta_c}\Big)^t \\ &\quad \text{where } \{a_1, \ldots, a_t\} \text{ are the terms associated to the Hermitian monomial } X\,. \end{aligned} \tag{2.12}$$

If the input satisfies (2.12)*, then so does the output:*

$$|\widehat{c}| \leqslant (1-\gamma)^{|\widehat{S} \cap (\operatorname{supp}(H_{\widehat{a}_1}) \cup \cdots \cup \operatorname{supp}(H_{\widehat{a}_{\widehat{t}}}))|} \Big(\frac{\beta}{\beta_c}\Big)^{\widehat{t}} \tag{2.13}$$

Proof. Because $\widehat{S} = S \setminus \operatorname{supp}(H_{a^*})$ and the manner in which a^* is chosen, $|\widehat{S}| < |S|$.

To prove the inductive invariant, it suffices to consider the term $(\widehat{c}_{\widehat{\ell}}, \widehat{X}_{\widehat{\ell}})$, which we denote $(\widehat{c}, \widehat{X})$ in the algorithm, since this is the only term that Algorithm 2.3.11 changes: the invariant will continue to hold provided that the term does not change. We first note that, in Line 7, Algorithm 2.3.5 is always run for $\beta \leftarrow \beta/2$ and $\mathcal{Q} \leftarrow \mathcal{E}_{(a^*)} \cap \mathcal{E}^{(S)}$ (see Definition 1.4.17). By definition of the dual interaction graph, $|\mathcal{Q}| \leqslant \mathfrak{d} + 1$. So, by Lemma 2.3.6,

$$\begin{aligned} &b_1 = 0 \text{ and } t_1 = 0 \text{ or } |b_1| \leqslant (\beta \cdot 2(\mathfrak{d}+1))^{t_1}, \text{ and} \\ &b_2 = 0 \text{ and } t_2 = 0 \text{ or } |b_2| \leqslant (\beta \cdot 2(\mathfrak{d}+1))^{t_2} \end{aligned}$$

We now show that $\widehat{c}$ is small for every choice of ξ. For this proof, we denote $c = c_{\widehat{\ell}}$ and $X = X_{\widehat{\ell}}$ to be the term which is to be changed. Further, we will abuse notation and, for a Hermitian monomial X with associated terms $\{a_1, \dots, a_t\}$, write $\operatorname{supp}(X) = \operatorname{supp}(H_{a_1}) \cup \cdots \cup \operatorname{supp}(H_{a_t})$. This set contains the "true" support of X, but can be larger in general. So, we need to show that, assuming the invariant (2.12) which states that $|c| \leqslant (1-\gamma)^{|S \cap \operatorname{supp}(X)|}(\frac{\beta}{\beta_c})^t$, then $|\widehat{c}| \leqslant (1-\gamma)^{|\widehat{S} \cap \operatorname{supp}(\widehat{X})|}(\frac{\beta}{\beta_c})^{\widehat{t}}$. Note that, by our choice of β_c, $10\mathfrak{K}\beta_c \cdot 2(\mathfrak{d}+1) \leqslant \frac{2}{5} = 1 - \mathfrak{K}\gamma$.

Case $\xi = 0$: Here, $\widehat{X} = X$, the associated terms are identical ($\widehat{a}_s = a_s$ for all $s \in [t]$) and $\widehat{c} = c/(1-\gamma)$. Then either $c = 0$, and the bound holds trivially, or a^* was chosen to neighbor some a_s associated to X. In this case, the scalar increases; however, this is accounted for in the potential, because $\operatorname{supp}(H_{a^*})$ is removed from S to get $\widehat{S}$:

$$|\widehat{c}| = |c|/(1-\gamma) \leqslant (1-\gamma)^{|S \cap \operatorname{supp}(X)|-1}\Big(\frac{\beta}{\beta_c}\Big)^t \leqslant (1-\gamma)^{|\widehat{S} \cap \operatorname{supp}(X)|}\Big(\frac{\beta}{\beta_c}\Big)^t.$$

The first inequality uses the induction hypothesis; the second uses that $\widehat{S} \subseteq S$.

Case $\xi = 1$, $\xi = 2$, $\xi = 5$: We first work out the $\xi = 1$ case. Here, $\widehat{X} = (E_1^\dagger + E_1)/2$, and so the monomial has degree $\widehat{t} = t_1$. If $t_1 = 0$, then $b_1 = 0$ and we are done.

Otherwise,

$$\begin{aligned}|\widehat{c}| = 6|b_1|/\gamma &\leqslant 10\mathfrak{K}(\beta \cdot 2(\mathfrak{d}+1))^{t_1} \leqslant (10\mathfrak{K}\beta_c \cdot 2(\mathfrak{d}+1))^{t_1}\Big(\frac{\beta}{\beta_c}\Big)^{t_1} \\ &\leqslant (1-\mathfrak{K}\gamma)^{t_1}\Big(\frac{\beta}{\beta_c}\Big)^{t_1} \leqslant (1-\gamma)^{\mathfrak{K}t_1}\Big(\frac{\beta}{\beta_c}\Big)^{t_1} \leqslant (1-\gamma)^{|\widehat{S}\cap \mathrm{supp}(\widehat{X})|}\Big(\frac{\beta}{\beta_c}\Big)^{\widehat{t}}.\end{aligned}$$

The $\xi = 2$ case is analogous, switching out E_1 and t_1 for E_2 and t_2. The $\xi = 5$ case is also analogous, with $E_1^\dagger E_2$ and $\widehat{t} = t_1 + t_2$.

Case $\xi = 3$, $\xi = 4$, $\xi = 6$: We first work out the $\xi = 3$ case. Here, $\widehat{X} = (E_1^\dagger X + X E_1)/2$, so the monomial has degree $t + t_1$, where t is the degree of the original monomial X, and the associated terms are the terms from X, along with t_1 additional terms from E_1. If $t_1 = 0$ then $\widehat{c} = 6cb_1/\gamma = 0$ so we are done. Otherwise,

$$\begin{aligned}|\widehat{c}| &= 6|c||b_1|/\gamma \\ &\leqslant 10\mathfrak{K}(1-\gamma)^{|S\cap \mathrm{supp}(X)|}\Big(\frac{\beta}{\beta_c}\Big)^{t}(\beta \cdot 2(\mathfrak{d}+1))^{t_1} \\ &\leqslant (1-\gamma)^{|S\cap \mathrm{supp}(X)|+\mathfrak{K}t_1}\Big(\frac{\beta}{\beta_c}\Big)^{t+t_1} \\ &\leqslant (1-\gamma)^{|\widehat{S}\cap \mathrm{supp}(\widehat{X})|}\Big(\frac{\beta}{\beta_c}\Big)^{\widehat{t}}.\end{aligned}$$

The $\xi = 4$ and $\xi = 6$ cases are analogous, in the same manner as the $\xi = 2$ and $\xi = 5$ cases. □

Theorem 2.3.17. *Let $H = \sum_a H_a$ be a Hamiltonian with locality $\mathfrak{K}$ and dual interaction graph $\mathfrak{G}$ with degree $\mathfrak{d}$, and let the temperature parameter β satisfy $\beta \leqslant \beta_c = 1/(100\mathfrak{K}\mathfrak{d})$. Then Algorithm 2.3.13 outputs a configuration $\mathcal{X}$ such that $\mathbb{E}[\sigma(\mathcal{X})] = e^{-\beta H}$ and, for all $(c, X) \in \mathcal{X}$, $|c| \leqslant (\beta/\beta_c)^t$, where t is the degree of the Hermitian monomial X.*

Proof. We show that through the while loop, Algorithm 2.3.13 maintains the two

invariants that

$$\mathbb{E}[e^{-\frac{\beta}{2}H^{(S)}}\sigma(\mathcal{X})e^{-\frac{\beta}{2}H^{(S)}}] = e^{-\beta H}$$

and, from Eq. (2.12), for all $(c, X) \in \mathcal{X}$ with degree t,

$$|c| \leqslant (1-\gamma)^{|S\cap \mathrm{supp}(X)|}\Big(\frac{\beta}{\beta_c}\Big)^t.$$

(We use the same notation as in Lemma 2.3.16, where $\mathrm{supp}(X)$ denotes the support of the terms associated to X.) Since the initialization is that $S = [\mathfrak{n}]$ and $\mathcal{X} = \varnothing$, so that $\sigma(\mathcal{X}) = I$, both invariants hold before the while loop begins. Lemma 2.3.15 implies that the first invariant is maintained, since the output of Algorithm 2.3.11 satisfies

$$e^{-\frac{\beta}{2}H^{(\widehat{S})}} \mathop{\mathbb{E}}_{\text{iteration}} [\sigma(\widehat{\mathcal{X}})] e^{-\frac{\beta}{2}H^{(\widehat{S})}} = e^{-\frac{\beta}{2}H^{(S)}}\sigma(\mathcal{X})e^{-\frac{\beta}{2}H^{(S)}},$$

where the randomness here is taken only over the call to the algorithm in Line 5. Lemma 2.3.16 implies that the second invariant is maintained. Finally, by Lemma 2.3.16, the size of S decreases every iteration, so the algorithm terminates after at most $\mathfrak{n}$ iterations. Upon termination, $S = \varnothing$, so the invariants imply that, for the output configuration $\mathcal{X}$, $\mathbb{E}[\sigma(\mathcal{X})] = e^{-\beta H}$ and, for every $(c, X) \in \mathcal{X}$ with degree t, $|c| \leqslant (\frac{\beta}{\beta_c})^t$. □

Because Hermitian monomials X satisfy $\|X\|_{\mathrm{op}} \leqslant 1$ and the scalars in the output configuration $\mathcal{X}$ always satisfy $|c| \leqslant 1$, $I + cX$, and thus $\sigma(\mathcal{X})$ is always positive semi-definite. Further, for reasonable classes of terms, $\sigma(\mathcal{X})$ can be shown to be separable, which means that $e^{-\beta H}$ and therefore the Gibbs state $e^{-\beta H}/\operatorname{tr} e^{-\beta H}$ is separable.

Lemma 2.3.18. *Let $H = \sum_a H_a \in \mathbb{C}^{2^{\mathfrak{n}}\times 2^{\mathfrak{n}}}$ be a Hamiltonian over $\mathfrak{n}$ qubits with Pauli terms, meaning that $H_a = \lambda_a E_a$ for $E_a \in \mathcal{P}$ and $-1 \leqslant \lambda_a \leqslant 1$. Let X be a Hermitian monomial of degree t with respect to the terms of H. Then $I + cX$ is separable when $-1 \leqslant c \leqslant 1$. Further, there is an algorithm which outputs a stabilizer product state $|\psi\rangle$ such that $\mathbb{E}[|\psi\rangle\langle\psi|] = (I + cX)/\operatorname{tr}(I + cX)$ in $\widetilde{\mathcal{O}}(\mathfrak{n} + t\mathfrak{K})$ time.*

Proof. Since X is a Hermitian monomial and every H_a is a tensor product of Paulis, X is either zero or a tensor product of Paulis, up to a sign. This can be proved by induction on the definition of a Hermitian monomial (Definition 2.3.9). With this, we can conclude that $I + cX$ is separable: for $X = X_1 \otimes \cdots \otimes X_{\mathfrak{n}}$, we can write it as a convex combination of tensor products of PSD matrices by writing $I + cX = (1 - |c|)I + |c|(I + \frac{c}{|c|}X)$, and then for the latter part, using

$$I + X = \frac{1}{2^{\mathfrak{n}}} \bigotimes_{i=1}^{\mathfrak{n}} \Big((I + X_i) + (I - X_i) \Big) + \frac{1}{2^{\mathfrak{n}}} \bigotimes_{i=1}^{\mathfrak{n}} \Big((I + X_i) - (I - X_i) \Big).$$

By expanding this, we conclude that, by drawing a uniformly random $s_1, \ldots, s_{\mathfrak{n}} \in \{\pm 1\}$ with even parity, i.e. $s_1 s_2 \ldots s_{\mathfrak{n}} = 1$, and preparing the state $\bigotimes_{i=1}^{\mathfrak{n}} (I + s_i X_i)$, this state is $I + X$ in expectation.

To turn this into an algorithm, we note that given X represented as a list of terms and operations, we can compute $X_1, \ldots, X_{\mathfrak{n}}$ efficiently. For the qubits i for which $X_i = I$, we can take $|\psi\rangle_i$ to be $|0\rangle$ or $|1\rangle$ with half probability, thereby making the corresponding density matrix $I/2$. On the support of X, we can draw a uniformly random string with even parity and prepare the corresponding state $I + s_i X_i$; since X_i is a non-identity Pauli, this is a pure eigenstate of a Pauli. This gives an algorithm for preparing $(I + cX)/\operatorname{tr}(I + cX)$ by either preparing $I + \frac{c}{|c|}X$ or I with the appropriate probabilities. □

With this, we can prove separability of high-temperature Gibbs states for qubit systems, as stated in the introduction.

Proof of Theorem 2.1.1. Theorem 2.3.17 implies that, for $\beta \leqslant 1/(100\mathfrak{d}\mathfrak{K})$, $e^{-\beta H}$ is a convex combination of matrices of the form $\sigma(\mathcal{X})$, where for every $(c, X) \in \mathcal{X}$, $|c| \leqslant 1$ and X is a Hermitian monomial (Definition 2.3.9). By Lemma 2.3.18, $(I + cX)_{\operatorname{supp}(X)}$ is separable (and expressible as a positive linear combination of stabilizer product states), and since $\sigma(\mathcal{X})$ is the tensor product of such matrices, it is also separable (and expressible as a positive linear combination of stabilizer product states). Since

separability is closed under convex combinations and scaling by a constant factor, $e^{-\beta H}/\operatorname{tr} e^{-\beta H}$ is separable as desired. Further, this argument shows it is expressible as a convex combination of stabilizer product states. □

2.4 Efficient Preparation of the Gibbs State

In this section, we prove Theorem 2.1.2. We first need to introduce some additional machinery for approximating quantum partition functions so that we can sample from the correct distribution and ensure that we output a normalized state.

2.4.1 Approximating the Partition Function

At high temperature, we can efficiently estimate the partition function of a Gibbs state up to small multiplicative error. This follows from cluster expansion; we defer most of the proof to the appendix. Prior work immediately gives this result for $\beta \ll 1/\mathfrak{d}^2$ [103]; in the appendix we combine these results with tighter analyses [88, 160] to improve the critical temperature to $\Theta(1/\mathfrak{d})$.

Theorem 2.4.1 (Estimating the log-partition function)**.** *Let $H = \sum_{a=1}^{m} \lambda_a E_a \in \mathbb{C}^{2^{\mathfrak{n}} \times 2^{\mathfrak{n}}}$ be a $\mathfrak{K}$-local qubit Hamiltonian with Pauli terms, i.e. where $-1 \leqslant \lambda_a \leqslant 1$ and $E_a \in \mathcal{P}$ is Pauli. Let $\mathfrak{d}$ be the degree of the dual interaction graph of H. Let $0 \leqslant \beta < \beta_c = 1/(100\mathfrak{d})$. Then given any $0 < \eta < 1$, we can compute an estimate $\widehat{z}$ such that*

$$\log(\operatorname{tr}(e^{-\beta H})) - \eta \leqslant \widehat{z} \leqslant \log(\operatorname{tr}(e^{-\beta H})) + \eta,$$

in time

$$\mathfrak{n} \cdot (\mathfrak{n}/\eta)^{\frac{4+\log(\mathfrak{d})}{\log(\beta_c/\beta)}} \cdot \mathfrak{K} \cdot \mathfrak{d}^2 \cdot \operatorname{polylog}(\mathfrak{n}/\eta).$$

Proof of Theorem 2.4.1. This follows from Lemma A.1.10. Let $\beta_* = 1/(e(e+1)(1 + e(\mathfrak{d}-1)))$ be the inverse temperature threshold from that lemma; since $\beta < \beta_c < \beta_*$, the results of that lemma hold for our choice of inverse temperature β. In particular,

we can write $\mathcal{L} = \log(\mathrm{tr}(e^{-\beta H}))$ as a power series

$$\mathcal{L} = \sum_{\ell \geqslant 0} \beta^\ell p_\ell(\lambda),$$

and when we estimate it by truncating at some level k,

$$\widetilde{\mathcal{L}} = \sum_{\ell=0}^{k} \beta^\ell p_\ell(\lambda),$$

the error is bounded $|\mathcal{L} - \widetilde{\mathcal{L}}| \leqslant \mathfrak{n} \frac{(\beta/\beta_*)^{k+1}}{1-\beta/\beta_*}$ by Part 2 of Lemma A.1.10. This difference can be made to be at most η by taking

$$k = \left\lfloor \frac{\log(n/((1-\beta/\beta_*)\eta))}{\log(\beta_*/\beta)} \right\rfloor. \tag{2.14}$$

To compute $\widetilde{\mathcal{L}}$, we enumerate the list of monomials to order k and then compute all corresponding coefficients. The running time of this is dominated by the task at order k, which, using parts A and B of Lemma A.1.10, is bounded by

$$\underbrace{\mathfrak{n}(e\mathfrak{d})^k}_{\#\text{ of clusters}} \cdot \underbrace{(8^k + \mathfrak{K})\,\mathrm{poly}(k)}_{\text{time to compute a coefficient}} + \underbrace{\mathcal{O}(k\mathfrak{d} \cdot \mathfrak{n}(e\mathfrak{d})^k)}_{\text{time to enumerate clusters}} \tag{2.15}$$

$$= \widetilde{\mathcal{O}}\Big(\mathfrak{n}(8e\mathfrak{d})^k \mathfrak{K}\mathfrak{d}\Big) \tag{2.16}$$

$$= \widetilde{\mathcal{O}}\Big(\mathfrak{n}\big(\frac{n}{(1-\beta/\beta_*)\eta}\big)^{\frac{\log(8e\mathfrak{d})}{\log(\beta_*/\beta)}} \mathfrak{K}\mathfrak{d}\Big) \tag{2.17}$$

This gives the desired bound. Since we are not optimizing the constant in β_c, for our choice of β_c, $\beta_c \leqslant \beta_*/2$, so we can bound $(\frac{1}{1-\beta/\beta_*})^{\frac{\log(8e\mathfrak{d})}{\log(\beta_*/\beta)}} = \mathcal{O}(\mathfrak{d})$. □

2.4.2 Random Walks on Trees

Our sampling algorithm will take Algorithm 2.3.13 and Algorithm 2.3.11 from the separability proof, and then argue that we can run a Markov chain on the "state tree" of these algorithms to get a distribution over output configuration which satisfies $\mathbb{E}[\sigma(\mathcal{X})/\,\mathrm{tr}(\sigma(\mathcal{X}))] \approx \rho$, instead of the original guarantee that $\mathbb{E}[\sigma(\mathcal{X})] = e^{-\beta H}$.

As a subroutine of our main sampling algorithm, we will design a (classical) random walk on a tree. In this section, we present some general machinery for analyzing the mixing times of random walks on trees. We begin with the definition of a *weighted tree*.

Definition 2.4.2 (Weighted tree)**.** Let $\mathcal{T}$ be a tree with a unique root such that all root-to-leaf paths have length at most n. A weighted tree $(\mathcal{T}, w)$ of depth n is obtained by assigning some non-negative weight w_v to each leaf v of the tree. For an interior node v, we let the weight assigned to v be the sum of the weights of the leaves in the sub-tree rooted at v.

Next, we assume access to the following sampling sub-routine:

Definition 2.4.3 (Sample query)**.** For a weighted tree $(\mathcal{T}, w)$, we define a sample query as querying a node u and outputting a random child of u, where the probability of outputting a child v is w_v/w_u.

Since the weights on each leaf are non-negative, they induce a probability distribution over the leaves, corresponding to the distribution formed by repeatedly performing sample queries starting from the root node.

Definition 2.4.4 (Leaf distribution)**.** For a weighted tree $(\mathcal{T}, w)$, its leaf distribution is the distribution over leaves where each leaf v is sampled proportional to its weight w_v.

We then recall the definition of a Markov chain and the corresponding stationary distribution.

Definition 2.4.5 (Transition matrix and stationary distribution)**.** A Markov chain on N states is given by a transition matrix P with entries P_{ij} for $i, j \in [N]$ given by the probability of moving from state i to state j. We define the stationary distribution, denoted by $\pi = (\pi_1, \ldots, \pi_N)$, such that $P\pi = \pi$.

Definition 2.4.6 (Ergodic and time-reversible Markov chain)**.** A Markov chain P

is *ergodic* if there exists a positive integer z such that P^z is entry-wise positive. It is *time-reversible* if its stationary distribution π satisfies

$$P_{ij}\pi_i = P_{ji}\pi_j$$

for all $i, j \in [N]$.

Next, we define the notion of conductance of a Markov chain.

Definition 2.4.7 (Conductance)**.** Given a Markov chain on N states with transition matrix P and stationary distribution π, for any subset $S \subseteq [N]$, the conductance Φ_S is defined by

$$\Phi_S = \frac{\sum_{i \in S, j \notin S} P_{ij}\pi_i}{\sum_{i \in S} \pi_i}\,.$$

The global conductance of the chain is defined by $\Phi = \min_{S:C_S \leqslant 1/2} \Phi_S$, where $C_S = \sum_{i \in S} \pi_i$.

A classical result of Jerrum and Sinclair [194] bounds the spectral gap of an ergodic, time-reversible Markov chain as a function of the conductance.

Lemma 2.4.8 (Spectral gap of a Markov chain [194])**.** *For an ergodic time-reversible Markov chain P, if we order the eigenvalues of P as $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_N$ where $\lambda_1 = 1$, then*

$$\lambda_1 \leqslant 1 - \frac{\Phi^2}{2}\,.$$

The main result in this section, stated in Theorem 2.4.9 below, is about relating two different weighted trees on the same vertex set, which we denote by $(\mathcal{T}, w)$ and $(\mathcal{T}, w')$. For a vertex v, the distortion between the two weight functions is just w_v/w'_v. Note that scaling a weight function by a constant factor doesn't affect any of the resulting distributions. Given any edge (u, v) on this tree, we assume that the distortion between w and w' along this edge is bounded: $0.1 \leqslant (w_u/w_v) \cdot (w'_u/w'_v) \leqslant 10$. With this assumption, we show that given sample access to w' and exact access to w_v/w'_v at the leaves, but only a constant approximate oracle for w_v/w'_v at interior nodes, we

can efficiently sample from the leaf distribution of w via a Markov chain that mixes quickly. This is closely related to the reduction from sampling to weak approximate counting for self-reducible problems in [194].

Theorem 2.4.9 (Sampling a leaf via a Markov chain)**.** *Let $(\mathcal{T}, w), (\mathcal{T}, w')$ be weighted trees of depth n on the same vertex set such that each vertex has at most k children. Assume that for any node $u \in \mathcal{T}$ with child $v \in \mathcal{T}$, $0.1 \leqslant (w_u w'_v)/(w_v w'_u) \leqslant 10$. Further, assume we are given an oracle that can perform the following types of queries.*

1. *For any internal node v, compute an estimate $\widehat{r}_v$ such that $0.1(w_v/w'_v) \leqslant \widehat{r}_v \leqslant 10(w_v/w'_v)$.*
2. *For any leaf node v, exactly compute $\widehat{r}_v = w_v/w'_v$.*
3. *For any internal node v, respond to a sample query for $(\mathcal{T}, w')$ (see Definition 2.4.3).*

Then, for any $0 < \varepsilon, \delta < 1$, there exists an algorithm that uses $\mathcal{O}(n^4 \log(nk/\varepsilon) \log(1/\delta))$ queries to the aforementioned oracle and outputs either a leaf node $v \in \mathcal{T}$ or $\perp$. The algorithm outputs a leaf with probability $\geqslant 1 - \delta$, and when the algorithm outputs a leaf, its distribution is ε-close to the leaf distribution of $(\mathcal{T}, w)$ in TV distance.

Proof. Consider the following random walk on $(\mathcal{T}, w)$.

Algorithm 2.4.10 (Tree random walk)**.**

Set v to be the root node;
for $\mathcal{O}(n^3 \log(nk/\varepsilon))$ iterations **do**
    Let u be the parent of v, if it exists;
    Sample an $\xi \in \{0, 1, 2\}$ with probability $\{0.01\widehat{r}_u/\widehat{r}_v, 0.01, 0.99 - 0.01\widehat{r}_u/\widehat{r}_v\}$, respectively.
    If $\xi = 0$, set v to u;

6: If $\xi = 1$, set v to a child of v chosen according to a sample query on (T, w'), if one exists;

7: If $\xi = 2$, keep v unchanged;

8: **output** v;

Note that for all vertices, we query the oracle once for $\widehat{r}_v$ and always use the same estimate throughout the random walk, so that we never query the same vertex again for a new estimate. By assumption, $\widehat{r}_u/\widehat{r}_v \leqslant 10$ so this walk is well-defined.

First, we prove that the stationary distribution of this walk has probability mass on each vertex proportional to $\widehat{r}_v w'_v$. In particular, since $\widehat{r}_v = \frac{w_v}{w'_v}$ on any leaf v, this is exactly the distribution proportional to w on the leaves.

We do this by verifying reversibility (see Definition 2.4.6). Consider any two vertices u, v such that u is a parent of v. Then we have

$$P_{uv}\pi_v = 0.01\frac{\widehat{r}_u}{\widehat{r}_v}\widehat{r}_v w'_v = 0.01\widehat{r}_u w'_v = 0.01\frac{w'_v}{w'_u}\widehat{r}_u w'_u = P_{vu}\pi_u$$

as desired. Thus we can conclude that the Markov chain is ergodic and time-reversible.

Next, we will lower bound the spectral gap of this walk by bounding the conductance and applying Lemma 2.4.8. To lower bound the conductance we show that it suffices to consider when the subset S is a sub-tree rooted at some vertex v. To see this, consider any cut (T_1, T_2) such that T_2 contains the root. Let $\mathcal{M} = \{v_i\}_{i\in[r]}$ be the maximal elements in T_1, i.e. for each v_i, the parent of v_i , denoted by u_i, is in T_2. Then,

$$\Phi_{T_1} \geqslant \frac{\sum_{v_i\in\mathcal{M}} P_{v_i u_i}\pi_{v_i}}{\sum_{v_i\in T_1}\pi_{v_i}} \geqslant \frac{\sum_{v_i\in\mathcal{M}} P_{v_i u_i}\pi_{v_i}}{\sum_{v_i\in\mathcal{M}}\sum_{v_j\in \text{ sub-tree } (v_i)}\pi_{v_j}} \geqslant \frac{P_{v_i u_i}\pi_{v_i}}{\sum_{v_j\in \text{ sub-tree } (v_i)}\pi_{v_j}}$$

Let S be the sub-tree rooted at v and let u be the parent. Then

$$\Phi_S \geqslant \frac{0.1\widehat{r}_u w'_v}{\sum_{v' \preceq v} \widehat{r}_v w'_v} \geqslant \frac{w_u w'_v}{40 w'_u (\sum_{v' \preceq v} w_v)} \geqslant \frac{w_u w'_v}{40 n w'_u w_v} \geqslant \frac{1}{80n}$$

where in the above, we used that $\sum_{v' \preceq v} w_v \leqslant n w_v$ from the definition of a weighted tree of depth n. Thus, by Lemma 2.4.8, the spectral gap of the Markov chain is $\Omega(1/n^2)$. Finally, the number of nodes in the tree is at most $(k+1)^n$ so after $O(n^3 \log(k/\varepsilon))$ steps of the Markov chain, the distribution will be ε-close to the stationary distribution (see for instance [150]). Note that the stationary distribution matches the leaf distribution of (T, w) on the leaves and also the probability of being at a leaf in the stationary distribution is at least

$$\frac{\sum_{v \text{ leaf}} \widehat{r}_v w'_v}{\sum_v \widehat{r}_v w'_v} \leqslant \frac{1}{4} \frac{\sum_{v \text{ leaf}} w_v}{\sum_v w_v} \geqslant \frac{1}{4n},$$

using that the tree is depth n. Thus, with $O(n^3 \log(nk/\varepsilon))$ steps of the Markov chain, we get $\varepsilon/(100n)$-close to the stationary distribution, and the output will be a leaf with probability $1/(4n)$. So, by running Algorithm 2.4.10 until the output is a leaf, with probability $1 - \delta$, we will hit a leaf within $O(n \log(1/\delta))$ epochs and this gives the desired output. □

2.4.3 Sample Tree

We want to apply the above algorithm to the sample tree corresponding to Algorithm 2.3.11. In particular, we consider a root node corresponding to an initial input, an empty configuration with zero pinned nodes, $S = [\mathfrak{n}]$ and $\mathcal{X} = \varnothing$. Then, the children of this node correspond to calling the algorithm on $(S, \mathcal{X})$, which we can interpret as pinning more unpinned sites and adding them to the configuration.

Definition 2.4.11 (Sample tree of a Hamiltonian)**.** Given a Hamiltonian $H = \sum_a H_a$ over n sites such that H is a $\mathfrak{K}$-local Hamiltonian with dual interaction graph with degree $\mathfrak{d}$, along with parameters β and ε, we construct its sample tree as follows:

- It has a root node with label $(S, \mathcal{X}) = ([n], \varnothing)$.
- For each node v such that its label $(S_v, \mathcal{X}_v)$ satisfies $S_v \neq \varnothing$, it has a child for every possible outcome of running Algorithm 2.3.11 on that node with input S_v, $\mathcal{X}_v$, and ε, labeled with the corresponding output $(\widehat{S}, \widehat{\mathcal{X}})$.
- For a node v, we associate its label $(S_v, \mathcal{X}_v)$ with the matrix

$$Q_v = e^{-\frac{\beta}{2}H^{(S_v)}} \sigma(\mathcal{X}_v) e^{-\frac{\beta}{2}H^{(S_v)}} .$$

Our overall strategy is as follows. Assigned to every node in the sample tree is a matrix $Q_v \in \mathbb{C}^{2^n \times 2^n}$; for the root node, this matrix is $e^{-\beta H}$, and for a leaf node, this matrix is a configuration $\sigma(\mathcal{X}_v)$. In between, we think of Q_v as a "posterior" Gibbs state: having pinned all qubits in $[\mathfrak{n}] \backslash S_v$ according to the (separable) configuration $\sigma(\mathcal{X}_v)$, Q_v is then some Gibbs-like state on S_v.

Going from a node to its children pins more sites. We first show in Lemma 2.4.18 that the expectation of $Q_{v'}$ over the children v' of a node v, with respect to the distribution given by the sampling algorithm Algorithm 2.3.11, is close to the matrix Q_v. When $\varepsilon = 0$, this expectation is exact by Lemma 2.3.15, but we have to take a non-zero ε to get an efficient algorithm. Then, by iterating this lemma, we show that the average over the leaves is close to the root matrix, $e^{-\beta H}$. In other words, iterating Algorithm 2.3.11 until $S = \varnothing$ gives a configuration $\mathcal{X}$ such that $\sigma(\mathcal{X}) \approx e^{-\beta H}$.

We would like to sample the Gibbs state $e^{-\beta H} / \operatorname{tr}(e^{-\beta H})$ by sampling a leaf v from a certain distribution, and then preparing the corresponding configuration $\sigma(\mathcal{X}_v) / \operatorname{tr}(\sigma(\mathcal{X}_v))$; we define this distribution in Definition 2.4.16. However, this distribution is not the same as the distribution naturally induced by running Algorithm 2.3.11; this distribution is defined in Definition 2.4.14. To prove our main result, we show that the ratio between these two distributions is bounded and can be efficiently approximated. We then use Theorem 2.4.9 to sample. Throughout this argument, we use the lemmas in Section 2.3.3, which describe the behavior of Algorithm 2.3.11, to analyze the sample

tree.

We begin by recording a few basic observations about the structure of the sample tree. In particular, we observe that, throughout, at most one element of $\mathcal{X}_v$ ever intersects with S_v, so that Q_v can be written as a tensor product of a set of "pinned" sites with a set of "active" sites.

Lemma 2.4.12. *The sample tree has the following properties. For a node $v = (S_v, \mathcal{X}_v)$ at depth d:*

1. *$|S_v| \leqslant \mathfrak{n} - d$, so the depth of the tree is at most $\mathfrak{n}$;*
2. *For any two children u, u' of v, $S_u = S_{u'}$;*
3. *For the root node u, $Q_u = e^{-\beta H}$, and for a leaf node v, $Q_v = \sigma(\mathcal{X}_v)$;*
4. *For an interior node v, Q_v decomposes into a tensor product $Q_v = (Y_v)_{\mathrm{supp}(Y_v)} \otimes (e^{-\frac{\beta}{2}H^{(S_v)}}(I + cX)e^{-\frac{\beta}{2}H^{(S_v)}})_{[\mathfrak{n}]\setminus\mathrm{supp}(Y_v)}$, where $(c, X) \in \mathcal{X}_v$ or $(c, X) = (0, I)$;*
5. *When $\beta \leqslant \frac{1}{200\mathfrak{d}\mathfrak{K}}$, for every node v and every $(c, X) \in \mathcal{X}_v$, $|c| \leqslant \frac{1}{2}$; in particular, Q_v is PSD.*

Proof. 1 and 2 follow from inspection of Algorithm 2.3.11; it always removes elements from S_v, and it makes this choice deterministically. 3 follows from the definition of Q_v. 4 follows from Lemma 2.3.14. 5 follows from Lemma 2.3.16. □

We will also need a bound on the running time of Algorithm 2.3.11.

Lemma 2.4.13 (Algorithm complexity)**.** *Let $H = \sum_{a\in[m]} H_a$ be a Hamiltonian and we assume access to it as in Definition 1.4.16. Then Algorithm 2.3.11 can be implemented to run in time $\widetilde{\mathcal{O}}(\log(\mathfrak{n}/\varepsilon)\mathfrak{K}\mathfrak{d})$, where the output configuration is described in terms of the descriptions of its Hermition monomials. Further, the number of possible outputs is $\mathcal{O}((40\log(\mathfrak{n}/\varepsilon)(\mathfrak{d}+1))^{20\log(\mathfrak{n}/\varepsilon)})$.*

Proof. We begin by considering Algorithm 2.3.3 with input H, $k \geqslant 0$, and $|\mathcal{Q}| \leqslant \mathfrak{d}+1$. Then this algorithm can be run in $\widetilde{\mathcal{O}}(k\mathfrak{d})$ time. The main technical consideration is maintaining $\mathcal{R}_t$, which can be done by adding the neighbors of the term being added to $\mathcal{R}_t$ to form $\mathcal{R}_{t+1}$. This costs $\widetilde{\mathcal{O}}(\mathfrak{d})$ time since there are at most that many neighbors. Further, the number of possible outputs $c_k, b^{(k)}$ is $\mathcal{O}(((k-1)!) \cdot (4(\mathfrak{d}+1))^k)$, since $|\mathcal{R}_t| \leqslant t(\mathfrak{d}+1)$.

Next, consider Algorithm 2.3.5, which calls Algorithm 2.3.3 for some k between 1 and $t_{\max}$. So, its running time is $\widetilde{\mathcal{O}}(t_{\max}\mathfrak{d})$ and the number of possible outputs is $\mathcal{O}((t_{\max}!) \cdot (4(\mathfrak{d}+1))^{t_{\max}})$.

Finally, consider Algorithm 2.3.11. This algorithm can be run in $\widetilde{\mathcal{O}}(t_{\max}\mathfrak{d}\mathfrak{K})$ time: it can be seen as taking an element of the configuration $(c_{\widehat{\ell}}, X_{\widehat{\ell}}) \in \mathcal{X}$ and changing it to have at most $2t_{\max}$ more terms in it. The step of finding an element of $\mathcal{E}^{(S)}$ neighboring a term in $X_{\widehat{\ell}}$ can be done by maintaining a list of such terms; then, when $X_{\widehat{\ell}}$ is changed, this list can be updated by removing elements when S is modified and adding elements corresponding to the terms being added. The algorithm is deterministic apart from the two calls to Algorithm 2.3.5 and the choice of ξ among seven options. We know that $|\mathcal{Q}| \leqslant \mathfrak{d}+1$, so the number of possible outputs is

$$\mathcal{O}(((t_{\max}!) \cdot (4(\mathfrak{d}+1))^{t_{\max}})^2) = \mathcal{O}((40\log(\mathfrak{n}/\varepsilon)(\mathfrak{d}+1))^{20\log(\mathfrak{n}/\varepsilon)}).$$

Note that, across all of these algorithms, we never use the matrix representations of the Hamiltonian terms, instead only using the information coming from the dual interaction graph and the support of every term. □

2.4.4 Weight Function Analysis

Now we define two different weighted trees on the sample tree. The first is the weight derived from the sampling process.

Definition 2.4.14 (Natural weight)**.** Given a Hamiltonian $H = \sum_a H_a$ over n sites

and parameters $\beta, \mathfrak{K}, \mathfrak{d}$ such that H is a $\mathfrak{K}$-local Hamiltonian with degree $\mathfrak{d}$, let $\mathcal{T}$ be its sample tree. We define the natural weight function ω to have for each node $v \in \mathcal{T}$, $\omega(v)$ is the probability of reaching v from the root by running the sampling process in Algorithm 2.3.11 at each intermediate node.

Remark 2.4.15. The weight at any intermediate node is equal to the sum of the weights of the leaves of its sub-tree, so ω indeed defines a valid weighted tree.

The second weight function, the true weight, is defined by adjusting the natural weight at each leaf by $\mathrm{tr}(\sigma(\mathcal{X}))$. We will show in Corollary 2.4.19 that to sample from the Gibbs state, it suffices to sample a leaf according to this true weight distribution.

Definition 2.4.16 (True weight)**.** Given a Hamiltonian $H = \sum_a H_a$ over n sites and parameters $\beta, \mathfrak{K}, \mathfrak{d}$ such that H is a $\mathfrak{K}$-local Hamiltonian with degree $\mathfrak{d}$, let T be its sample tree. We define the true weight function κ as follows: for each leaf node v, we set $\kappa(v) = \mathrm{tr}(\sigma(\mathcal{X}_v))\omega(v)$ and for each intermediate node v, $\kappa(v)$ is the sum of the weights of the leaves in its sub-tree.

Remark 2.4.17. By Lemma 2.4.12, when $\beta \leqslant 1/(200\mathfrak{d}\mathfrak{K})$, all of the $\sigma(\mathcal{X}_v)$ are PSD and thus this is a valid weighted tree.

First, we prove that Q_v is close to the average of $Q_{v'}$ over its children v', according to the natural weight. Since these matrices are not trace-normalized, it will be important to ensure that our error bounds are "at the right scale". We do this by bounding our errors multiplicatively in PSD ordering.

Lemma 2.4.18. *Given a Hamiltonian $H = \sum_a H_a$ over n sites and parameters $\beta, \mathfrak{K}, \mathfrak{d}$ such that H is a $\mathfrak{K}$-local Hamiltonian with degree $\mathfrak{d}$ and $\beta \leqslant 1/(200\mathfrak{d}\mathfrak{K})$, let $\mathcal{T}$ be its sample tree as defined in Definition 2.4.11. Let $v \in \mathcal{T}$ be a node in the tree. Then we have*

$$\Big(1 - \frac{\varepsilon}{20n}\Big) Q_v \preceq \sum_{v' \text{ child of } v} \frac{\omega(v')}{\omega(v)} Q_{v'} \preceq \Big(1 + \frac{\varepsilon}{20n}\Big) Q_v.$$

Proof. Consider the execution of Algorithm 2.3.11 with input $(S_v, \mathcal{X}_v)$. Then by

Lemma 2.3.15, over the expectation of the children v' of v,

$$\mathop{\mathbb{E}}_{v'}[\sigma(\mathcal{X}_{v'})] = T_{t_{\max},\beta/2}(H^{(S_v)}, H^{(S_v)}_{(S_v\setminus S_{v'})})^\dagger \sigma(\mathcal{X}_v) T_{t_{\max},\beta/2}(H^{(S_v)}, H^{(S_v)}_{(S_v\setminus S_{v'})})\,.$$

Here, we are using Lemma 2.4.12, that every child v' of v has the same $S_{v'}$. By applying Lemma 2.3.8, we have that

$$\mathop{\mathbb{E}}_{v'}[\sigma(\mathcal{X}_{v'})] \preceq \Big(1 + \frac{100}{C^{t_{\max}}}\Big) e^{\frac{\beta}{2}(H^{(S_v)} - H^{(S_v)}_{(S_v\setminus S_{v'})})} e^{-\frac{\beta}{2}H^{(S_v)}} \sigma(\mathcal{X}_v) e^{-\frac{\beta}{2}H^{(S_v)}} e^{\frac{\beta}{2}(H^{(S_v)} - H^{(S_v)}_{(S_v\setminus S_{v'})})}$$

Using that $t_{\max} = 10\log(\mathfrak{n}/\varepsilon)$ and left and right multiplying the above by $e^{-\frac{\beta}{2}(H^{(S_v)} - H^{(S_v)}_{(S_v\setminus S_{v'})})}$ and applying Lemma 2.3.8, we get

$$\mathop{\mathbb{E}}_{v'}[Q_{v'}] \preceq \Big(1 + \frac{\varepsilon}{10\mathfrak{n}}\Big) Q_v.$$

Here, we use that $H^{(S_v)} - H^{(S_v)}_{(S_v\setminus S_{v'})} = H^{(S_{v'})}$. This gives us the upper bound, since averaging over the children of v with weights $\omega(v')/\omega(v)$ is exactly the same as taking the expectation over the execution of Algorithm 2.3.11: $\mathbb{E}_{v'}[Q_{v'}] = \sum_{v' \text{ child of } v} \frac{\omega(v')}{\omega(v)} Q_{v'}$. By applying the other side of Lemma 2.3.8, we get the corresponding lower bound. □

By iterating Lemma 2.4.18, we can relate the average at the leaves to the Gibbs state. Specifically, since the true weight distribution on the leaves is exactly defined to be equal to the natural weight distribution arising from the sampling process and then distorted by a factor of α_v at each leaf v, we get that the average of the (normalized) states at the leaves according to the true weight distribution is close to the Gibbs state $e^{-\beta H}/\operatorname{tr}(e^{-\beta H})$.

Corollary 2.4.19 (Average of the leaves is close to the Gibbs state)**.** *Given a Hamiltonian $H = \sum_a H_a$ over n sites and parameters $\beta, \mathfrak{K}, \mathfrak{d}$ such that H is a $\mathfrak{K}$-local*

Hamiltonian with degree $\mathfrak{d}$ *and* $\beta \leqslant 1/(200\mathfrak{d}\mathfrak{K})$*, let* $\mathcal{T}$ *be its sample tree. Then*

$$\Big\| \frac{e^{-\beta H}}{\operatorname{tr}(e^{-\beta H})} - \sum_{v \text{ leaf of } \mathcal{T}} \frac{\kappa(v)}{\sum_{v' \text{ leaf of } \mathcal{T}} \kappa(v')} \frac{\sigma(\mathcal{X}_v)}{\operatorname{tr}(\sigma(\mathcal{X}_v))} \Big\|_1 \leqslant \frac{\varepsilon}{2} .$$

Proof. By repeatedly applying Lemma 2.4.18 starting from the root to the leaves, we have

$$\Big(1 - \frac{\varepsilon}{10}\Big) e^{-\beta H} \preceq \sum_{v \text{ leaf of } \mathcal{T}} \kappa(v) \frac{\sigma(\mathcal{X}_v)}{\operatorname{tr}(\sigma(\mathcal{X}_v))} \preceq \Big(1 + \frac{\varepsilon}{10}\Big) e^{-\beta H} ,$$

which follows from recalling that $\kappa(v) = \operatorname{tr}(\sigma(\mathcal{X}_v))\omega(v)$. Thus

$$\begin{aligned} &\Big\| \frac{e^{-\beta H}}{\operatorname{tr}(e^{-\beta H})} - \sum_{v \text{ leaf of } \mathcal{T}} \frac{\kappa(v)}{\sum_{v' \text{ leaf of } \mathcal{T}} \kappa(v')} \frac{\sigma(\mathcal{X}_v)}{\operatorname{tr}(\sigma(\mathcal{X}_v))} \Big\|_1 \\ &\leqslant 2 \Big\| \frac{e^{-\beta H}}{\operatorname{tr}(e^{-\beta H})} - \sum_{v \text{ leaf of } \mathcal{T}} \frac{\kappa(v)}{\operatorname{tr}(e^{-\beta H})} \frac{\sigma(\mathcal{X}_v)}{\operatorname{tr}(\sigma(\mathcal{X}_v))} \Big\|_1 \leqslant \frac{\varepsilon}{2} \end{aligned}$$

as desired. □

To sample from the distribution induced by κ on the leaves, we will need to approximate the ratio $\kappa(v)/\omega(v)$ in order to then apply Theorem 2.4.9 (since we can run Algorithm 2.3.11 to sample from the weighted tree ω). We do this below.

Corollary 2.4.20 (Bounded weight ratio)**.** *Given a Hamiltonian* $H = \sum_a H_a$ *over* n *sites and parameters* $\beta, \mathfrak{K}, \mathfrak{d}$ *such that* H *is a* $\mathfrak{K}$*-local Hamiltonian with degree* $\mathfrak{d}$ *and* $\beta \leqslant 1/(200\mathfrak{d}\mathfrak{K})$*, let* $\mathcal{T}$ *be its sample tree. Let* v *be a node with associated label* $(S_v, \mathcal{X}_v)$*. Then we have*

$$0.9 \operatorname{tr}(Q_v) \leqslant \frac{\kappa(v)}{\omega(v)} \leqslant 1.1 \operatorname{tr}(Q_v) .$$

Proof. By repeatedly applying Lemma 2.4.18, we have

$$\Big(1 - \frac{\varepsilon}{20n}\Big)^{|S_v|} Q_v \preceq \sum_{\substack{v' \text{ a leaf} \\ v' \text{ descendant of } v}} \frac{\omega(v')}{\omega(v)} \sigma(\mathcal{X}_{v'}) \preceq \Big(1 + \frac{\varepsilon}{20n}\Big)^{|S_v|} Q_v .$$

Now take the trace of the above to get the desired bound:

$$\left(1-\frac{\varepsilon}{20n}\right)^{|S_v|}\operatorname{tr}(Q_v) \leqslant \frac{\kappa(v)}{\omega(v)} \leqslant \left(1+\frac{\varepsilon}{20n}\right)^{|S_v|}\operatorname{tr}(Q_v) \qquad \square$$

Now we can put everything together to prove our our main theorem, Theorem 2.1.2.

Proof of Theorem 2.1.2. We will apply Theorem 2.4.9 on the sample-tree with $w' \leftarrow \omega$ and $w \leftarrow \kappa$. First, we verify the hypotheses of Theorem 2.4.9. By Lemma 2.4.13, the number of children of each node is at most $k = \mathcal{O}((40\log(\mathfrak{n}/\varepsilon)(\mathfrak{d}+1))^{20\log(\mathfrak{n}/\varepsilon)})$. Consider a node u with child v, and consider their associated labels $(S_u, \mathcal{X}_u)$ and $(S_v, \mathcal{X}_v)$. Note this means $S_v = S_u \backslash R$ for some $R = \operatorname{supp}(H_a)$. By Corollary 2.4.20,

$$0.8\frac{\operatorname{tr}(Q_u)}{\operatorname{tr}(Q_v)} \leqslant \frac{w_u w'_v}{w_v w'_u} \leqslant 1.25\frac{\operatorname{tr}(Q_u)}{\operatorname{tr}(Q_v)}.$$

Recall that $Q_u = e^{-\frac{\beta}{2}H^{(S_u)}}\sigma(\mathcal{X}_u)e^{-\frac{\beta}{2}H^{(S_u)}}$, and respectively for Q_v. $(S_v, \mathcal{X}_v)$ is a possible outcome of running Algorithm 2.3.11 on $(S_u, \mathcal{X}_u)$. So, because of Lemma 2.3.14, we can write $Q_u = Y_{\operatorname{supp}(Y)} \otimes (e^{-\frac{\beta}{2}H^{(S_u)}}(I+cX)e^{-\frac{\beta}{2}H^{(S_u)}})_{[\mathfrak{n}]\backslash\operatorname{supp}(Y)}$ and $Q_v = Y_{\operatorname{supp}(Y)} \otimes (e^{-\frac{\beta}{2}H^{(S_v)}}(I+\widehat{c}\widehat{X})e^{-\frac{\beta}{2}H^{(S_v)}})_{[\mathfrak{n}]\backslash\operatorname{supp}(Y)}$ for (c, X) and $(\widehat{c}, \widehat{X})$ corresponding to the $(c_{\widehat{\ell}}, X_{\widehat{\ell}})$ and $(\widehat{c}, \widehat{X})$ in Algorithm 2.3.11. Here, we are treating the unchanged part of the configuration of $\mathcal{X}_u$ as Y.

$$\frac{\operatorname{tr}(Q_u)}{\operatorname{tr}(Q_v)} = \frac{\operatorname{tr}(e^{-\frac{\beta}{2}H^{(S_u)}}(I+cX)e^{-\frac{\beta}{2}H^{(S_u)}})}{\operatorname{tr}(e^{-\frac{\beta}{2}H^{(S_v)}}(I+\widehat{c}\widehat{X})e^{-\frac{\beta}{2}H^{(S_v)}})} \leqslant 3\frac{\operatorname{tr}(e^{-\beta H^{(S_u)}})}{\operatorname{tr}(e^{-\beta H^{(S_v)}})} \leqslant 4.5$$

The first inequality follows from Lemma 2.4.12, which implies that $|c|, |\widehat{c}| \leqslant 1/2$, so $0.5I \preceq I + cX \preceq 1.5I$. The second follows from Lemma 2.3.7 which implies that $\frac{3}{4} \leqslant \operatorname{tr}(e^{-\beta H^{(S_u)}})/\operatorname{tr}(e^{-\beta H^{(S_v)}}) \leqslant \frac{3}{2}$. Combining this with the analogous lower bound, we deduce

$$0.1 \leqslant \frac{w_u w'_v}{w_v w'_u} \leqslant 10\,.$$

Next, we show how to implement the types of queries required in Theorem 2.4.9. For the first type of query, estimating w_v/w'_v to multiplicative error, by Corol-

lary 2.4.20 it suffices to be able to estimate $\operatorname{tr}(Q_v)$. By Lemma 2.4.18, we can write $\operatorname{tr}(Q_v) = \operatorname{tr}((Y_v)_{\operatorname{supp}(Y_v)}) \operatorname{tr}((e^{-\frac{\beta}{2}H^{(S_v)}}(I + cX)e^{-\frac{\beta}{2}H^{(S_v)}})_{[\mathfrak{n}]\setminus\operatorname{supp}(Y_v)})$. Computing the trace associated with Y_v can be done exactly, as Y_v is a tensor product of Hermitian monomials from the configuration $\mathcal{X}_v$. Using that $|c| \leqslant 1/2$, the second trace is between $0.5 \operatorname{tr}((e^{-\beta H^{(S_v)}})_{[\mathfrak{n}]\setminus\operatorname{supp}(Y_v)})$ and $1.5 \operatorname{tr}((e^{-\beta H^{(S_v)}})_{[\mathfrak{n}]\setminus\operatorname{supp}(Y_v)})$, and this can be estimated via Theorem 2.4.1 with $\eta \leftarrow 0.01$. The running time of answering this query is

$$O\left(\mathfrak{n} \cdot (100\mathfrak{n})^{\frac{4+\log(\mathfrak{d})}{\log(\beta_c/\beta)}} \cdot \mathfrak{K} \cdot \mathfrak{d}^2 \cdot \operatorname{polylog}(\mathfrak{n})\right)$$

For the third type of queries, we simply run Algorithm 2.3.11. By Lemma 2.4.13, the running time is $\widetilde{\mathcal{O}}(\log(n/\varepsilon) \operatorname{poly}(\mathfrak{K}, \mathfrak{d}))$. Note that when running the Markov chain in Theorem 2.4.9, we can store the labels $(S_u, \mathcal{X}_u)$ of all of the nodes that we have visited so far. When visiting a new node, computing the new label $(S, \mathcal{X})$ involves just one execution of Algorithm 2.3.11 on its parent, which we must have already visited. Thus, whenever we visit a leaf v, can compute $\operatorname{tr}(\sigma(\mathcal{X}_v)) = \kappa(v)/\omega(v)$ exactly, allowing us to answer the second type of query whenever we need to (which is only when the Markov chain visits a leaf). Thus, we have verified all of the hypotheses of Theorem 2.4.9. Putting everything together, we get that with probability $1 - \delta$, we get a sample from a distribution that is $\varepsilon/4$-close in TV to the leaf distribution of κ in time

$$\widetilde{\mathcal{O}}\Big((100\mathfrak{n})^{5+\frac{\cdot(4+\log(\mathfrak{d}))}{\log(\beta_c/\beta)}} \log^2(1/\varepsilon) \log(1/\delta) \operatorname{poly}(\mathfrak{K}, \mathfrak{d})\Big).$$

By setting $\delta \leftarrow \varepsilon/4$, we get that the output distribution including failure is $\varepsilon/2$-close in TV distance to the leaf distribution of κ.

A leaf v corresponds to a configuration $\mathcal{X}_v$; by applying Lemma 2.3.18 to each $(c, X) \in \mathcal{X}_v$, in $\widetilde{\mathcal{O}}(\mathfrak{n} \log(1/\varepsilon)\mathfrak{K})$ time we can prepare a corresponding stabilizer product state $|\psi\rangle$ such that $\mathbb{E}[|\psi\rangle\langle\psi|] = \sigma(\mathcal{X}_v)/\operatorname{tr}(\sigma(\mathcal{X}_v))$. Note that, here we use that the sum of the degrees of the monomials involved in $\mathcal{X}_v$ is bounded by $\mathcal{O}(\mathfrak{n}t_{\max})$, the number of iterations multiplied by the largest the monomial degree can increase per-iteration. Finally,

$$\left\| \frac{e^{-\beta H}}{\operatorname{tr}(e^{-\beta H})} - \mathbb{E}[|\psi\rangle\langle\psi|] \right\|_1 \leqslant \left\| \frac{e^{-\beta H}}{\operatorname{tr}(e^{-\beta H})} - \sum_{v \text{ leaf of } T} \frac{\kappa(v)}{\sum_{v' \text{ leaf of } T} \kappa(v')} \frac{\sigma(\mathcal{X}_v)}{\operatorname{tr}(\sigma(\mathcal{X}_{v'}))} \right\|_1$$
$$+ \left\| \sum_{v \text{ leaf of } T} \frac{\kappa(v)}{\sum_{v' \text{ leaf of } T} \kappa(v')} \frac{\sigma(\mathcal{X}_v)}{\operatorname{tr}(\sigma(\mathcal{X}_{v'}))} - \mathbb{E}\Big[\frac{\sigma(\mathcal{X}_v)}{\operatorname{tr}(\sigma(\mathcal{X}_v))}\Big] \right\|_1 \leqslant \varepsilon.$$

where the expectation is over the execution of the Markov chain in Theorem 2.4.9, conditioned on a valid output, and the call to Lemma 2.3.18. Note that in the last step above we used Corollary 2.4.19 and the fact that the distribution of our output is $\varepsilon/2$ close to the distribution induced by κ on the leaves. Thus, we can simply output $\widehat{\rho} = |\psi\rangle\langle\psi|$ and this completes the proof. The bound we required on the temperature is that $\beta \leqslant 1/(200\mathfrak{d}\mathfrak{K})$; we set β_c sufficiently small so that the exponent in the running time can be bounded $5 + \frac{4+\log(\mathfrak{d})}{\log(\beta_c/\beta)} \leqslant 6 + \frac{\log(\mathfrak{d})}{\log(\beta_c/\beta)}$, and that the $100^{\frac{\log(\mathfrak{d})}{\log(\beta_c/\beta)}}$ can be folded into the $\operatorname{poly}(\mathfrak{d})$. □

Chapter 3

Learning Quantum Hamiltonians

The previous chapter focused broadly on the question of understanding the properties of the Gibbs state of a given local Hamiltonian. In this chapter, we study the inverse problem, called Hamiltonian learning — given access to measurements of the Gibbs state, can we recover information about the underlying Hamiltonian? Hamiltonian learning is central towards gaining a mechanistic understanding of how particle-to-particle interactions drive interesting quantum phenomena and plays a key role in controlling and engineering quantum devices [72, 73, 29, 15].

Whereas previous algorithms for Hamiltonian learning were intractable in the challenging, but especially important low-temperature regime [15, 103, 7, 13], we present a new algorithm that overcomes these previously conjectured barriers, addressing a key computational bottleneck towards understanding and harnessing the power of quantum systems. Crucial to overcoming these barriers, our algorithm combines a suite of modern tools from computational learning theory including a new polynomial approximation and the powerful sum-of-squares semidefinite programming hierarchy.

3.1 Results

Throughout this chapter, we will work with a $(\mathfrak{d}, \mathfrak{K})$-low intersection Hamiltonian H (recall Definition 1.4.14) and we will assume that the terms of the Hamiltonian H_a are written in the form $\lambda_a E_a$ where E_a is a non-identity Pauli operator supported on at most $\mathfrak{K}$ qubits and $|\lambda_a| \leqslant 1$.

We assume that the structure of the Hamiltonian i.e. the terms E_a are known, but the coefficients λ_a are not. Because for a fixed support set, the Pauli operators with that support form an orthonormal basis for the space of operators with that support, this decomposition into Pauli operators is equivalent to saying that the "geometry", i.e. the subsets of particles that interact, is known, but the interaction strengths are unknown.

We will use the shorthand $N = 2^{\mathfrak{n}}$. The main question that we seek to address is stated below.

Problem 1 (Hamiltonian learning)**.** Let $H = \sum_{a=1}^{m} \lambda_a E_a \in \mathbb{C}^{N \times N}$ be a $(\mathfrak{d}, \mathfrak{K})$-low intersection Hamiltonian on $\mathfrak{n}$ qubits whose terms E_a are known, distinct, non-identity Pauli operators, and whose coefficients $\lambda_a \in \mathbb{R}$ satisfy $|\lambda_a| \leqslant 1$.

Given $\varepsilon, \delta > 0$, along with copies of the Gibbs state $\rho = \frac{e^{-\beta H}}{\operatorname{tr} e^{-\beta H}}$ corresponding to H at a known inverse temperature $\beta > 0$, find estimates $\hat{\lambda}_a$ such that, with probability $\geqslant 1 - \delta$, $(\hat{\lambda}_a - \lambda_a)^2 \leqslant \varepsilon^2$ for all $a \in [m]$.

We are interested in both the number of copies of ρ that we need, which is called the sample complexity, as well as the running time of the algorithm. Our main focus for Hamiltonian learning will be in the *low-temperature* regime, where β is an arbitrarily large constant.

Our main result is a computationally efficient algorithm for Hamiltonian learning that works at all temperatures.

Theorem 3.1.1 (Efficiently learning a quantum Hamiltonian, Theorem 3.7.1 (infor-

mal)). *Given $\varepsilon > 0$, $\beta \geqslant \beta_c$, for a fixed universal constant $\beta_c > 0$, and $\mathfrak{n}$ copies of the Gibbs state of a low-intersection Hamiltonian, $H = \sum_{a\in[m]} \lambda_a E_a$, there exists an algorithm that runs in time $\mathfrak{n}^{\mathcal{O}(1)}$ and outputs estimates $\left\{ \hat{\lambda}_a \right\}_{a\in[m]}$ such that with probability at least 99/100, for all $a \in [m]$, $\left|\lambda_a - \hat{\lambda}_a\right| \leqslant \varepsilon$, whenever $\mathfrak{n} \geqslant \mathrm{poly}\left(m, (1/\varepsilon)^{2^{\mathcal{O}(\beta)}}\right)$.*

Remark 3.1.2 (On temperature)**.** For our algorithm, we only need to know an upper bound on β, as we can consider a Gibbs state at temperature β to be a Gibbs state at temperature, say, 2β with Hamiltonian $H/2$. Our requirement that $\beta > \beta_c$ is for simplicity, and β_c can be any constant bounded away from zero. In particular, we can take β_c to be the temperature at which the high-temperature algorithm of [103] fails; so, when $\beta < \beta_c$, we can simply appeal to [103] to achieve a sample and time complexity of $\frac{\log m}{\beta^2\varepsilon^2}$ and $\frac{m \log m}{\beta^2\varepsilon^2}$, respectively.

As noted by prior work [15], Hamiltonian learning is a generalization of the classical and well-studied problem of learning undirected graphical models, specifically parameter learning of these models. This classical problem requires $\frac{e^{\mathcal{O}(\beta)} m \log(m)}{\beta^2\varepsilon^2}$ time (and there is an algorithm matching the lower bound), so exponential dependence on β is necessary [103].* Analogies with the classical setting turn out to be of limited use, since the non-commutativity and non-locality inherent in the quantum setting rules out generalizations of classical ideas. However, with the classical setting we can identify barriers to designing a time-efficient algorithm.

A key challenge of time-efficient Hamiltonian learning is that we cannot work directly with the partition function. The only previous approach to low temperature [15] only used its copies of ρ to estimate $\mathrm{tr}(E_a\rho)$ for all $a \in [m]$. It is known in the classical literature that taking just these estimates and using them to compute the parameters λ_a is as hard as computing the partition function [163]. To avoid this barrier, we take a richer set of expectations $\mathrm{tr}(P\rho)$ that allows us to reduce learning to a tractable, but fairly involved, optimization problem instead. Along the way we develop several new tools of independent interest, and ultimately give a semi-definite

*It is an interesting open question to improve our doubly exponential dependence to singly exponential.

programming algorithm based on the sum-of-squares hierarchy. Consequently, we show that sophisticated modern tools in optimization theory lead to a surprising resolution of the Hamiltonian learning problem.

3.1.1 Additional Related Work

Hamiltonian learning. Hamiltonian learning is a broad topic studied both in experimental and theoretical contexts. This work fits into a large body of algorithmic research about learning properties of quantum states modeling physical systems [72, 13]. Here, we point to a few lines of related work in this field.

Hamiltonian learning often focuses on the real-time evolution setting, where one can allow the system to evolve with respect to H, applying the unitary $e^{-\mathrm{i}Ht}$ [73, 212, 121]. Some algorithms consider taking time derivatives (i.e. taking $t \to 0$), which are similar to small-β algorithms in the Gibbs state setting [224, 103, 97]. There is some research on learning from (zero-temperature) ground states [182], but the algorithmic work is limited because the ground state of a Hamiltonian need not determine the Hamiltonian. We study the finite temperature case, which is both the typical temperature at which experiments are run and, in the $\beta \to \infty$ limit, a rich approximation to the much less computationally tractable ground state [7, 91].

Though our algorithm is not practical, we use constraint systems that bear some similarity to the "correlation matrix" linear constraint systems analyzed heuristically and experimentally in prior work [29, 182]. In fact, our constraint system contains these constraints for technical reasons. Our work places these works on a rigorous basis, as we prove that, though the linear constraint systems might not uniquely identify the true Hamiltonian, adding more, similar constraints eventually fully constrains the Hamiltonian.

Bounding correlations in Gibbs states. Though classical Gibbs states have extremely good locality properties, these become much weaker in the quantum setting. A series of works aims at bounding the non-locality in quantum Gibbs states with

various different measures and in various different regimes [133, 145, 146], often with the goal of concluding that simulating or learning these systems can be done time-efficiently. It is an interesting open problem whether one can extract a new kind of "locality" statement from our algorithm, to understand how general our approach is for learning quantum systems. Our polynomial approximation is inspired by proofs of the Lieb–Robinson bound [152, 110], and can be viewed as a "low-degree" form of this bound. This could be of independent interest.

Parameter learning of graphical models. There is a rich body of work on the problem of learning graphical models. Our setting is that of learning Markov random fields; the literature on this topic focuses on the task of *structure learning*, which in our setting corresponds to learning the terms $\{E_a\}_{a\in[m]}$, given the guarantee that they form an (unknown) dual interaction graph with bounded degree [53, 52, 105, 138]. The problem we consider, learning the parameters with known terms, is easy in the classical setting [103, Appendix B], because classical Gibbs states satisfy the Hammersley–Clifford theorem [106], also known as the Markov property. A consequence of the Markov property is that estimating a parameter on a $\mathfrak{K}$-body term can be done by computing conditional marginals on the support of this term. It is not clear how to generalize this argument to the quantum setting, since the Markov property does not hold for low-temperature quantum Hamiltonians, even approximately [145].

The sum-of-squares meta-algorithm. The sum-of-squares hierarchy has been used to analyze several problems in quantum information, including best state separation [77, 51, 37, 39], optimizing fermionic Hamiltonians [112, 111], and a quantum analogue of max-cut [175, 209]. Additionally, the proofs-to-algorithms perspective, introduced in [38, 35], has been extensively used to design efficient algorithms for estimation and learning tasks, with many recent advances especially in mixture models and robust statistics [116, 142, 139, 154, 30].

3.2 Proof Overview

The recipe for quantum Hamiltonian learning introduced by Anshu, Arunachalam, Kuwahara, and Soleimanifar [15] is based on matching the local marginals of the Gibbs state ρ, which we can estimate with copies of ρ. Specifically, for two Hamiltonians $H = \sum \lambda_a E_a$ and $H' = \sum \lambda'_a E_a$ with respective Gibbs states

$$\rho = \frac{e^{-\beta H}}{\mathrm{tr}(e^{-\beta H})} \text{ and } \rho' = \frac{e^{-\beta H'}}{\mathrm{tr}(e^{-\beta H'})},$$

they show that $H = H'$ (and so $\lambda_a = \lambda'_a$ for all $a \in [m]$) if and only if ρ and ρ' are identical on local marginals i.e. $\mathrm{tr}(E_a\rho) = \mathrm{tr}(E_a\rho')$ for all $a \in [m]$ [15, Proposition 4]. This does not imply a bound on sample complexity, because with copies of ρ, we can only compute $\mathrm{tr}(E_a\rho)$ approximately, with noise introduced from sampling error. The key structural result of [15] is that this equivalence can be made robust, so that if H' only approximately matches marginals, then the corresponding coefficients λ'_a approximately match the true coefficients λ_a.

However, the last step of this algorithm is to invert the map $\{\lambda_a\}_{a\in[m]} \mapsto \{\mathrm{tr}(E_a\rho)\}_{a\in[m]}$, which is a computationally hard problem. Formally, for a classical Hamiltonian,[†] $\mathrm{tr}(E_a\rho)$ are *sufficient statistics* of a graphical model and it is known that estimating the parameters of a graphical model from these sufficient statistics is computationally intractable [163]. This doesn't mean that the problem is hopeless, but rather that to find a tractable algorithm, we should be looking for the opportunity to use a richer family of statistics.

Designing a new system of constraints. We interpret the previous argument as defining and then solving a constraint system in the set of unknowns, $\{\lambda'_a\}_{a\in[m]}$. The structural result in [15] shows that an approximate solution to this system will be close to the true parameters λ_a. However, this system is computationally hard to solve.

[†]A classical Hamiltonian is a Hamiltonian that is diagonal, i.e. its terms are tensor products of the identity and σ_z (Definition 1.4.11). For a classical Hamiltonian, the state ρ is a sample from the Gibbs distribution, and $\mathrm{tr}(E_a\rho)$ is a $\mathfrak{K}$-point correlation function.

Our starting point is to define a larger set of constraints which $\{\lambda_a\}_{a\in[m]}$ must satisfy, which can be verified by measuring expectations of observables slightly less local than the terms $\{E_a\}_{a\in[m]}$. Let $\mathcal{P}_{\text{local}}$ be the set of Pauli matrices whose support is K-local for some large constant K. We begin with the following system of constraints:

$$\left\{\begin{array}{ll} \forall a \in [m] & -1 \leqslant \lambda'_a \leqslant 1 \\ & H' = \sum_{a\in[m]} \lambda'_a \cdot E_a \\ \forall P, Q \in \mathcal{P}_{\text{local}}, & \operatorname{tr}\Big(Qe^{-\beta H'} P e^{\beta H'} \rho\Big) = \operatorname{tr}(PQ\rho) \end{array}\right\}, \tag{3.1}$$

The constraints above are indeed satisfied for the true parameters ($\lambda' = \lambda$) since by assumption $|\lambda_a| \leqslant 1$ for all $a \in [m]$ and moreover

$$\operatorname{tr}\big(Qe^{-\beta H} P e^{\beta H} \rho\big) = \operatorname{tr}\left(Qe^{-\beta H} P e^{\beta H} \frac{e^{-\beta H}}{\operatorname{tr}(e^{-\beta H})}\right) = \operatorname{tr}(PQ\rho)$$

which follows from the cyclic property of the trace. Two main challenges remain: Must a solution to this system be close to the true coefficients? And how can we efficiently solve the system? Eventually we will derive a convex relaxation for it that is based on

(A) replacing the last constraint in Eq. (3.1) which involves the matrix exponential with low degree polynomial constraints on the indeterminates ($\{\lambda'_a\}_{a\in[m]}$) instead and

(B) showing that any choice of λ' that satisfy the constraints must also approximately match the true coefficients λ.

In general solving systems of polynomial equations is computationally hard, but because our analysis in (B) will be based on sum-of-squares proofs, there is by now standard machinery for turning it into an efficient algorithm (see Section 3.3.4 for detailed explanation).

Identifying an equivalence between nested commutators and polynomials. Working towards the goal of replacing the term $\operatorname{tr}\big(Qe^{-\beta H'} P e^{\beta H'} \rho\big)$ with a low-degree

polynomial in the variables λ', we begin by recalling the Hadamard formula:[‡]

$$e^{-\beta H'} P e^{\beta H'} = \sum_{\ell=0}^{\infty} \beta^\ell \frac{[H', P]_\ell}{\ell!}, \tag{3.2}$$

where $[H', P]_\ell = [H', [H', \ldots, [H', P] \ldots]]$ is the ℓ-th nested commutator. A natural first step is to truncate this series at d terms and observe that

$$\operatorname{tr}\left(Q \left(\sum_{\ell=0}^{d} \beta^\ell \frac{[H', P]_\ell}{\ell!} \right) \rho \right)$$

is a low-degree polynomial in the variables λ'. For instance, observe for the order-2 nested commutator, we have

$$\begin{aligned} \operatorname{tr}\Big(\Big[\sum_{a \in [m]} \lambda'_a E_a, P \Big]_2 \rho \Big) &= \operatorname{tr}\Big(\Big[\sum_{a \in [m]} \lambda'_a E_a, \sum_{a \in [m]} \lambda'_a [E_a, P] \Big] \rho \Big) \\ &= \sum_{a,b \in [m]} \lambda'_a \lambda'_b \operatorname{tr}([E_a, [E_b, P]]\, \rho), \end{aligned}$$

which is a degree-2 polynomial in the λ'_i indeterminates. However, the series in Eq. (3.2) only converges quickly when β is sufficiently small [103], so we cannot use it.[§]

Nevertheless, from this observation we can develop a general formalism for constructing polynomial approximations of evolutions of operators. We observe that in the eigenbasis of H',

$$[H', P]_\ell = P \circ \left\{ \, (\sigma_i - \sigma_j)^\ell \, \right\}_{ij},$$

where $\{ \, \sigma_i \, \}_{i \in [N]}$ are the eigenvalues of H' and $\circ$ denotes the Hadamard product

[‡]This can be derived from the Baker–Campbell–Hausdorff formula,

$$\exp(A) \exp(B) = \exp\Big(A + B + \frac{1}{2}[A, B] + \frac{1}{12}([A, [A, B]] + [B, [B, A]]) + \ldots \Big).$$

[§]This expansion does converge after $\beta \|H\|$ terms, but our running time is exponential in the degree, so this would be far too large.

(Definition 1.4.1). Similarly,

$$e^{-\beta H'} P e^{\beta H'} = P \circ \left\{ e^{-\beta(\sigma_i - \sigma_j)} \right\}_{ij},$$

and thus we can focus our attention on designing polynomials that approximate the scalar quantity $e^{-\beta(\sigma_i - \sigma_j)}$ with low-degree polynomials in $(\sigma_i - \sigma_j)$. Further, any degree-d polynomial $p(z) = \sum_{\ell=0}^{d} c_\ell z^\ell$ can be extended to commutators as follows:

$$p(H' \mid P) = P \circ \{ p(\sigma_i - \sigma_j) \}_{ij} = \sum_{\ell=0}^{d} c_\ell [H', P]_\ell.$$

This allows us to translate between matrix series expansions involving nested commutators and univariate polynomials. We note that for technical reasons we need to extend our equivalence to nested commutators with two distinct operators X, Y appearing in an arbitrary order, such as $[X, [Y, [X, \ldots] \ldots], E_{a_1}]$ and bi-variate polynomials $p(x, y)$ (Section 3.4). The translation between bi-variate polynomials and nested commutators incurs additive error depending on $[X, Y]$ due to re-ordering of the X and Y operators as expected (Theorem 3.4.9). In our full algorithm (Section 3.7) we introduce an additional constraint to drive this additive error to zero. Focusing on the scalar polynomial approximation to the exponential function, we now formalize the notion of approximation that we require.

Constructing a new, flat approximation to the exponential. Recall that we want a polynomial such that, working in the eigenbasis of H',

$$p(H' \mid P) = P \circ \{ p(\sigma_i - \sigma_j) \}_{ij} \approx P \circ \left\{ e^{-\beta(\sigma_i - \sigma_j)} \right\}_{ij} = e^{-\beta H'} P e^{\beta H'}, \tag{3.3}$$

where "$\approx$" denotes an unusual notion of approximation which, for the purposes of this discussion, we can consider to mean that the matrices are close in some norm. The Taylor series approximation to the exponential would be Eq. (3.2), which we established is too high degree.

Our key insight is that we choose a better polynomial approximation. We begin by observing that an operator with small support is approximately band-diagonal in the basis of eigenvectors of H, which is a property of local terms proved by Arad, Kuwahara, and Landau [20]. We state a weak version of this here: let P be a Pauli operator with support size $\mathcal{O}(1)$, and let $H = \sum_i d_i v_i v_i^\dagger$ be an eigendecomposition of H. Then, considering P in the eigenbasis of H,

$$|P_{ij}| = |v_i^\dagger P v_j| \leqslant e^{-\Omega(|d_i - d_j|)}. \qquad \text{(Corollary 3.3.10)}$$

A consequence of this is that entries of $P \circ \left\{ \, p(\sigma_i - \sigma_j) - e^{-\beta(\sigma_i - \sigma_j)} \, \right\}_{ij}$, which denote the error of the polynomial approximation, are *weighted inverse exponentially in* $\sigma_i - \sigma_j$. Therefore, our polynomial approximation need not be equally good for all $\sigma_i - \sigma_j$; rather, our approximation should be ε-good in a small range but is allowed to diverge at a sufficiently slow exponential rate outside that range. We call this a *flat* approximation. In particular, given parameters $\beta \geqslant 0$, $0 < \varepsilon, \eta < 1$, we construct p such that

$$\begin{cases} \left| p(z) - e^{-\beta z} \right| \leqslant \varepsilon & \text{if } z \in [-1, 1] \\ |p(z)| \leqslant \max\left(1, e^{-\beta z}\right) \cdot e^{\eta \beta |z|} & \text{if } z \notin [-1, 1] \end{cases} \tag{3.4}$$

The key difficulty in satisfying the above constraints is satisfying $|p(z)| \leqslant e^{\eta\beta|z|}$ for $z \geqslant \beta$, as standard approximations like Taylor series truncations and Chebyshev series truncations fail this condition (Remark 3.5.3). In Section 3.5, we explicitly construct a degree-$\left(2^{\mathcal{O}(1/\eta)} \cdot (\beta + \log(1/(\varepsilon\eta)))\right)$ polynomial that satisfies Eq. (3.4). This construction is inspired by the iterative "peeling" of the exponential used in proofs of Lieb-Robinson bounds [152, 110]. We can write

$$e^{-\beta z} = \underbrace{e^{-\beta_c z} \cdots e^{-\beta_c z}}_{\beta/\beta_c}$$

for a fixed small constant β_c and then truncate the Taylor series expansion of $e^{-\beta_c z}$ at different scales for all of the $\beta/\beta_c = O(\beta)$ copies in the product so that the tails of the different truncations don't "interfere".

We show that when p is a flat approximation of the above form for some sufficiently small η, then $Qp(H|P)\rho$ is a good approximation to $Qe^{-\beta H}Pe^{\beta H}\rho$. In other words, the polynomial approximation is good when $H' = H$ and we right multiply by ρ; this is crucial for the polynomial system that we set up next to be feasible.

Formulating a polynomial system. We now have all the tools to describe a polynomial system that captures the Hamiltonian learning problem. The constraint system we describe in this section is an informal treatment of the system that appears in Section 3.7, and avoids several technical details. We show that using our flat approximation to the exponential, we can obtain a polynomial p such that

$$\operatorname{tr}\big(Qe^{-\beta H}Pe^{\beta H}\rho\big) \approx \operatorname{tr}(Qp(H \mid P)\rho).$$

Then, we can re-write Eq. (3.1) as the following polynomial constraint system:

$$\left\{\begin{array}{ll} \forall a \in [m] & -1 \leqslant \lambda'_a \leqslant 1 \\ & H' = \displaystyle\sum_{a\in[m]} \lambda'_a E_a \\ \forall P, Q \in \mathcal{P}_{\text{local}}, & |\operatorname{tr}(Qp(H' \mid P)\rho) - \operatorname{tr}(PQ\rho)| \leqslant \varepsilon \end{array}\right\}, \tag{3.5}$$

and observe that the last constraint encodes a relaxation of the last constraint in Eq. (3.1) and is satisfied when $H' = H$. Further, all of the constraints are indeed succinctly representable as low-degree polynomials in the indeterminates, $\{\, \lambda'_a \,\}_{a\in[m]}$, as discussed earlier. Finally, the coefficients, such as $\operatorname{tr}([E_a, [E_b, P]])$, are expectations of the Gibbs state against a slightly larger set of local observables, which are the richer class of test functions we desired. We can obtain estimates of these expectations through quantum measurements (Section 3.6). Computing these estimates is the only quantum part of our algorithm, and the rest of the algorithm is entirely classical.

Feasibility of the polynomial system. Recall that to show that the polynomial system in (3.5) is feasible, we need to argue that

$$\operatorname{tr}\big(Qe^{-\beta H}Pe^{\beta H}\rho\big) \approx \operatorname{tr}(Qp(H \mid P)\rho)$$

for all P, Q. Working in the eigenbasis of H, let its eigenvalues be $\{ \sigma_i \}_{i\in[2^n]}$. The key tool that we leverage is from [20] (see Corollary 3.3.10) which roughly states that any local term E must be approximately diagonal in the eigenbasis of H, with off-diagonal entries decaying as $|E_{ij}| \leqslant e^{-\Omega(|\sigma_i-\sigma_j|)}$. Thus, we can decompose the matrices Q, P into two parts – parts indexed by i, j where $|\sigma_i - \sigma_j| \leqslant \beta$ and parts indexed by i, j where $|\sigma_i - \sigma_j| \geqslant \beta$. Then we use the fact that $p(x)$ is a good approximation to e^{-x} on $[-\beta, \beta]$ to prove that the error on the first part is small. We then appeal to the exponential decay of the off-diagonals to argue that the contribution from the second part in both $\operatorname{tr}\big(Qe^{-\beta H}Pe^{\beta H}\rho\big)$ and $tr(Qp(H \mid P)\rho)$ is small. Our flat approximation to the exponential is designed to ensure that it does not overwhelm the exponential decay in the off-diagonal entries in P, Q in any regime.

Efficiently optimizing polynomial systems. Now that we know that our polynomial system is feasible, we consider a convex relaxation of this system. In particular, we consider a degree-d sum-of-squares relaxation, which can be efficiently optimized by expressing it as a semi-definite program (see Section 3.3.4 for details), with $d = \log(1/\varepsilon) \cdot 2^{\mathcal{O}(\beta)}$. Since we have m variables and $2^{\mathcal{O}(\beta)}$ constraints, and each constraint is a degree-d polynomial, we can solve the degree-$2d$ sum-of-squares relaxation of Eq. (3.5) in $m^{\left(\log(1/\varepsilon)\cdot 2^{\mathcal{O}(\beta)}\right)}$ time. The main challenge in analyzing the sum-of-squares relaxation is to show that we can *round* it to estimates $\left\{ \tilde{\lambda}'_a \right\}_{a\in[m]}$ such that they are close to the true parameters. Here, we adopt the so-called *proofs-to-algorithms* philosophy, where we instead work with the dual object to the sum-of-squares relaxation, namely sum-of-squares proofs (see [35, 87], and references therein). This perspective states that if the true parameters are identifiable only using the sum-of-squares proof system, then we immediately obtain an efficient algorithm and we show that we can

easily and accurately *round* the solution.

We then provide a proof of identifiability, i.e. for all $a \in [m]$, the inequality $(\lambda'_a - \lambda_a) \leqslant \varepsilon$ can be derived using the system of polynomial constraints and other basic inequalities that admit sum-of-squares proofs (we refer the reader to Section 3.10 for a detailed exposition). At a high level, the proof works by arguing that when $H' - H$ is large, there are witnesses P, Q such that

$$|\operatorname{tr}(Qp(H'|P)\rho) - \operatorname{tr}(Qp(H|P)\rho)|$$

is large. Since we know that H is a feasible solution, this would imply that H' cannot be a feasible solution so any feasible solution must have $H' - H$ be small. The construction of the witnesses relies on an additional property of the polynomial p that we construct, namely that it is strongly monotone (in some appropriate quantitative sense).

For the identifiability proof, we crucially use an additional important property of local Hamiltonians. It deals with the quantity $\operatorname{tr}(A^2\rho)$, where $A = \sum_b \sigma_b P_b$ is a Hermitian linear combination of Pauli matrices with small support. Thinking of ρ as a distribution, $\operatorname{tr}(A^2\rho)$ is a second moment term with respect to ρ; we can prove this is not much smaller than $\operatorname{tr}(A^2 I/\dim) = \sum_b \sigma_b^2$, the second moment against the uniform distribution: for some constant $c > 0$,

$$\operatorname{tr}(A^2\rho) \geqslant c^{\mathcal{O}(\beta)} \max_b \sigma_b^2. \qquad \text{(Theorem 3.3.7)}$$

Intuitively, this shows that ρ is not close to zero in any local direction. This was first shown by [15] for quasi-local operators; we adjust their proof to hold for just local operators and give a tighter bound. We show that we can obtain a slightly weaker statement of this form in the sum-of-squares proof system by formulating it as a quadratic inequality. This inequality can be used to remove the dependence on ρ in expressions appearing in the proof; for example, it is used to relate $\operatorname{tr}([H, H']^2\rho)$ to the size of $[H, H']$ itself.

Finally, we observe that our identifiability proof does not use the full power of a degree-$2d$ sum-of-squares relaxation and therefore, it should suffice to solve a significantly smaller semi-definite program. We show that we can execute our proof of identifiability by only appealing to a sparse subset of monomials of degree at most $2d$ and invoke a linearization theorem by Steurer and Teigel [198] to obtain a final running time of $\mathrm{poly}(m) \cdot (1/\varepsilon)^{2^{\mathcal{O}(\beta)}}$, as desired.

3.3 Basic Tools

We define the following notation to extract the piece of an operator that acts on a particular qubit.

Definition 3.3.1 (Localizing an operator)**.** For an operator $O \in \mathbb{C}^{N \times N}$, define

$$O_{(i)} = O - \mathrm{tr}_i(O) \otimes \frac{I_i}{2}, = O - \int \mathrm{d}\mu_i(U) U^\dagger O U,$$

where I_i denotes the identity operator on the ith qubit, tr_i denotes the partial trace operation with respect to the ith qubit, and μ_i denotes the Haar measure over the set of unitaries only supported on qubit i. In other words, $[\cdot]_{(i)} : (\mathbb{C}^{2\times 2})^{\otimes \mathfrak{n}} \to (\mathbb{C}^{2\times 2})^{\otimes \mathfrak{n}}$ is the linear map on operators that is the identity on every qubit but i, and on the ith qubit maps $M \mapsto M - \frac{1}{2}\mathrm{tr}(M) I$ for $M \in \mathbb{C}^{2\times 2}$.

For a tensor product of Pauli matrices, $P \in \mathcal{P}$, $P_{(i)}$ is P when $i \notin \mathrm{supp}(P)$ and 0 otherwise. So, for a linear combination of Paulis, $A = \sum_{P \in \mathcal{P}} \lambda_P P$, applying this map restricts the sum to Pauli matrices that interact with qubit i:

$$A_{(i)} = \sum_{P : i \in \mathrm{supp}(P)} \lambda_P P \tag{3.6}$$

So, $|\mathrm{supp}(A_{(i)})| \leqslant (\mathfrak{d}+1)\mathfrak{K}$.

Definition 3.3.2 (Projector onto eigenspaces)**.** For a Hermitian matrix X and an interval $\mathcal{I} \subset \mathbb{R}$, $\Pi_{\mathcal{I}}^{(X)}$ denotes the orthogonal projector onto the subspace spanned by

eigenvectors of X with eigenvalues in $\mathcal{I}$.

3.3.1 Local Operators

We will be considering operators other than the terms, so we define a notion of locality with respect to the dual interaction graph that generalizes to such operators. Concretely, what we need about ℓ-local operators for $\ell \geqslant \mathfrak{K}$ is that (1) they contain the Hamiltonian terms E_a, (2) the dimension of the subspace spanned is $O(\mathfrak{n})$, and (3) they contain nested commutators involving $\mathfrak{K}$-local operators.

Definition 3.3.3 (Local operator with respect to the dual interaction graph)**.** Consider a $\mathfrak{K}$-local Hamiltonian $H = \sum_{a=1}^{m} \lambda_a E_a$ with dual interaction graph $\mathfrak{G}$. For P a tensor product of Paulis, we say it is $\mathfrak{K}\ell$-$\mathfrak{G}$-local if there is some $S \subset [m]$ of size ℓ such that $\mathrm{supp}(P) \subset \cup_{a \in S} \mathrm{supp}(E_a)$ and S is connected in $\mathfrak{G}$.

We define $\mathcal{P}_{\mathfrak{K}\ell}$ to be the set of $\mathfrak{K}\ell$-$\mathfrak{G}$-local Pauli matrices. Generally, we call an operator $M \in \mathbb{C}^{N \times N}$ $\mathfrak{K}\ell$-$\mathfrak{G}$-local if it is equal to a linear combination of elements in $\mathcal{P}_{\mathfrak{K}\ell}$.

By this definition, if E_a and E_b are terms of a Hamiltonian, then E_b is $\mathfrak{K}$-$\mathfrak{G}$-local, and $[E_a, E_b]$ is $2\mathfrak{K}$-$\mathfrak{G}$-local.

Pauli matrices behave straightforwardly under commutation: the commutator of two Pauli matrices is another Pauli matrix up to a scalar. We state the form of nested commutators of Pauli matrices below.

Lemma 3.3.4. *Let $P_1 \in \mathcal{P}_{k_1}, \ldots, P_a \in \mathcal{P}_{k_a}$, and $Q \in \mathcal{P}_\ell$ (with respect to some background dual interaction graph $\mathfrak{G}$). Then the nested commutator is either zero or also a $\mathfrak{G}$-local Pauli matrix,*

$$\frac{\mathrm{i}^a}{2^a}[P_1, [P_2, \ldots [P_a, Q] \ldots]] \in \mathcal{P}_{k_1 + \cdots + k_a + \ell},$$

possibly up to a factor of ± 1 where $\mathrm{i} = \sqrt{-1}$.

Proof. Let $a = 1$. By properties given in Definition 1.4.11, for $P \in \mathcal{P}_{k_1}$ and $Q \in \mathcal{P}_\ell$,

PQ is a tensor product of Pauli matrices, up to fourth root of unity. Consequently, $[P, Q] = PQ - (PQ)^\dagger$ is either zero (if PQ is Hermitian) or $2PQ$ (if $\mathrm{i}PQ$ is Hermitian). Further, $\mathrm{supp}([P, Q]) \subset \mathrm{supp}(PQ) \subset \mathrm{supp}(P) \cup \mathrm{supp}(Q)$, which shows that $[P, Q]$ is $(k_1 + \ell)$-$\mathfrak{G}$-local. This proves the lemma for $a = 1$; the general statement follows by iterating this case. □

3.3.2 Properties of Local Operators on Quantum Systems

For a Hamiltonian describing a quantum system, we consider getting access to the associated state attained by thermalizing it at a particular inverse temperature β. This is known as a Gibbs state.

Definition 3.3.5 (Gibbs state)**.** The *Gibbs state* of the Hamiltonian $H = \sum \lambda_a E_a$ at inverse temperature $\beta > 0$ is given by

$$\rho = \frac{\exp(-\beta H)}{\operatorname{tr} \exp(-\beta H)} = \exp\left(-\beta \sum_a \lambda_a E_a\right) \Big/ \operatorname{tr} \exp\left(-\beta \sum_a \lambda_a E_a\right). \tag{3.7}$$

A useful piece of intuition is to think of ρ as a distribution over its eigenspaces with probability proportional to the eigenvalue. In that sense, we can think about expectations and variances against this distribution.

Definition 3.3.6 (Expectation against the Gibbs state)**.** For a Gibbs state ρ of a Hamiltonian H and an operator $A \in \mathbb{C}^{N \times N}$, we define $\langle A \rangle = \operatorname{tr}(A\rho)$.

A key result in the prior work of Anshu, Arunachalam, Kuwahara, and Soleimanifar gives a lower bound on the variance of a local operator with respect to the energy distribution defined by the Gibbs state. The authors prove this for the more general class of quasi-local operators, but we only need it for local operators, for which the result can be tightened. We give this tighter version below; its proof is in Appendix B.1.

Theorem 3.3.7 ([15, Theorem 33])**.** *Let H be a $\mathfrak{K}$-local Hamiltonian with dual interaction graph $\mathfrak{G}$ with max degree $\mathfrak{d}$. Let $A = \sum_b \sigma_b P_b$ be a $\mathfrak{K}$'-local operator where*

P_b are products of Pauli matrices and $-1 \leqslant \sigma_b \leqslant 1$ and whose dual interaction graph has max degree $\mathfrak{d}'$. For $\beta > 0$, let ρ be the corresponding Gibbs state of H. Then

$$\langle A^2 \rangle = \operatorname{tr}(A^2 \rho) \geqslant \max_{i \in [\mathfrak{n}]} \Big(c \operatorname{tr}(A_{(i)}^2 / N) \Big)^{6 + c'\beta}.$$

Here, c and c' are positive constants that depend on $\mathfrak{K}$, $\mathfrak{d}$, $\mathfrak{K}'$, and $\mathfrak{d}'$.

After some manipulation we can conclude that, if a local operator has small variance with respect to ρ, then the operator itself must be close to zero. Some may recognize this as the quantum analogue of the statement that bounded-degree graphical models have local marginals bounded away from zero (see e.g. [52]).

Corollary 3.3.8 ("No small local marginals")**.** *Let H be a $\mathfrak{K}$-local Hamiltonian with dual interaction graph $\mathfrak{G}$ with max degree $\mathfrak{d}$. Let $A = \sum_{P \in \mathcal{P}} \sigma_P P$ be a sum over Paulis with support at most $\mathfrak{K}'$ and with a dual interaction graph $\mathfrak{G}'$ with max degree $\mathfrak{d}'$, with coefficients $\sigma_P \in \mathbb{R}$. Then*

$$\operatorname{tr}(A^2 \rho) \geqslant \exp(-c_{\mathfrak{K},\mathfrak{d},\mathfrak{K}',\mathfrak{d}'} - c'_{\mathfrak{K},\mathfrak{d},\mathfrak{K}',\mathfrak{d}'} \beta) \max_{Q \in \mathcal{P}} \sigma_Q^2$$

where $c_{\mathfrak{K},\mathfrak{d},\mathfrak{K}',\mathfrak{d}'}, c'_{\mathfrak{K},\mathfrak{d},\mathfrak{K}',\mathfrak{d}'}$ are constants depending only on $\mathfrak{K}, \mathfrak{d}, \mathfrak{K}', \mathfrak{d}'$.

Proof. Let $g^2 = \max_{Q \in \mathcal{P}_\ell} \sigma_Q^2$.

$$\begin{aligned}
\operatorname{tr}(X^2 \rho) &= g^2 \operatorname{tr}((X/g)^2 \rho) \\
&\geqslant g^2 \max_{i \in [\mathfrak{n}]} \Big(c \operatorname{tr}((X_{(i)}/g)^2 / N) \Big)^{6 + c'\beta} && \text{by Theorem 3.3.7} \\
&= g^2 \max_{i \in [\mathfrak{n}]} \Big(c \operatorname{tr}\Big(\Big(\sum_{\substack{P \in \mathcal{P}_\ell \\ i \in \operatorname{supp}(P)}} \frac{\sigma_P}{g} P \Big)^2 / N \Big) \Big)^{6 + c'\beta} && \text{by Eq. (3.6)} \\
&= g^2 \max_{i \in [\mathfrak{n}]} \Big(c \sum_{\substack{P \in \mathcal{P}_\ell \\ i \in \operatorname{supp}(P)}} \frac{\sigma_P^2}{g^2} \Big)^{6 + c'\beta} && \text{by Definition 1.4.11} \\
&\geqslant g^2 c^{6 + c'\beta}.
\end{aligned}$$

This gives the desired statement. □

Another key statement we use about local operators is that a local operator is approximately diagonal when considered in the basis of eigenvectors of another local operator.

Theorem 3.3.9 ([20, Theorem 2.1])**.** *Let $H = \sum_{S \subset [\mathfrak{n}]} h_S$ be a Hamiltonian where all its terms h_S are positive semi-definite, supported on at most $\mathfrak{K}$ qubits, and the terms interacting with any particular site have bounded norm: $\sum_{S:i \in S} \|h_S\| \leqslant g$ for all $i \in [\mathfrak{n}]$. Let A be an operator, and define $R = \sum_{X \in \mathcal{C}} \|h_X\|$, where $\mathcal{C} = \{S \subset [\mathfrak{n}] \mid [h_S, A] \neq 0\}$ is the set of terms that don't commute with A. Then*

$$\|\Pi^{(H)}_{[\sigma,\infty)} A \Pi^{(H)}_{[0,\varsigma]}\| \leqslant \|A\| e^{-\frac{1}{2g\mathfrak{K}}(\sigma - \varsigma - 2R)}.$$

The assumption that the h_S are PSD can be removed: if the h_S are not PSD, we can apply the theorem to $h_S + I\|h_S\|$, which only affects the final bound by inflating g and R by a factor of two. Note that this just leads to shifting H by a factor of the identity, and so only shifts the spectrum.

Corollary 3.3.10. *Let $H = \sum_{S \subset [\mathfrak{n}]} h_S$ be a Hamiltonian where all its terms h_S are supported on at most $\mathfrak{K}$ qubits, and the terms interacting with any particular site has bounded norm: $\sum_{S:i \in S} \|h_S\| \leqslant g$ for all $i \in [\mathfrak{n}]$. Let A be an operator, and define $R = \sum_{X \in \mathcal{C}} \|h_X\|$, where $\mathcal{C} = \{S \subset [\mathfrak{n}] \mid [h_S, A] \neq 0\}$ is the set of terms that don't commute with A. Then*

$$\|\Pi^{(H)}_{[\sigma,\infty)} A \Pi^{(H)}_{[-\infty,\varsigma]}\| \leqslant \|A\| e^{-\frac{1}{4g\mathfrak{K}}(\sigma - \varsigma - 4R)}.$$

When A satisfies $|\mathrm{supp}(A)| \leqslant \mathfrak{K}'$, we can take R to be the number of terms that intersect with the support of A. In our setup, where the terms are Pauli matrices with a dual interaction graph with maximum degree $\mathfrak{d}$, we can take $g = \mathfrak{d} + 1$. If the support of A is contained in the support of a term, we can take $R = \mathfrak{d} + 1$; otherwise,

we can take $R = \mathfrak{K}'\mathfrak{d}$.

3.3.3 Bounds on Nested Commutators

One key property about commutators $[A, B]$ is that they compose well when the inputs A, B are a sum of local terms (recall Lemma 3.3.4). In this section, we will make these composition properties more precise when we have a nested commutator of the form

$$[H_1, [H_2, [\dots [H_\ell, A] \dots]]],$$

where $H_1, \dots, H_\ell$ are local operators and A has small support. We use a cluster expansion argument to bound these terms.

Definition 3.3.11. Let $M_1, \dots, M_\ell \in \mathbb{C}^{N \times N}$ be operators. We say that the ordered ℓ-tuple $(M_1, \dots, M_\ell)$ forms a cluster if, for all a, the support of M_{a+1} has nonempty intersection with $\operatorname{supp}(M_1) \cup \cdots \cup \operatorname{supp}(M_a)$.

Lemma 3.3.12. *Let $\mathcal{E} \subset \mathcal{P}$ be a set of Pauli terms where every $P \in \mathcal{E}$ satisfies $\operatorname{supp}(P) \leqslant \mathfrak{K}$ and the dual interaction graph associated with $\mathcal{E}$ has max degree $\mathfrak{d}$. Let $H_1, \dots, H_\ell$ be linear combinations of elements of $\mathcal{E}$, written as $H_i = \sum_{P \in \mathcal{E}} \lambda_{i,P} P$ for all i. Let $A \in \mathcal{E}$. Then we can write $[H_1, [H_2, [\dots [H_\ell, A] \dots]]]$ in the form*

$$2^\ell \sum_{\substack{P_1, P_2, \dots, P_\ell \in \mathcal{E} \\ (A, P_\ell, \dots, P_1) \text{ is a cluster}}} c_{P_1,\dots,P_\ell} Q_{P_1,\dots,P_\ell} \prod_{j=1}^{\ell} \lambda_{j,P_j},$$

where the constants $c_{P_1,\dots,P_m} \in \{0, \pm 1, \pm i\}$ and

(a) $Q_{P_1,\dots,P_\ell} \in \mathcal{P}_{(\ell+1)k}$ (as defined in Definition 3.3.3) and is distance ℓ from A in $\mathfrak{G}$-distance;

(b) The number of terms in the sum is at most $\ell!(\mathfrak{d}+1)^\ell$.

Proof. We can expand the nested commutator $[H_1, [H_2, [\dots [H_\ell, A] \dots]]]$, using that

the commutator is bilinear, into a sum of individual terms

$$\sum_{P_1,\dots,P_\ell} [P_1, [P_2, [\dots [P_\ell, A] \dots]]] \prod_{a=1}^{\ell} \lambda_{a,P_a} .$$

Now we argue about which of the terms in the sum above are nonzero. Note that for the commutator to be nonzero, P_ℓ must intersect the support of A, $P_{\ell-1}$ must intersect $\operatorname{supp}(P_\ell) \cup \operatorname{supp}(A)$, and so on. This condition is equivalent to $A, P_\ell, \dots, P_1$ being a cluster. Now by Lemma 3.3.4, we have

$$[P_1, [P_2, [\dots [P_\ell, A] \dots]]] \in 2^\ell c_{P_1,\dots,P_\ell} \mathcal{P}_{\ell+1}$$

for some $c_{P_1,\dots,P_\ell} \in \{0, \pm 1, \pm i\}$. Because $(A, P_\ell, \dots, P_1)$ is a cluster, we can also conclude that the elements of $\mathcal{P}_{\ell+1}$ that appear in the sum have distance $\leqslant \ell$ from A. Now it remains to count the number of clusters. When choosing P_a, we have a choices for which of $(A, P_1, \dots, P_{a-1})$, to intersect with, and $\mathfrak{d} + 1$ choices for elements of $\mathcal{E}$ that intersect with that element of $\mathcal{E}$. This gives an upper bound on the total number of clusters of

$$(\mathfrak{d} + 1) \cdot 2(\mathfrak{d} + 1) \cdots \ell(\mathfrak{d} + 1) = \ell!(\mathfrak{d} + 1)^\ell.$$

This completes the proof. □

We will also need the following lemma from [103] that counts the number of distinct multisets of terms that form a cluster (under some ordering).

Lemma 3.3.13 ([103, Proposition 3.6])**.** *Consider a set of terms $\mathcal{E} \subset \mathcal{P}$ with dual interaction graph $\mathfrak{G}$ with max degree $\mathfrak{d} \geqslant 2$. For a fixed $E_{a_1} \in \mathcal{E}$, the number of multisets of terms $E_{a_2}, \dots, E_{a_\ell} \in \mathcal{E}$ such that there is some ordering π where $(E_{a_{\pi(1)}}, \dots, E_{a_{\pi(\ell)}})$ forms a cluster is at most $e\mathfrak{d}(1 + e(\mathfrak{d} - 1))^{\ell-1} \leqslant (3\mathfrak{d})^\ell$.*

As a consequence, we can bound the number of distinct elements in $\mathcal{P}_{k\ell}$.

Corollary 3.3.14. *We have $|\mathcal{P}_{k\ell}| \leqslant m(10^k \mathfrak{d})^\ell$. Furthermore, the number of elements*

of $\mathcal{P}_{k\ell}$ that have support intersecting with a fixed term $P \in \mathcal{P}_k$ is at most $(10^k \mathfrak{d})^{\ell+1}$.

3.3.4 Sums-of-Squares Polynomials

We will need some preliminaries about polynomials and the sum-of-squares (SoS) framework. First, we introduce a notion of polynomials with sum-of-squares representations with bounded coefficients. We maintain coefficient bounds because later, certain sampling and approximation errors are multiplied by these coefficients. The general SoS framework is not concerned with such tight bounds on the coefficients and thus we need to make a few definition outside of the general framework.

Definition 3.3.15 (Sum-of-squares polynomial)**.** A polynomial $p(x_1, \ldots, x_m) \in \mathbb{R}[x_1, \ldots, x_m]$ is a *sum-of-squares polynomial* if we can write

$$p(x_1, \ldots, x_m) = q_1(x_1, \ldots, x_m)^2 + \cdots + q_k(x_1, \ldots, x_m)^2$$

for some polynomials $q_1, \ldots, q_k \in \mathbb{R}[x_1, \ldots, x_m]$.

We sometimes abbreviate sum-of-squares as SoS.

Definition 3.3.16 (Bounded polynomial)**.** We say a polynomial $p(x_1, \ldots, x_m) \in \mathbb{R}[x_1, \ldots, x_m]$ is (d, C)*-bounded* if

(a) p has degree at most d, and

(b) for each monomial of degree $d' \leqslant d$ in p, its coefficient has magnitude at most $C/(d'!)$.

We say p is a (k, d, C)*-bounded sum-of-squares polynomial* if p is a sum-of-squares polynomial, $p = q_1^2 + \cdots + q_k^2$, and each of the q_i's are (d, C)-bounded.

Claim 3.3.17 (Composition of bounded SoS polynomials)**.** Let $p_1(x_1, x_2)$ be a (k_1, d_1, C_1)-bounded SoS polynomial and $p_2(x_1, x_2)$ be a (k_2, d_2, C_2)-bounded SoS polynomial. Then

(a) $p_1 + p_2$ is a $(k_1 + k_2,\ \max(d_1, d_2),\ \max(C_1, C_2))$-bounded SoS polynomial;

(b) $p_1 p_2$ is a $(k_1 k_2,\ d_1 + d_2,\ (d_1 + d_2 + 1)2^{d_1+d_2} C_1 C_2)$-bounded SoS polynomial;

(c) For any $t \in [0, 1]$, $p_1((1-t)x_1 + ty_1, (1-t)x_2 + ty_2)$ is a (k_1, d_1, C_1)-bounded SoS polynomial in x_1, y_1, x_2, y_2.

Proof. For this proof, we let $[x_1^i x_2^j]p(x_1, x_2)$ denote the coefficient corresponding to $x_1^i x_2^j$ in p.

The first statement is obvious as we can simply combine the two representations as sums of squares. The second also follows immediately by taking the two representations as sums of squares and expanding the product. For this, we use that, when $r(x_1, x_2)$ is (d_1, C_1)-bounded and $s(x_1, x_2)$ is (d_2, C_2)-bounded, rs is $(d_1 + d_2, (d_1 + d_2 + 1)2^{d_1+d_2} C_1 C_2)$-bounded.

$$[x_1^i x_2^j](rs) = \sum_{\substack{0 \leqslant i' \leqslant i \\ 0 \leqslant j' \leqslant j}} |[x_1^{i'} x_2^{j'}]r||[x_1^{i-i'} x_2^{j-j'}]s|$$

$$\leqslant \sum_{\substack{0 \leqslant i' \leqslant i \\ 0 \leqslant j' \leqslant j}} \frac{C_1}{(i'+j')!} \frac{C_2}{(i+j-i'-j')!} = \frac{C_1 C_2}{(i+j)!} \sum_{\substack{0 \leqslant i' \leqslant i \\ 0 \leqslant j' \leqslant j}} \binom{i+j}{i'+j'} \leqslant \frac{C_1 C_2}{(i+j)!}(i+j+1)2^{i+j}$$

For the final statement, write $p_1(x_1, x_2) = q_1(x_1, x_2)^2 + \cdots + q_{k_1}(x_1, x_2)^2$. Now we simply substitute in the change of variables. A coefficient of the new polynomial is

$$\begin{aligned}
&[x_1^{i_1} y_1^{j_1} x_2^{i_2} y_2^{j_2}] q_\ell((1-t)x_1 + ty_1, (1-t)x_2 + ty_2) \\
&= [x_1^{i_1+j_1} x_2^{i_2+j_2}] q_\ell(x_1, x_2) \cdot \binom{i_1 + j_1}{i_1} (1-t)^{i_1} t^{j_1} \binom{i_2 + j_2}{i_2} (1-t)^{i_2} t^{j_2} \\
&\leqslant [x_1^{i_1+j_1} x_2^{i_2+j_2}] q_\ell(x_1, x_2)
\end{aligned}$$

and this shows that after the change of variables, $p_1((1-t)x_1 + ty_1, (1-t)x_2 + ty_2)$ is still a (k_1, d_1, C_1)-bounded SoS polynomial. □

The Sum-of-Squares Framework

We now provide an overview of the sum-of-squares proof system. We closely follow the exposition as it appears in the lecture notes of Barak [35].

Pseudo-Distributions. A discrete probability distribution over $\mathbb{R}^m$ is defined by its probability mass function, $D\colon \mathbb{R}^m \to \mathbb{R}$, which must satisfy $\sum_{x \in \mathrm{supp}(D)} D(x) = 1$ and $D \geqslant 0$. We extend this definition by relaxing the non-negativity constraint to merely requiring that D passes certain low-degree non-negativity tests. We call the resulting object a pseudo-distribution.

Definition 3.3.18 (Pseudo-distribution)**.** A *degree-ℓ pseudo-distribution* is a finitely-supported function $D : \mathbb{R}^m \to \mathbb{R}$ such that $\sum_x D(x) = 1$ and $\sum_x D(x)p(x)^2 \geqslant 0$ for every polynomial p of degree at most $\ell/2$, where the summation is over all x in the support of D.

Next, we define the related notion of pseudo-expectation.

Definition 3.3.19 (Pseudo-expectation)**.** The *pseudo-expectation* of a function f on $\mathbb{R}^m$ with respect to a pseudo-distribution μ, denoted by $\widetilde{\mathbb{E}}_{\mu(x)}[f(x)]$, is defined as

$$\widetilde{\mathbb{E}}_{\mu(x)}[f(x)] = \sum_x \mu(x) f(x).$$

We use the notation $\widetilde{\mathbb{E}}_{\mu(x)}\left[(1, x_1, x_2, \ldots, x_m)^{\otimes \ell}\right]$ to denote the degree-ℓ moment tensor of the pseudo-distribution μ. In particular, each entry in the moment tensor corresponds to the pseudo-expectation of a monomial of degree at most ℓ in x.

Definition 3.3.20 (Constrained pseudo-distributions)**.** Let

$$\mathcal{A} = \{\, p_1 \geqslant 0, p_2 \geqslant 0, \ldots, p_r \geqslant 0 \,\}$$

be a system of r polynomial inequality constraints of degree at most d in m variables. Let μ be a degree-ℓ pseudo-distribution over $\mathbb{R}^m$. We say that μ *satisfies* $\mathcal{A}$ at

degree $\ell \geqslant 1$ if for every subset $\mathcal{S} \subset [r]$ and every sum-of-squares polynomial q such that $\deg(q) + \sum_{i \in \mathcal{S}} \max(\deg(p_i), d) \leqslant \ell$, $\widetilde{\mathbb{E}}_\mu \left[q \prod_{i \in \mathcal{S}} p_i \right] \geqslant 0$. Further, we say that μ *approximately satisfies* the system of constraints $\mathcal{A}$ if the above inequalities are satisfied up to additive error $\widetilde{\mathbb{E}}_\mu \left[q \prod_{i \in \mathcal{S}} p_i \right] \geqslant -2^{-n^\ell} \|q\| \prod_{i \in \mathcal{S}} \|p_i\|$, where $\|\cdot\|$ denotes the Euclidean norm of the coefficients of the polynomial, represented in the monomial basis.

Crucially, there's an efficient separation oracle for moment tensors of constrained pseudo-distributions. Below gives the unconstrained statement; the constraint statement follows analogously.

Fact 3.3.21 ([192, 166, 177, 94])**.** *For any $m, \ell \in \mathbb{N}$, the following convex set has a $m^{\mathcal{O}(\ell)}$-time weak separation oracle, in the sense of [96]:*[¶]*:*

$$\left\{ \widetilde{\mathbb{E}}_{\mu(x)} \left[(1, x_1, x_2, \ldots, x_m)^{\otimes \ell} \right] \middle| \ \mu \text{ is a degree-}\ell \text{ pseudo-distribution over } \mathbb{R}^m \right\}$$

This fact, together with the equivalence of weak separation and optimization [96] forms the basis of the sum-of-squares algorithm, as it allows us to efficiently approximately optimize over pseudo-distributions.

Given a system of polynomial constraints, denoted by $\mathcal{A}$, we say that it is *explicitly bounded* if it contains a constraint of the form $\{\|x\|^2 \leqslant 1\}$. Then, the following fact follows from Fact 3.3.21 and [96]:

Theorem 3.3.22 (Efficient optimization over pseudo-distributions)**.** *There exists an $(m + r)^{O(\ell)}$-time algorithm that, given any explicitly bounded and satisfiable system $\mathcal{A}$ of r polynomial constraints in m variables, outputs a degree-ℓ pseudo-distribution that satisfies $\mathcal{A}$ approximately, in the sense of Definition 3.3.20.*[‖]

Remark 3.3.23 (Bit complexity and approximate satisfaction)**.** We will eventually

[¶]A separation oracle of a convex set $S \subset \mathbb{R}^M$ is an algorithm that can decide whether a vector $v \in \mathbb{R}^M$ is in the set, and if not, provide a hyperplane between v and S. Roughly, a weak separation oracle is a separation oracle that allows for some ε slack in this decision.

[‖]Here, we assume that the bit complexity of the constraints in $\mathcal{A}$ is $(m + t)^{O(1)}$.

apply this result to a constraint system that can be defined with numbers with $\log(t)$ bits, where t is the sample complexity of the algorithm (scaling polynomially with m). Consequently, we can run this algorithm efficiently, and the errors incurred here (which is exponentially small in $\mathfrak{n}$) can be thought of as a "machine precision" error, and is dominated by the sampling errors incurred elsewhere. We can therefore safely ignore precision issues in the rest of our proof.

The pseudo-distribution D found will satisfy $\mathcal{A}$ only approximately, but provided the bit complexity of the sum-of-squares proof of $\mathcal{A} \vdash_{r'} B$, i.e. the number of bits required to write down the proof, is bounded by $m^{\mathcal{O}(\ell)}$ (assuming that all numbers in the input have bit complexity $m^{\mathcal{O}(1)}$), we can compute to sufficiently good error in polynomial time that the soundness will hold approximately. All of our sum-of-squares proofs will have this bit complexity.

We now state some standard facts for pseudo-distributions, which extend facts that hold for standard probability distributions. These can be found in the prior works listed above.

Fact 3.3.24 (Cauchy–Schwarz for pseudo-distributions)**.** *Let f, g be polynomials of degree at most d in the variables $x \in \mathbb{R}^m$. Then, for any degree-d pseudo-distribution μ, $\widetilde{\mathbb{E}}_\mu [fg] \leqslant \sqrt{\widetilde{\mathbb{E}}_\mu [f^2]} \cdot \sqrt{\widetilde{\mathbb{E}}_\mu [g^2]}$.*

Fact 3.3.25 (Hölder's inequality for pseudo-distributions)**.** *Let f, g be polynomials of degree at most d in the variables $x \in \mathbb{R}^m$. Fix $t \in \mathbb{N}$. Then, for any degree-dt pseudo-distribution μ,*

$$\widetilde{\mathbb{E}}_\mu [f^{t-1} g] \leqslant \left(\widetilde{\mathbb{E}}_\mu [f^t]\right)^{\frac{t-1}{t}} \cdot \left(\widetilde{\mathbb{E}}_\mu [g^t]\right)^{\frac{1}{t}}.$$

In particular, when t is even, $\widetilde{\mathbb{E}}_\mu [f]^t \leqslant \widetilde{\mathbb{E}}_\mu [f^t]$.

Sum-of-squares proofs. Up to minor technical details, our algorithm is to set up a polynomial constraint system and then call Theorem 3.3.22 to get a pseudo-distribution

μ over the indeterminates $\{\lambda'_i \mid i \in [m]\}$ which approximately satisfies the constraints. With this pseudo-distribution, we will output $\widetilde{\mathbb{E}}_\mu [\lambda']$ as our estimated Hamiltonian coefficients. To show that these estimates are close to the true coefficients λ, we use that, under μ, the polynomial constraints hold. That is, for a constraint $p \geqslant 0$, we have $\widetilde{\mathbb{E}}_\mu [p] \geqslant 0$. If we can use these constraints to derive that $|\widetilde{\mathbb{E}}_\mu [\lambda'_a] - \lambda_a| \leqslant \varepsilon$ for every $a \in [m]$, then the algorithm is correct. So, we provide such a proof: this proof will be in the sum-of-squares proof system.

Let $f_1, f_2, \ldots, f_r$ and g be multivariate polynomials in the indeterminates $x \in \mathbb{R}^m$. Given the constraints $\{f_1 \geqslant 0, \ldots, f_r \geqslant 0\}$, a *sum-of-squares proof* of the identity $\{g \geqslant 0\}$ is a set of polynomials $\{p_S\}_{S \subseteq [r]}$ such that

$$g = \sum_{S \subseteq [r]} p_S^2 \cdot \prod_{i \in S} f_i.$$

As its name suggests, the existence of such an SoS proof shows that if the constraints $\{f_i \geqslant 0 \mid i \in [r]\}$ are satisfied, then the identity $g \geqslant 0$ is also satisfied. We say that this SoS proof has *degree* ℓ if for every set $S \subseteq [r]$, the polynomial $p_S^2 \Pi_{i \in S} f_i$ has degree at most ℓ. If there is a degree-ℓ SoS proof that $\{f_i \geqslant 0 \mid i \in [r]\}$ implies $\{g \geqslant 0\}$, we write

$$\{f_i \geqslant 0 \mid i \in [r]\} \left|\frac{x}{\ell}\right. \{g \geqslant 0\}. \tag{3.8}$$

We will sometimes drop the indeterminate in $\left|\frac{x}{\ell}\right.$ when this causes no confusion. For all polynomials $f, g \colon \mathbb{R}^m \to \mathbb{R}$ and for all coordinate-wise polynomials $F \colon \mathbb{R}^m \to \mathbb{R}^{m_F}$, $G \colon \mathbb{R}^m \to \mathbb{R}^{m_G}$, $H \colon \mathbb{R}^{m_H} \to \mathbb{R}^m$, we have the following inference rules.**

Sum-of-squares proofs allow us to deduce properties of pseudo-distributions that satisfy some constraints.

Fact 3.3.26 (Soundness)**.** *Let μ be a degree-ℓ pseudo-distribution. If μ is consistent with the set of degree-d_A polynomial constraints $\mathcal{A}$, denoted $\mu \left|\frac{=}{d_A}\right. \mathcal{A}$, and there is a*

**This notation should be read in the following way: given the proofs above the bar line, we can derive the proof below the bar line.

$$\text{Addition Rule}\quad \frac{\mathcal{A}\left|\frac{}{\ell}\right. \{f \geqslant 0, g \geqslant 0\}}{\mathcal{A}\left|\frac{}{\ell}\right. \{f+g \geqslant 0\}} \qquad \text{Multiplication Rule}\quad \frac{\mathcal{A}\left|\frac{}{\ell}\right. \{f \geqslant 0\}, \quad \mathcal{A}\left|\frac{}{\ell'}\right. \{g \geqslant 0\}}{\mathcal{A}\left|\frac{}{\ell+\ell'}\right. \{f \cdot g \geqslant 0\}}$$

$$\text{Transitivity Rule}\quad \frac{\mathcal{A}\left|\frac{}{\ell}\right. \mathcal{B}, \quad \mathcal{B}\left|\frac{}{\ell'}\right. C}{\mathcal{A}\left|\frac{}{\ell\cdot\ell'}\right. C} \qquad \text{Substitution Rule}\quad \frac{\{F \geqslant 0\}\left|\frac{}{\ell}\right. \{G \geqslant 0\}}{\{F(H) \geqslant 0\}\left|\frac{}{\ell\cdot\deg(H)}\right. \{G(H) \geqslant 0\}}$$

degree-d_B sum-of-squares proof that $\mathcal{A}\left|\frac{}{d_B}\right. \mathcal{B}$, and $\ell \geqslant d_A d_B$, then $\mu\left|\frac{}{d_A d_B}\right. \mathcal{B}$.

We also have a converse to Fact 3.3.26: every property of low-level pseudo-distributions can be derived by low-degree sum-of-squares proofs.

Fact 3.3.27 (Completeness)**.** *Let $d \geqslant r \geqslant r'$. Suppose $\mathcal{A}$ is a collection of polynomial constraints with degree at most r, and $\mathcal{A}\left|\frac{x}{}\right. \{\sum_{i=1}^m x_i^2 \leqslant 1\}$. Let $\{g \geqslant 0\}$ be a polynomial constraint. If every degree-d pseudo-distribution that satisfies $D\left|\frac{}{r}\right. \mathcal{A}$ also satisfies $D\left|\frac{}{r'}\right. \{g \geqslant 0\}$, then for every $\varepsilon > 0$, there is a sum-of-squares proof $\mathcal{A}\left|\frac{}{d}\right. \{g \geqslant -\varepsilon\}$.*

Basic sum-of-squares proofs. Now, we recall some basic facts about sum-of-squares proofs. First, any univariate polynomial inequality admits a sum-of-squares proof over the reals.

Fact 3.3.28 (Univariate polynomial inequalities admit SoS proofs [148])**.** *Let p be a polynomial of degree d. If $p(x) \geqslant 0$ for all $x \geqslant 0$, we have $\left|\frac{x}{d}\right. \{\, p(x) \geqslant 0 \,\}$. If $p(x) \geqslant 0$ for all $x \in [a, b]$, then $\{\, x \geqslant a, x \leqslant b \,\}\left|\frac{x}{d}\right. \{\, p(x) \geqslant 0 \,\}$.*

Second, if $p \geqslant 0$ and p is a quadratic, then this admits a sum-of-squares proof.

Fact 3.3.29 (Quadratic polynomial inequalities admit SoS proofs)**.** *Let p be a polynomial in the indeterminates $x \in \mathbb{R}^m$ such that p has degree 2 and $p \geqslant 0$ for all $x \in \mathbb{R}^m$. Then $\left|\frac{x}{2}\right. \{\, p(x) \geqslant 0 \,\}$.*

Proof. Let M be the unique $(m+1) \times (m+1)$ Hermitian matrix such that, for

$v(x) = (1, x_1, \ldots, x_m)^\dagger$,

$$p(x_1, \ldots, x_m) = v(x)^\dagger M v(x)\,.$$

The inequality $p \geqslant 0$ implies that M is PSD: consider a vector $v = (v_1, \ldots, v_{m+1}) \in \mathbb{R}^{m+1}$. If $v_1 \neq 0$, then $v^\dagger M v = p(w) \geqslant 0$ for $w_1 = v_2/v_1, \ldots, w_m = v_{m+1}/v_1$. If $v_1 = 0$, then $v^\dagger M v = \lim_{c\to\infty} p(c \cdot w) \geqslant 0$ for $w_1 = v_2, \ldots, w_m = v_{m+1}$. This shows that M must be PSD, so we can write $M = \sum_{i=1}^{m+1} u_i u_i^\dagger$ for some vectors $u_i \in \mathbb{R}^{m+1}$. Thus,

$$p(x_1, \ldots, x_m) = v(x)^\dagger M v(x) = \sum_{i=1}^{m+1} \langle u_i, v(x)\rangle^2$$

which is a degree-2 SoS polynomial and we are done. □

We also use the following basic sum-of-squares proofs. For further details, we refer the reader to a recent monograph [87].

Fact 3.3.30 (Operator norm bound)**.** *For a symmetric matrix $A \in \mathbb{R}^{d\times d}$ and a vector $v\mathbb{R}^d$,*

$$\left|\frac{v}{2}\right. \Big[v^\dagger A v \leqslant \|A\|\|v\|^2\Big].$$

Fact 3.3.31 (Almost triangle inequality)**.** *Let $f_1, f_2, \ldots, f_r$ be indeterminates. Then*

$$\left|\frac{f_1, f_2, \ldots, f_r}{2t}\right. \left\{ \Big(\sum_{i\leqslant r} f_i\Big)^{2t} \leqslant r^{2t-1}\Big(\sum_{i=1}^{r} f_i^{2t}\Big) \right\}.$$

Fact 3.3.32 (SoS Hölder's inequality)**.** *Let $w_1, \ldots w_n$ be indeterminates and let $f_1, \ldots f_n$ be polynomials of degree d in the variables $x \in \mathbb{R}^m$. Let k be a power of 2. Then*

$$\{\, w_i^2 = w_i, \forall i \in [n] \,\} \left|\frac{x, w}{2kd}\right. \left\{ \Big(\frac{1}{n}\sum_{i=1}^{n} w_i f_i\Big)^k \leqslant \Big(\frac{1}{n}\sum_{i=1}^{n} w_i\Big)^{k-1}\Big(\frac{1}{n}\sum_{i=1}^{n} f_i^k\Big) \right\}.$$

Fact 3.3.33 (Almost square-root)**.** *Given a scalar indeterminate v, $\{\, v^2 \leqslant 1 \,\} \left|\frac{v}{2}\right.$*

$\{ -1 \leqslant v \leqslant 1 \}$.

Proof. We know that $\{ (1-v)^2 = 1 + v^2 - 2v \geqslant 0 \}$ and $\{ (1+v)^2 = 1 + v^2 + 2v \geqslant 0 \}$. By assumption, $\{ 1 - v^2 \geqslant 0 \}$. So, from the addition rule we have $\{ 2 + 2v \geqslant 0 \}$ and $\{ 2 - 2v \geqslant 0 \}$. Rearranging yields the claim. □

3.4 Translating Between Polynomials and Nested Commutators

In this section, we relate nested commutators of matrices to polynomials of their associated eigenvalues. We begin with the following basic observation mentioned in the proof overview.

Lemma 3.4.1 (Nested commutators to eigen-polynomials)**.** *For matrices $A, B \in \mathbb{C}^{n \times n}$, in the basis where A is diagonal with entries $A_{ii} = \alpha_i$, $AB = B \circ \{\alpha_i\}_{ij}$ and $BA = B \circ \{\alpha_j\}_{ij}$. Consequently,*

$$[A, B]_k = B \circ \{(\alpha_i - \alpha_j)^k\}_{ij}.$$

Further, by linearity, for a polynomial $q(x) = \sum_{k=0}^{d} c_k x^k$,

$$\sum c_k [A, B]_k = B \circ \{q(\alpha_i - \alpha_j)\}_{ij}$$

In light of the above, we make the following definition associating a polynomial to an expression involving commutators of matrices.

Definition 3.4.2 (Univariate "commutator polynomials")**.** For a polynomial $p(x) = a_0 + a_1 x + \cdots + a_d x^d$, given square matrices X, A of the same size, define

$$p(X|A) = a_0 A + a_1 [X, A]_1 + \cdots + a_d [X, A]_d \,.$$

We will need a generalization of the above that associates bivariate polynomials $p(x, y)$ to expressions with nested commutators involving two matrices X, Y. We will primarily be interested in the case where X, Y commute or are close to commuting.

We begin by generalizing the nested commutator.

Definition 3.4.3 (Bivariate nested commutators)**.** Let $S \in \{0,1\}^\ell$ and $X, Y, A \in \mathbb{C}^{N \times N}$ be matrices. Consider a sequence $Z_1, Z_2, \ldots, Z_\ell$ of length ℓ where each $Z_i \in \{X, Y\}$ and $Z_i = X$ if and only if the ith entry of S is 0. We define

$$[(X, Y)_S, A] = [Z_1, [Z_2, [\ldots [Z_\ell, A] \ldots]]] .$$

For a monomial $x^i y^j$, we wish to associate it to a nested commutator $[(X, Y)_S, A]$ where the number of 0's and 1's in S is i and j, respectively. There are many different such commutators, reflecting that X and Y need not commute. We will show that when X, Y are close to commuting, the nested commutators are also close, so the ordering in S does not matter. When $|S| = 2$, this follows from the identity below.

Fact 3.4.4 (Jacobi identity)**.** *We have the identity* $[X, [Y, A]] - [Y, [X, A]] = [[X, Y], A]$.

We extend this to higher-order commutators. Below, and throughout the remainder of this chapter, for a sequence $S \in \{0,1\}^*$, we will use $\mathrm{len}(S)$ to denote its length.

Lemma 3.4.5 (Reordering bivariate nested commutators)**.** *For any two sequences $S, S' \in \{0,1\}^\ell$ with the same number of 0's and 1's, let $t \leqslant \ell^2$ be the number of adjacent swaps needed to transform S to S'. Then there are some coefficients $c_1, \ldots, c_t \in \{-1, 1\}$, and sequences $S_1, T_1, \ldots, S_t, T_t$ where $\mathrm{len}(S_i) + \mathrm{len}(|T_i|) = \ell - 2$ such that*

$$[(X, Y)_S, A] - [(X, Y)_{S'}, A] = \sum_{i=1}^{t} c_i \left[(X, Y)_{S_i}, [[X, Y], [(X, Y)_{T_i}, A]]\right] .$$

Proof. Consider when S, S' differ exactly by a single swap of two adjacent elements.

In this case, by Fact 3.4.4, the difference on the LHS is equal to exactly one term of the form

$$[(X,Y)_{S_i}, [[X,Y], [(X,Y)_{T_i}, A]]]$$

where S_i is the prefix up to the point where S, S' differ and T_i is the suffix. Now we can repeatedly apply this to swap adjacent elements of S until it matches S'. Each of the residual terms is of the form given on the RHS so we are done. □

Now we define the correspondence between bivariate polynomials and bivariate nested commutators by arbitrarily choosing an order for the S associated to every monomial.

Definition 3.4.6 (Bivariate "commutator polynomials")**.** Given a polynomial of degree d in two variables x, y, $p(x,y) = \sum_{i+j \leqslant d} a_{ij} x^i y^j$, we define its associated matrix commutator polynomial for matrices $X, Y, A \in \mathbb{C}^{n \times n}$ as

$$p(X, Y|A) = \sum_{i+j \leqslant d} a_{ij} [X, [Y, A]_j]_i$$

The key property that we will need is that certain polynomial identities (in the original bivariate polynomials) are essentially preserved under this translation. We begin by showing this for monomials. The following fact shows a relation between a commutator polynomial on A to a commutator polynomial on B.

Fact 3.4.7. *For any Hermitian matrix X and matrices A, B, ρ, we have the identity*

$$\operatorname{tr}\big([X,A]B^\dagger\rho\big) - \operatorname{tr}\big(A[X,B]^\dagger\rho\big) = -\operatorname{tr}\big(AB^\dagger[X,\rho]\big)\,.$$

Proof.

$$\begin{aligned}
\operatorname{tr}\big([X,A]B^\dagger\rho\big) - \operatorname{tr}\big(A[X,B]^\dagger\rho\big) &= \operatorname{tr}\big(XAB^\dagger\rho - AXB^\dagger\rho\big) - \operatorname{tr}\big(AB^\dagger X\rho - AXB^\dagger\rho\big) \\
&= \operatorname{tr}\big(XAB^\dagger\rho\big) - \operatorname{tr}\big(AB^\dagger X\rho\big) \\
&= \operatorname{tr}\big(AB^\dagger\rho X\big) - \operatorname{tr}\big(AB^\dagger X\rho\big)
\end{aligned}$$

$$= -\operatorname{tr}\big(AB^\dagger[X,\rho]\big) \qquad \square$$

This can be extended to general monomials.

Lemma 3.4.8 (Commutator monomial equivalences)**.** *Let $p(x,y) = x^{i_1}y^{i_2}$, $q(x,y) = x^{j_1}y^{j_2}$, and $r(x,y) = p(x,y)q(x,y)$. Let $d = \deg(r)$. Then for some $\ell \leqslant d^2$, we can write*

$$\operatorname{tr}\Big(p(X,Y|A)q(X,Y|B)^\dagger\rho\Big) - \operatorname{tr}\Big(A\cdot r(X,Y|B)^\dagger\rho\Big) = \sum_{i=1}^{\ell} Z_i,$$

where every Z_i is one of the following three types of errors:

1. $\pm\operatorname{tr}\Big(\big[(X,Y)_S, A\big]\big[(X,Y)_T, B\big]^\dagger[X,\rho]\Big)$, *where* $\operatorname{len}(S)+\operatorname{len}(T) = d-1$;
2. $\pm\operatorname{tr}\Big(\big[(X,Y)_S, A\big]\big[(X,Y)_T, B\big]^\dagger[Y,\rho]\Big)$, *where* $\operatorname{len}(S)+\operatorname{len}(T) = d-1$;
3. $\pm\operatorname{tr}\Big(A\big[(X,Y)_S, [[X,Y],[(X,Y)_T,B]]\big]^\dagger\rho\Big)$, *where* $\operatorname{len}(S)+\operatorname{len}(T) = d-2$.

Proof. Our goal is to express

$$\operatorname{tr}\Big([X,[Y,A]_{i_2}]_{i_1}[X,[Y,B]_{j_2}]^\dagger_{j_1}\rho\Big) - \operatorname{tr}\Big(A[X,[Y,B]_{i_2+j_2}]^\dagger_{i_1+j_1}\rho\Big)$$

as a sum of errors. Observe that Fact 3.4.7 allows us to remove one copy of X or Y from the commutator in front of A and move it onto the commutator in front of B at the cost of an error term of type 1 or 2. Thus, we can repeatedly apply Fact 3.4.7 to move all of the X's and Y's in the commutator in front of A onto the commutator in front of B and write

$$\operatorname{tr}\Big([X,[Y,A]_{i_2}]_{i_1}[X,[Y,B]_{j_2}]^\dagger_{j_1}\rho\Big) - \operatorname{tr}\Big(A[Y,[X,[Y,B]_{j_2}]_{j_1+i_1}]^\dagger_{i_2}\rho\Big)$$

as a sum of i_1 terms of type 1 and i_2 terms of type 2. Next, we can apply Lemma 3.4.5 to reorder the sequence of X and Y on the commutator in front of B at the cost of

error terms of type 3. This allows us to write

$$\mathrm{tr}\Big(A[Y,[X,[Y,B]_{j_2}]_{j_1+i_1}]^\dagger_{i_2}\rho\Big) - \mathrm{tr}\Big(A[X,[Y,B]_{i_2+j_2}]^\dagger_{i_1+j_2}\rho\Big)$$

as a sum of $i_2(j_1+i_1)$ terms of type 3. Together, this gives the desired bound. $\square$

Theorem 3.4.9 (Polynomial identities to nested commutator identities)**.** *Consider a formal polynomial identity in two variables*

$$p_1(x,y)q_1(x,y) + \cdots + p_k(x,y)q_k(x,y) = 0$$

where each of the polynomials p_i, q_i is (d,C)-bounded. Let $X, Y \in \mathbb{C}^{N\times N}$ be Hermitian matrices and $\rho, A, B \in \mathbb{C}^{N\times N}$ be arbitrary matrices. Then we can write

$$\mathrm{tr}\left(\left(p_1(X,Y|A)q_1(X,Y|B)^\dagger + \cdots + p_k(X,Y|A)q_k(X,Y|B)^\dagger\right)\rho\right) = \sum_{i=1}^{t} c_i Z_i,$$

where $t \leqslant 4kd^6$, the coefficients c_i satisfy $|c_i| \leqslant C^2 2^{2d}$, and every Z_i is one of the following three types of errors:

1. $\frac{\pm 1}{(\mathrm{len}(S)+\mathrm{len}(T))!}\,\mathrm{tr}\Big(\left[(X,Y)_S, A\right]\left[(X,Y)_T, B\right]^\dagger[X,\rho]\Big)$, *where* $\mathrm{len}(S)+\mathrm{len}(T) \leqslant 2d$;
2. $\frac{\pm 1}{(\mathrm{len}(S)+\mathrm{len}(T))!}\,\mathrm{tr}\Big(\left[(X,Y)_S, A\right]\left[(X,Y)_T, B\right]^\dagger[Y,\rho]\Big)$, *where* $\mathrm{len}(S)+\mathrm{len}(T) \leqslant 2d$;
3. $\frac{\pm 1}{(\mathrm{len}(S)+\mathrm{len}(T))!}\,\mathrm{tr}\Big(A\left[(X,Y)_S, [[X,Y],[(X,Y)_T,B]]\right]^\dagger\rho\Big)$, *where* $\mathrm{len}(S)+\mathrm{len}(T) \leqslant 2d$.

Proof. Let $r_\ell(x,y) = p_\ell(x,y)q_\ell(x,y)$. By the assumed formal polynomial identity,

$$\mathrm{tr}\left(\left(Ar_1(X,Y|B)^\dagger + \cdots + Ar_k(X,Y|B)^\dagger\right)\rho\right) = 0.$$

So, it suffices to express each product

$$\mathrm{tr}\left(p_\ell(X,Y|A)q_\ell(X,Y|B)^\dagger\rho\right) - \mathrm{tr}\left(Ar_\ell(X,Y|B)^\dagger\rho\right)$$

as a linear combination of errors. Let $p_\ell = \sum_{i_1,i_2} a_{\ell,i_1,i_2} x^{i_1} y^{i_2}$ and $q_\ell = \sum_{j_1,j_2} b_{\ell,j_1,j_2} x^{j_1} y^{j_2}$. We can expand the above expression into its commutator monomials:

$$\begin{aligned}
&\operatorname{tr}\left(p_\ell(X,Y|A) q_\ell(X,Y|B)^\dagger \rho\right) - \operatorname{tr}\left(A r_\ell(X,Y|B)^\dagger \rho\right) \\
&= \sum_{i_1,i_2,j_1,j_2} a_{\ell,i_1,i_2} \cdot b_{\ell,j_1,j_2} \Big(\operatorname{tr}([X,[Y,A]_{i_2}]_{i_1} \cdot [X,[Y,B]_{j_2}]^\dagger_{j_1} \rho) - \operatorname{tr}(A \cdot [X,[Y,B]_{i_2+j_2}]^\dagger_{i_1+j_1} \rho) \Big) .
\end{aligned}$$

Lemma 3.4.8 shows how to expand summand into error terms; note that the degree of the product is $i_1 + j_1 + i_2 + j_2 \leqslant 2d$. There are at most $(2d)^2$ error terms per summand, and there are d^4 summands per product $p_\ell q_\ell$. This gives a total bound on the number of error terms as $4kd^6$ as desired. All that is left is to bound the size of the coefficients. For an error term of any type associated to a particular summand, the coefficient is $a_{\ell,i_1,i_2} b_{\ell,j_1,j_2}$. Let S, T be the sequences associated to the error term. Because p_ℓ and q_ℓ are (d, C)-bounded, we can bound

$$|a_{\ell,i_1,i_2} b_{\ell,j_1,j_2}| \leqslant \frac{C^2}{(i_1+i_2)!(j_1+j_2)!} \leqslant \frac{C^2 2^{2d}}{(i_1+i_2+j_1+j_2)!} \leqslant \frac{C^2 2^{2d}}{(\operatorname{len}(S)+\operatorname{len}(T))!}.$$

The last inequality holds because, in all error types, $\operatorname{len}(S) + \operatorname{len}(T)$ is at most the degree of the product. □

3.5 Polynomial Approximations of the Exponential

Our goal for this section is to construct a polynomial approximation to e^x that satisfies the properties necessary for our algorithm. In particular, we need a polynomial that approximates e^x on an interval $[-\kappa, \kappa]$, but with the additional property that the polynomial does not grow too quickly outside the area of approximation.

Definition 3.5.1 (Flat approximation of the exponential)**.** Given $\varepsilon, \eta \in (0,1)$ and $\kappa \geqslant 1$, we say a polynomial $p(x)$ is a $(\kappa, \eta, \varepsilon)$-*flat exponential approximation* if

1. $|p(x) - e^x| \leqslant \varepsilon$ for $x \in [-\kappa, \kappa]$, and
2. $|p(x)| \leqslant \max(1, e^x) e^{\eta|x|}$.

The canonical polynomial approximation of e^x is its Taylor series truncation.

Definition 3.5.2. Let $s_\ell(x) = \sum_{k=0}^{\ell} \frac{x^k}{k!}$ denote the degree-ℓ truncation of the Taylor series of e^x.

Remark 3.5.3. The Taylor series truncation $s_\ell(x)$ does not satisfy our desired approximation properties from Definition 3.5.1, even for large ℓ. This is because the truncation blows up too quickly on the negative tail: $|s_\ell(-\ell)| \approx e^\ell > e^{\eta\ell}$. The same issue occurs with conventional approximations of the exponential, like truncations of the Chebyshev series and QSVT-style "bounded" approximations of the Chebyshev series [92, 200]: these truncations at degree ℓ fail in the region around $-\ell$ (so, notably, increasing the degree does not improve the flatness parameter of these approximations).

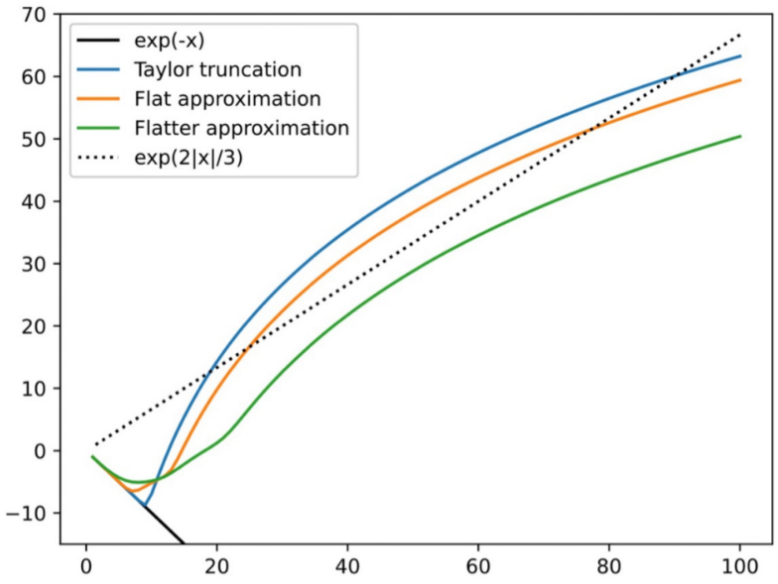

Figure 3-1: Plots (on log log scale) of the Taylor series truncation and our flat polynomial approximations to the exponential e^{-x}. Observe that the Taylor series fails the flat approximation property because it exceeds the reference line $e^{2x/3}$.

While standard techniques for approximating the exponential do not suffice, we construct a modified polynomial that is a flat approximation to the exponential.

Definition 3.5.4 (Iteratively truncated Taylor series)**.** For positive integers k, ℓ, we

define

$$p_{k,\ell}(x) = s_{2\ell}(x/k) \cdot s_{4\ell}(x/k) \cdots s_{2^k\ell}(x/k)$$

to be the product of k geometrically increasing Taylor series truncations. Next, we define a shifted integral of $p_{k,\ell}(x)$ as follows:

$$q_{k,\ell}(x) = 1 + \int_0^x p_{k,\ell}(y)\,\mathrm{d}y\,.$$

The degree of $p_{k,\ell}(x)$ and $q_{k,\ell}$ is $(2^{k+1}-1)\ell$ and $(2^{k+1}-1)\ell + 1$, respectively.

It will also be important that $p_{k,\ell}$ is always positive and thus $q_{k,\ell}$ is monotonically increasing. The main theorems that we show in this section are as follows:

Theorem 3.5.5 (Flat approximations to the exponential)**.** *Given $\varepsilon, \eta \in (0,1)$ and $\kappa \geqslant 1$, let $k \geqslant 5/\eta$ and $\ell \geqslant 100(\kappa + \log k/\varepsilon)$. Then the polynomials $p_{k,\ell}, q_{k,\ell}$ are $(\kappa, \eta, \varepsilon)$-flat exponential approximation.*

Theorem 3.5.6 (Monotonicity of the flat approximations)**.** *For any positive integers k, ℓ, we have the following:*

$$\vdash_{\frac{x,y}{2^{k+1}\ell+20}} \left\{\, 0.5(x-y)(1+0.25(x-y)^2)(q_{k,\ell}(x) - q_{k,\ell}(y)) - 0.00025(x-y)^2 p_{k,\ell}(x) \geqslant 0 \,\right\}.$$

Moreover, the LHS is a $(10^{2^k\ell},\, 2^k\ell + 10,\, 200^{2^k\ell})$-bounded sum-of-squares polynomial in x and y.

Remark 3.5.7. Note that Theorem 3.5.6 stems from the intuition that roughly we should have

$$q_{k,\ell}(x) - q_{k,\ell}(y) = \int_y^x p_{k,\ell}(y)\,\mathrm{d}y \geqslant (x-y)(p_{k,\ell}(x) - p_{k,\ell}(y))\,.$$

However, the above isn't quite true, so the additional terms in Theorem 3.5.6 are there to account for this. The fact that the difference is not only non-negative, but also a sum of square polynomials will be crucial later on in the analysis of our algorithm.

We will prove Theorem 3.5.5 in this section. The proof of Theorem 3.5.6 is deferred to the appendix as it is quite long and computational. We begin by establishing some basic facts about the truncated Taylor series and our polynomials p and q. The following fact follows from Taylor's theorem, which implies that for every $x \in \mathbb{R}$ there exists some $c \in [0,1]$ such that

$$e^x - s_\ell(x) = e^{cx} \frac{x^{\ell+1}}{(\ell+1)!}\,. \tag{3.9}$$

Fact 3.5.8 (Bounds on truncated Taylor series of e^x)**.** *For $x \geqslant 0$, $e^x \geqslant s_\ell(x)$ for all ℓ. For $x < 0$, we have $s_{2\ell-1}(x) \leqslant e^x \leqslant s_{2\ell}(x)$ for all ℓ.*

Corollary 3.5.9 (Even truncations are non-negative)**.** *For any $\ell \in \mathbb{N}$, for all $x \in \mathbb{R}$, $s_{2\ell}(x) \geqslant 0$.*

Lemma 3.5.10 (Taylor series truncation of e^x)**.** *Given $\varepsilon \in (0,1)$ and $b \geqslant 0$, let $\ell \geqslant 10b + \log(1/\varepsilon)$. Then $|s_\ell(x) - e^x| \leqslant \varepsilon$ for $x \in [-b, b]$.*

Proof. This statement also follows from Taylor's theorem (3.9). From this, for all x, the approximation error of the Taylor series truncation is

$$|s_{\ell-1}(x) - e^x| \leqslant \frac{e^{|x|}|x|^\ell}{\ell!} \leqslant e^{|x|}\Big(\frac{e|x|}{\ell}\Big)^\ell.$$

Taking $\ell \geqslant e^2|x| + \log(e^{|x|}/\varepsilon)$ makes the above expression bounded by ε.[††] The final bound comes from taking $|x| \leqslant b$. □

Claim 3.5.11. The polynomial $p_{k,\ell}$ is $((2^{k+1}-1)\ell, 1)$-bounded and the polynomial $q_{k,\ell}$ is $((2^{k+1}-1)\ell + 1, 1)$-bounded.

Proof. Note that the coefficients of the monomials in $p_{k,\ell}$ are all positive and are all

[††]This can be improved by a log log factor [92, Lemma 59], but this does not improve the degree for the parameter setting in our application.

less than the coefficients of the monomials in

$$\left(1 + \frac{(x/k)}{1!} + \frac{(x/k)^2}{2!} + \dots\right)^k$$

where we replace each truncated sum with the full infinite sum. However the above is exactly equal to $1 + x + \frac{x^2}{2!} + \dots$. Thus for any degree d, the coefficient of x^d in $p_{k,\ell}(x)$ is at most $\frac{1}{d!}$ so $p_{k,\ell}$ is $((2^{k+1}-1)\ell, 1)$-bounded since its degree is $(2^{k+1}-1)\ell$. Recall that $q_{k,\ell}$ is obtained by integrating p and thus we immediately get the same bound on the coefficients of $q_{k,\ell}$ i.e. it is is $((2^{k+1}-1)\ell + 1, 1)$-bounded and we are done. □

Now we move onto the proof of Theorem 3.5.5.

Proof of Theorem 3.5.5. We prove the statement for $p_{k,\ell}$ first. By Lemma 3.5.10 (with $\varepsilon \leftarrow \varepsilon/(k2^{2\kappa})$)), we have that for all $x \in [-\kappa, \kappa]$

$$\left|\frac{s_{2^j\ell}(x/k)}{e^{x/k}} - 1\right| \leqslant \frac{\varepsilon}{k \cdot e^{2\kappa}}$$

and thus,

$$1 - \frac{\varepsilon}{e^{2\kappa}} \leqslant \frac{p_{k,\ell}(x)}{e^x} - 1 \leqslant 1 + \frac{\varepsilon}{e^{2\kappa}}$$

so we have $|p_{k,\ell}(x) - e^x| \leqslant \varepsilon/\kappa$ on the interval $[-\kappa, \kappa]$. Now we immediately get the same guarantee for $q_{k,\ell}(x)$ by integrating the above.

Now we prove the second condition for $p_{k,\ell}$. For $x \geqslant 0$, we simply incur the e^x upper bound by Fact 3.5.8. For $x \leqslant 0$, let j_0 be the smallest positive integer such that $-x < 2^{j_0+1}k\ell$ then for all $1 \leqslant j < j_0$,

$$s_{2^j\ell}(x/k) \leqslant \frac{(x/k)^{2^j\ell}}{(2^j\ell)!}\,.$$

Also, for all $j > j_0 + 2$,

$$s_{2^j\ell}(x/k) = s_{(2^j\ell)-1}(x/k) + \frac{(x/k)^{2^j\ell}}{(2^j\ell)!} \leqslant e^{x/k} + \left(\frac{3x}{2^j\ell k}\right)^{2^j\ell} \leqslant 1 \tag{3.10}$$

where we used Fact 3.5.8. Finally, for $j_0 \leqslant j \leqslant j_0 + 2$ we have

$$s_{2^j\ell}(x/k) = s_{(2^j\ell)-1}(x/k) + \frac{(x/k)^{2^j\ell}}{(2^j\ell)!} \leqslant 1 + \frac{(x/k)^{2^j\ell}}{(2^j\ell)!} \leqslant 2\max\left(\frac{(x/k)^{2^j\ell}}{(2^j\ell)!}, 1\right) \quad (3.11)$$

Overall, combining the above, we conclude

$$p_{k,\ell}(x) = s_{2\ell}(x/k)\cdots s_{2^k\ell}(x/k) \leqslant 8\prod_{j=1}^{k}\max\left(\frac{(x/k)^{2^j\ell}}{(2^j\ell)!}, 1\right) \quad (3.12)$$

where the factor of 8 is because we need to apply (3.11) on at most 3 terms. Note that there must be some $k_0 \in \{\, j_0, j_0+1, j_0+2 \,\}$, such that

$$\frac{(x/k)^{2^j\ell}}{(2^j\ell)!} > 1,$$

exactly when $j \leqslant k_0$. Then, we can upper bound Equation (3.12) as follows:

$$p_{k,\ell}(x) \leqslant \frac{8(x/k)^{(2^{k_0+1}-2)\ell}}{(2\ell)!\cdots(2^{k_0}\ell)!} \quad (3.13)$$

Further, we have that

$$(2\ell)!\cdots(2^{k_0}\ell)! \geqslant \frac{((2^{k_0+1}-2)\ell)!}{4^{(2^{k_0+1}-2)\ell}}.$$

Substituting in Equation (3.13), we get

$$p_{k,\ell}(x) \leqslant \frac{(5x/k)^{(2^{k_0+1}-2)\ell}}{((2^{k_0+1}-2)\ell)!} \leqslant e^{5|x|/k}\,. \quad (3.14)$$

The desired bound for q then follows as well from the definition as q is obtained by simply integrating p. □

3.6 Accessing Gibbs States: The Quantum Piece

We now describe how our algorithm uses its copies of ρ. This is the only location in the algorithm where we access ρ; the rest of the algorithm is entirely classical. By making measurements on copies of ρ, we can estimate expectations of observables $\mathrm{tr}(X\rho)$ for various choices of the matrix X. These estimates will then be used in the learning algorithm.

Lemma 3.6.1. *Let $X = UDU^\dagger \in \mathbb{C}^{N\times N}$ be a unitarily diagonalizable matrix, and suppose we are given copies of a quantum state with density matrix $\rho \in \mathbb{C}^{N\times N}$. Then we can estimate $\mathrm{tr}(X\rho)$ to $\varepsilon\|X\|$ error with probability $\geqslant 1-\delta$ using $\mathcal{O}(\frac{1}{\varepsilon^2}\log\frac{1}{\delta})$ copies of ρ. The running time is the number of samples times the cost of applying $U^\dagger$ to a quantum state and measuring in the computational basis.*

Proof. Consider taking ρ and applying $U^\dagger$ to form $U^\dagger\rho U$, then measuring in the computational basis. We see the outcome i with probability $\langle u_i|\,\rho\,|u_i\rangle$, where u_i is the ith column of U. Let z be the random variable attained by performing this measurement and taking $z = D_{i,i}$; then $\mathbb{E}[z] = \sum_i D_{i,i}\langle u_i|\,\rho\,|u_i\rangle = \mathrm{tr}(X\rho)$, and z is always bounded by $\max_i|D_{i,i}| = \|X\|$. By a Chernoff bound we conclude that averaging $\mathcal{O}(\frac{1}{\varepsilon^2}\log\frac{1}{\delta})$ independent copies of this estimator gives the desired error bound. □

We wish to estimate $\mathrm{tr}(X\rho)$ for many different Pauli observables X, so we could simply run Lemma 3.6.1 for each one. However, we can use a classical shadows-like procedure [120] to estimate them all at once.

Lemma 3.6.2 (Computing expectations of observables of Gibbs states)**.** *Let $H = \sum_{a=1}^m \lambda_a E_a$ be a $\mathfrak{K}$-local Hamiltonian on $\mathfrak{n}$ qubits whose dual interaction graph $\mathfrak{G}$ has maximum degree $\mathfrak{d}$. Consider the set of ℓ-$\mathfrak{G}$-local Paulis, $\mathcal{P}_\ell$. There is a quantum algorithm that, with probability $\geqslant 1-\delta$, outputs estimates of $\mathrm{tr}(P_1P_2P_3\rho)$ to ε error for all $P_1, P_2, P_3 \in \mathcal{P}_\ell$. This algorithm uses $t = \mathcal{O}(\frac{4^{3\ell}}{\varepsilon^2}\log\frac{|\mathcal{P}_\ell|^3}{\delta})$ samples, $\mathcal{O}(\mathfrak{n}t)$ quantum gates, and $\mathrm{poly}(t, \mathfrak{n}, |P_\ell|)$ classical post-processing time.*

Proof. Consider the following procedure, where we take as input a copy of ρ, rotate into the eigenbasis of a Pauli matrix $P \in \mathcal{P}$ chosen uniformly at random, and then measure in the computational basis to get a measurement outcome $b \in \{0,1\}^{\mathfrak{n}}$. Take $t = \mathcal{O}(\frac{4^{3\ell}}{\varepsilon^2} \log \frac{|\mathcal{P}_\ell|^3}{\delta})$ of these samples (P, b) and denote the collection $\mathcal{S}$.

Now, consider estimating some $\mathrm{tr}(X\rho)$ where $X = UDU^\dagger \in \mathcal{P}$ and $\mathrm{supp}(X) \leqslant 3\ell$; $P_1P_2P_3$ is such an X, up to a root of unity. Since

$$\mathrm{tr}(X\rho) = \mathrm{tr}_{\mathrm{supp}(X)}(X_{(\mathrm{supp}(X))}\, \mathrm{tr}_{[n]-\mathrm{supp}(X)}(\rho)),$$

it suffices to consider ρ on the support of X. Let $\mathcal{T} \subset \mathcal{S}$ be the collection of samples whose Pauli matrix agrees with X on the support of X. Then we can use this subsample to run Lemma 3.6.1 and estimate $\mathrm{tr}(X\rho)$, as all that is needed to use it are measurements in a basis where X is diagonal. Since the probability that any given sample is in a basis where X is diagonal is $\geqslant 4^{-3\ell}$, with t samples we can guarantee that we get an estimate $\mathrm{tr}(X\rho)$ to ε error with failure probability $\delta/|\mathcal{P}_\ell|^3$. By union bound, we can then conclude that with our set of samples we can get ε-good estimates of every $\mathrm{tr}(P_1P_2P_3\rho)$ with probability $\geqslant 1-\delta$.

The running time can be seen to be $\mathrm{poly}(t, \mathfrak{n}, |\mathcal{P}_\ell|)$. Note that, for any product of Paulis $P \in \mathcal{P}$, we can apply the quantum gate to rotate into its eigenbasis with $\mathcal{O}(\mathfrak{n})$ gates, so the quantum gate complexity is $\mathcal{O}(\mathfrak{n}t) = \mathcal{O}(\frac{4^{3\ell}\mathfrak{n}}{\varepsilon^2} \log \frac{|\mathcal{P}_\ell|^3}{\delta})$. □

3.7 Algorithm and Analysis

Now we are ready to present our learning algorithm. We will show:

Theorem 3.7.1 (Efficiently learning a quantum Hamiltonian)**.** *Let $H = \sum_{a=1}^m \lambda_a E_a \in \mathbb{C}^{N\times N}$ be a $(\mathfrak{d}, \mathfrak{K})$-low intersection Hamiltonian on $\mathfrak{n}$ qubits (as in Problem 1). Suppose we are given the terms $\{E_a\}_{a\in[m]}$, $\varepsilon > 0$, $\delta > 0$, and copies of the Gibbs state ρ at a known inverse temperature $\beta > 0$. Then there exists an algorithm that can output estimates $\tilde{\lambda}_a$ such that, with probability $\geqslant 1-\delta$, $(\hat{\lambda}_a - \lambda_a)^2 \leqslant \varepsilon^2$ for all $a \in [m]$. This*

algorithm uses

$$\mathcal{O}\Big((m^6/\varepsilon^{e^{f(\mathfrak{K},\mathfrak{d})\beta}})\log(m/\delta) + (f(\mathfrak{K},\mathfrak{d})/(\beta^2\varepsilon^2))\log(m/\delta)\Big)$$

copies of the Gibbs state and running time

$$\mathrm{poly}(m, \log(1/\delta)) \cdot (1/\varepsilon)^{e^{f(\mathfrak{K},\mathfrak{d})\beta}} + f(\mathfrak{K},\mathfrak{d}) \cdot (m/(\beta^2\varepsilon^2))\log(m/\delta),$$

where $f(\mathfrak{K},\mathfrak{d})$ is a positive function depending only on $\mathfrak{K}$ and $\mathfrak{d}$.

Specifically, we prove that, when $\beta > g(\mathfrak{K},\mathfrak{d})$ for some positive g, there exists an algorithm that succeeds with sample complexity $(m^6/\varepsilon^{e^{f(\mathfrak{K},\mathfrak{d})\beta}})\log(m/\delta)$ and time complexity $\mathrm{poly}(m, \log(1/\delta)) \cdot (1/\varepsilon)^{e^{f(\mathfrak{K},\mathfrak{d})\beta}}$ for some positive $f(\mathfrak{K},\mathfrak{d})$. We get Theorem 3.7.1 by taking g to be the threshold at which the high-temperature algorithm [103] stops working and then combining our complexity bounds with that of the high-temperature algorithm. That β is lower bounded by a constant is just a simplification; the same algorithm and analysis should work for any temperature upon appropriate adjustments.

We are concerned with the setting where $\mathfrak{K}$ and $\mathfrak{d}$ are constant (that is, when the Hamiltonian is low-intersection), so we do not try to optimize dependence on these parameters. Further, we assume that $\mathfrak{n} = \mathcal{O}(m)$; this is without loss of generality, as $m \geqslant \mathfrak{n}/\mathfrak{K}$ is necessary for the support of the Hamiltonian to be all $\mathfrak{n}$ qubits.

Our main theorem follows from showing that a low-degree sum-of-squares relaxation of a carefully chosen polynomial system can be rounded to obtain accurate estimates of the true coefficients $\{\lambda_a\}_{a\in[m]}$. We first set up the system, and then build up to the full algorithm (Algorithm 3.7.2).

Presenting the polynomial system. We begin by setting up the precise polynomial system that we will then solve. Let $\varepsilon_0 = \frac{\varepsilon^{10^{C_{\mathfrak{K},\mathfrak{d}}\beta}}}{m^3}$ for some sufficiently large constant $C_{\mathfrak{K},\mathfrak{d}}$ depending only on $\mathfrak{K}$ and $\mathfrak{d}$. Next, let $\ell_0 = 2^{C_{\mathfrak{K},\mathfrak{d}}\beta}\log(1/\varepsilon), \ell_1 = 4\mathfrak{K}$ and define $\mathcal{A} = \mathcal{P}_{4^{C_{\mathfrak{K},\mathfrak{d}}\beta}\ell_0}$ and $\mathcal{B} = \mathcal{P}_{\ell_1}$ (as in Definition 3.3.3).

Let $\lambda' = [\lambda'_1, \lambda'_2, \ldots, \lambda'_m]$ be a set of indeterminates. We will solve a polynomial system in these indeterminates, and then use it to extract estimates to the true coefficients. In the system, and throughout, we denote $H' = \sum_a \lambda'_a E_a$. This object H' is merely notational convenience: we do not optimize with this exponential sized object. We only ever need to work with traces of H' and thus our polynomial system admits a succinct (polynomial sized) representation.

We begin by making measurements of ρ to construct ε_0-accurate estimates to $\operatorname{tr}(A_1 A_2 A_3 \rho)$ for all $A_1, A_2, A_3 \in \mathcal{A}$ using Lemma 3.6.2. Let $\widetilde{\operatorname{tr}}$ denote the map from $A_1 A_2 A_3 \rho$ to our estimate of $\operatorname{tr}(A_1 A_2 A_3 \rho)$, extended by linearity to linear combinations of such matrices that we have estimated.[‡‡] Again, this is notational convenience for our polynomial system.

By Corollary 3.3.14, $|\mathcal{A}| \leqslant m(10\mathfrak{d})^{4^{C_{\mathfrak{K},\mathfrak{d}}\beta}\ell_0} \leqslant m(1/\varepsilon)^{10^{C_{\mathfrak{K},\mathfrak{d}}\beta}}$ so we can produce these estimates in running time $\operatorname{poly}_{\beta,\mathfrak{K},\mathfrak{d}}(m, 1/\varepsilon)$ and sample complexity

$$\mathcal{O}\left(\frac{4^{3\cdot 4^{C_{\mathfrak{K},\mathfrak{d}}\beta}\ell_0}}{\varepsilon_0^2} \log \frac{|\mathcal{P}_{4^{C_{\mathfrak{K},\mathfrak{d}}\beta}\ell_0}|}{\delta}\right) = \mathcal{O}\left(\frac{m^6}{\varepsilon^{20^{C_{\mathfrak{K},\mathfrak{d}}\beta}}} \log \frac{m}{\delta}\right). \tag{3.15}$$

This is the sample complexity of the entire algorithm, as this is all of the information we need of ρ. Then, we write the following polynomial system:

$$\mathcal{C}_{\lambda'} = \left\{ \begin{aligned} &\forall a \in [m] && -1 \leqslant \lambda'_a \leqslant 1 \\ & && H' = \sum_{a\in[m]} \lambda'_a \cdot E_a \\ &\forall A \in \mathcal{A} && |\widetilde{\operatorname{tr}}(A(H'\rho - \rho H'))|^2 \leqslant \varepsilon_0^2 \\ &\forall B_1, B_2 \in \mathcal{B}, && \left|\widetilde{\operatorname{tr}}\big(B_2 q_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}(-\beta H'|B_1)\rho\big) - \widetilde{\operatorname{tr}}(B_1 B_2 \rho)\right|^2 \leqslant \varepsilon^2 \end{aligned} \right\}, \tag{3.16}$$

where q is the polynomial from Definition 3.5.4 extended to commutators using Defi-

[‡‡]To make this well-defined, we need to ensure that we only estimate a set of $A_1 A_2 A_3$ that are linearly independent. This simply involves removing duplicates up to phase.

nition 3.4.6. With this set of parameters, by Theorem 3.5.5,

$$\deg(q_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}) \leqslant 2 \cdot 4^{C_{\mathfrak{K},\mathfrak{d}}\beta} \log(1/\varepsilon) + 1 \text{ and} \tag{3.17}$$

$$q_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0} \text{ is a } \left(0.001 \cdot 2^{C_{\mathfrak{K},\mathfrak{d}}\beta} \log(1/\varepsilon), \frac{5}{C_{\mathfrak{K},\mathfrak{d}}\beta}, 0.001\varepsilon\right)\text{-flat exponential approximation.} \tag{3.18}$$

Representing the polynomial system. It is important to understand how the above can be represented as a polynomial system in the λ' with real coefficients and polynomial (in m) size. The key is that we can write the expressions inside the traces in the form $(\sum_\alpha (\lambda')^\alpha M_\alpha)\,\rho$ where α ranges over low-degree monomials and M_α are rescaled Pauli matrices, which we can explicitly and succinctly represent. Then by linearity of $\widetilde{\mathrm{tr}}$, we can write the system by plugging in our estimates for $\widetilde{\mathrm{tr}}(M_\alpha\rho)$. For example, for the first constraint, we can write

$$\widetilde{\mathrm{tr}}(A_1A_2(H'\rho - \rho H')) = \widetilde{\mathrm{tr}}((A_1A_2H' - H'A_1A_2)\rho) = \sum_a \lambda'_a(\widetilde{\mathrm{tr}}(A_1A_2E_a\rho) - \widetilde{\mathrm{tr}}(E_aA_1A_2\rho))\,.$$

We can then plug in our estimates $\widetilde{\mathrm{tr}}(A_1A_2E_a\rho)$ and $\widetilde{\mathrm{tr}}(E_aA_1A_2\rho)$ obtained from Lemma 3.6.2. These trace estimates may be complex-valued, so that $\widetilde{\mathrm{tr}}(A_1A_2E_a\rho) - \widetilde{\mathrm{tr}}(E_aA_1A_2\rho) = \psi_a + \mathrm{i}\chi_a$ for some $\psi_a, \chi_a \in \mathbb{R}$. We can rewrite this,

$$\begin{aligned} &\left|\sum_a \lambda'_a(\chi_a + \mathrm{i}\psi_a)\right|^2 \leqslant \varepsilon_0^2 \\ \iff &\left(\sum_a \lambda'_a\chi_a\right)^2 + \left(\sum_a \lambda'_a\psi_a\right)^2 \leqslant \varepsilon_0^2, \end{aligned}$$

revealing that we have a polynomial constraint in λ' with real coefficients. For the second constraint, $q_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}(-\beta H'|B_1)$ is a linear combination of nested commutators $[E_{a_\ell}, [E_{a_{\ell-1}}, [\ldots [E_{a_1}, B_1]]]]$ where $\ell \leqslant 2 \cdot 4^{C_{\mathfrak{K},\mathfrak{d}}\beta} \log(1/\varepsilon) + 1$, so by Lemma 3.3.4, such a commutator is an element of $\mathcal{A}$. The coefficient associated with this nested commutator

is $(-\beta)^\ell \lambda'_{a_\ell} \dots \lambda'_{a_1}$, so that for some polynomials $p_A(\lambda')$ we can write

$$\widetilde{\mathrm{tr}}\big(B_2 q_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}(-\beta H'|B_1)\rho\big) = \sum_{A\in\mathcal{A}} p_A(\lambda')\widetilde{\mathrm{tr}}\left(B_2 A\rho\right).$$

Since we can plug in our estimates with $\widetilde{\mathrm{tr}}$ as before, this expression is well-defined. We can further conclude that the associated constraint can be written as a polynomial constraint in λ' with real coefficients. Recall that $|\mathcal{A}| \leqslant m(1/\varepsilon)^{10^{C_{\mathfrak{K},\mathfrak{d}}\beta}}$. Furthermore, by Lemma 3.3.13, there are at most $m(10\mathfrak{d})^{4^{C_{\mathfrak{K},\mathfrak{d}}\beta}\ell_0} \leqslant m(1/\varepsilon)^{10^{C_{\mathfrak{K},\mathfrak{d}}\beta}}$ distinct monomials in the λ' that appear in all of the p_A and thus this entire representation has size $\mathrm{poly}(m, 1/\varepsilon)$.

Algorithm 3.7.2 (Learning a Hamiltonian from Gibbs states, Theorem 3.7.1)**.**

Input: Description of $\mathfrak{K}$-local Hamiltonian terms $\{E_a\}_{a\in[m]}$ with dual interaction graph degree $\mathfrak{d}$; accuracy and failure probability parameters $\varepsilon, \delta \in (0,1)$; inverse temperature $\beta > 0$; $(m^6/\varepsilon^{e^{f(\mathfrak{K},\mathfrak{d})\beta}})\log(m/\delta)$ copies of the Gibbs state $\rho = \frac{\exp(-\beta H)}{\mathrm{tr}\exp(-\beta H)}$ for an unknown Hamiltonian $H = \sum_a \lambda_a E_a$.

Operation:

1. Set $\varepsilon_0 = \frac{\varepsilon^{10^{\gamma\beta}}}{m^3}$ where γ is a sufficiently large constant depending only on $\mathfrak{K}, \mathfrak{d}$;
2. Set $\ell_0 = 2^{\gamma\beta}\log(1/\varepsilon), \ell_1 = 4\mathfrak{K}$;
3. Define $\mathcal{A} = \mathcal{P}_{4^{\gamma\beta}\ell_0}, \mathcal{B} = \mathcal{P}_{\ell_1}$;
4. For all $A \in \mathcal{A}$, compute estimates $\widetilde{\mathrm{tr}}(A\rho)$ of $\mathrm{tr}(A\rho)$ to ε_0 error using Lemma 3.6.2;

5. Consider the following constraint system:

$$\mathcal{C}_{\lambda'} = \begin{Bmatrix} \forall a \in [m] & -1 \leqslant \lambda'_a \leqslant 1 \\ & H' = \sum_{a\in[m]} \lambda'_a \cdot E_a \\ \forall A \in \mathcal{A} & |\widetilde{\mathrm{tr}}(A(H'\rho - \rho H'))|^2 \leqslant \varepsilon_0^2 \\ \forall B_1, B_2 \in \mathcal{B}, & \left|\widetilde{\mathrm{tr}}\big(B_2 q_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}(-\beta H'|B_1)\rho\big) - \widetilde{\mathrm{tr}}(B_1 B_2 \rho)\right|^2 \leqslant \varepsilon^2 \end{Bmatrix},$$

6. Compute a degree-$\mathcal{O}(2^{\gamma\beta}\ell_0)$ pseudo-distribution μ consistent with $\mathcal{C}_{\lambda'}$;

7. Set $\hat{\lambda} = \widetilde{\mathbb{E}}_\mu[\lambda']$;

Output: $\hat{\lambda}$ such that with probability at least $1-\delta$, $|\hat{\lambda}_a - \lambda_a|^2 \leqslant \varepsilon^2$ for all $a \in [m]$.

Analyzing the polynomial system. We have now described the bulk of the algorithm (Algorithm 3.7.2). To obtain Theorem 3.7.1, we prove the following intermediate lemmas. In light of Lemma 3.6.2, we may assume that all of our estimates are accurate to within ε_0 with high probability.

Assumption 3.7.3 (Measuring Gibbs states)**.** For all $A_1, A_2, A_3 \in \mathcal{A}$, our estimate satisfies

$$|\widetilde{\mathrm{tr}}(A_1A_2A_3\rho) - \mathrm{tr}(A_1A_2A_3\rho)| \leqslant \varepsilon_0\,.$$

The first lemma states that when our estimates are accurate, the true coefficients i.e. $\lambda' = \lambda$ are a feasible solution to the system.

Lemma 3.7.4 (Feasibility of the constraint system)**.** *If Assumption 3.7.3 holds, then the constraints in the system $\mathcal{C}_{\lambda'}$ in Equation* (3.16) *are satisfied when $\lambda' = \lambda$ (so that $H' = H$).*

Next, we need to prove soundness. The proof of soundness is broken into two key

steps. First, we prove that any solution to the system must have $H' = \sum_a \lambda'_a E_a$ nearly commute with H.

Remark 3.7.5 (On sum-of-squares proof degree)**.** All of our sum-of-squares proofs will be of degree $\log(1/\varepsilon)^{2^{f(\mathfrak{K},\mathfrak{d})\beta}}$. For the sake of brevity, we omit the precise degrees from all subsequent statements.

Lemma 3.7.6 (SoS finds an *almost-commuting* state)**.** *Assume that Assumption 3.7.3 holds. Let* $H' = \sum_a \lambda'_a E_a$ *and write* $\mathrm{i}[H, H'] = \sum \gamma_b F_b$ *for some* $2\mathfrak{K}$*-*$\mathfrak{G}$*-local terms* F_b *where each of* γ_b *are linear expressions in the* λ'_a *(such a representation exists by Lemma 3.3.4). Then*

$$\{ \mathcal{C}_{\lambda'} \} \Big|\!\!\frac{\lambda'}{} \{ -e^{\mathcal{O}_{\mathfrak{K},\mathfrak{d}}(\beta)} m^{1.5} \sqrt{\varepsilon_0} \leqslant \gamma_b \leqslant e^{\mathcal{O}_{\mathfrak{K},\mathfrak{d}}(\beta)} m^{1.5} \sqrt{\varepsilon_0} \} .$$

Using the above lemma, we can derive identifiability of each parameter in the sum-of-squares proof system:

Lemma 3.7.7 (A sum-of-squares proof of identifiability)**.** *Suppose Assumption 3.7.3 holds. Given the constraint system* $\mathcal{C}_{\lambda'}$*, for any term* $a^* \in [m]$*,*

$$\{ \mathcal{C}_{\lambda'} \} \Big|\!\!\frac{\lambda'}{} \{ (\lambda_{a^*} - \lambda'_{a^*})^2 \leqslant 2^{\gamma\beta} \varepsilon \} .$$

Combining the above lemmas, we can almost prove Theorem 3.7.1. We first show how to immediately get the same learning guarantee with the same sample complexity but a slightly worse running time. Later on, we show in Section 3.11 how to improve this runtime and complete the proof of the stronger Theorem 3.7.1.

Theorem 3.7.8 (Weaker version of Theorem 3.7.1)**.** *Let* $H = \sum_{a=1}^m \lambda_a E_a \in \mathbb{C}^{N \times N}$ *be a* $\mathfrak{K}$*-local Hamiltonian on* $\mathfrak{n}$ *qubits whose dual interaction graph has degree* $\mathfrak{d}$ *(as in Problem 1). Suppose we are given the terms* $\{E_a\}_{a \in [m]}$*,* $\varepsilon > 0$*,* $\delta > 0$*, and copies of the Gibbs state* ρ *at a known inverse temperature* $\beta > (10(\mathfrak{d}+1))^{-10}$*. Then there exists an algorithm that can output estimates* $\tilde{\lambda}_a$ *such that, with probability* $\geqslant 1 - \delta$*,*

$(\hat{\lambda}_a - \lambda_a)^2 \leqslant \varepsilon^2$ *for all* $a \in [m]$. *This algorithm uses*

$$\mathcal{O}\Big((m^6/\varepsilon^{e^{f(\mathfrak{K},\mathfrak{d})\beta}})\log(m/\delta)\Big)$$

copies of the Gibbs state and running time

$$(m/\varepsilon)^{\log(1/\varepsilon)\cdot\exp(f(\mathfrak{K},\mathfrak{d})\cdot\beta)}.$$

where $f(\mathfrak{K},\mathfrak{d})$ *is a positive function depending only on* $\mathfrak{K}$ *and* $\mathfrak{d}$.

Proof of Theorem 3.7.8. Given $\mathcal{O}((m^6/\varepsilon^{e^{f(\mathfrak{K},\mathfrak{d})\beta}})\log(m/\delta))$ copies of the Gibbs state ρ, it follows from Lemma 3.6.2 that with probability at least $1-\delta$, Assumption 3.7.3 holds. Conditioning on this event, it follows from Lemma 3.7.4 that the constraint system $\mathcal{C}_{\lambda'}$ is feasible. Given a degree $t = \Omega\big(2^{\gamma\beta}\log(1/\varepsilon)\big) = \Omega\big(2^{\gamma\beta}\log(1/\varepsilon)\big)$ pseudo-distribution μ, where γ is a fixed universal constant that only depends on $\mathfrak{K},\mathfrak{d}$, it follows from Lemma 3.7.7 (and redefining ε appropriately as $(\varepsilon/2^{\gamma\beta})^2$) that for all $a \in [m]$,

$$\Big(\lambda_{a^*} - \widetilde{\mathbb{E}}_\mu\left[\lambda'_{a^*}\right]\Big)^2 \leqslant \widetilde{\mathbb{E}}_\mu\left[(\lambda_{a^*} - \lambda'_{a^*})^2\right] \leqslant \varepsilon^2.$$

The running time is dominated by computing a degree-t pseudo-distribution over m indeterminates and $|\mathcal{A}| + |\mathcal{B}| \leqslant m^{\mathcal{O}(1)}(1/\varepsilon)^{\exp(\mathcal{O}(\gamma\beta))}$ many constraints. It follows from Theorem 3.3.22 that the overall running time is at most

$$(m/\varepsilon)^{\log(1/\varepsilon)\exp(\mathcal{O}(\gamma\beta))},$$

which concludes the proof. □

3.8 A Proof of Feasibility (Lemma 3.7.4)

Our goal in this section is to prove Lemma 3.7.4, which states that our constraint system Eq. (3.16) is satisfied by the ground truth, $\lambda' = \lambda$, when Assumption 3.7.3 is

satisfied. Since we are working with the *ground truth* Hamiltonian H, we can relax, as our proofs need not be in the sum-of-squares system.

We first use the assumption to bound the error of replacing $\widetilde{\mathrm{tr}}$ with tr in the constraints of the sum-of-squares system. This constraint error comes from the sampling error of our estimates of observables in Section 3.6.

Lemma 3.8.1. *When Assumption 3.7.3 holds, for all $B_1, B_2 \in \mathcal{B}$,*

$$|\widetilde{\mathrm{tr}}(B_1 B_2 \rho) - \mathrm{tr}(B_1 B_2 \rho)| \leqslant \tfrac{1}{2}\varepsilon$$
$$|\widetilde{\mathrm{tr}}(B_2 q_{\gamma\beta,\ell_0}(-\beta H|B_1)\rho) - \mathrm{tr}(B_2 q_{\gamma\beta,\ell_0}(-\beta H|B_1)\rho)| \leqslant \tfrac{1}{2}\varepsilon$$

Proof. Assumption 3.7.3 implies that

$$|\widetilde{\mathrm{tr}}(B_1 B_2 \rho) - \mathrm{tr}(B_1 B_2 \rho)| \leqslant 0.5\varepsilon$$

for all B_1, B_2. Next, by Lemma 3.3.12 and Claim 3.5.11, we can write $q_{\gamma\beta,\ell_0}(-\beta H|B_1)$ as a linear combination of the form

$$\sum_{A\in\mathcal{A}} c_A A$$

where $\mathcal{A} \subset \mathcal{P}_{2^{\gamma\beta+1}\ell_0 \mathfrak{K}}$ and the coefficients satisfy $\sum_{A\in\mathcal{A}} |c_A| \leqslant ((1+\beta)\mathfrak{d})^{2^{\gamma\beta+5}\ell_0}$. Thus, conditioned on the event that Assumption 3.7.3 holds, we have

$$|\widetilde{\mathrm{tr}}(B_2 q_{\gamma\beta,\ell_0}(-\beta H|B_1)\rho) - \mathrm{tr}(B_2 q_{\gamma\beta,\ell_0}(-\beta H|B_1)\rho)| \leqslant 0.5\varepsilon\,. \qquad \square$$

We now prove that the first two constraints are satisfied, when $\widetilde{\mathrm{tr}}$ is replaced by tr.

Lemma 3.8.2 (Feasibility of commutators)**.** *Let A be a matrix and let ρ be a Gibbs state associated with the $\mathfrak{K}$-local Hamiltonian $H = \sum_{a\in[m]} \lambda_a E_a$. Then for all $a \in [m]$, $|\lambda_a| \leqslant 1$ and $\mathrm{tr}(A(H\rho - \rho H)) = 0$.*

Proof. By assumption, $|\lambda_a| \leqslant 1$. Next, since H commutes with ρ, for all A,

$$\operatorname{tr}(A(H\rho - \rho H)) = 0. \qquad \square$$

It remains to prove that H satisfies the final set of constraints in Eq. (3.16), which states that $\operatorname{tr}(B_2 q(-\beta H|B_1)\rho) \approx \operatorname{tr}(B_1 B_2 \rho)$ when q is a good flat approximation to the exponential e^x (recall Definition 3.5.1). To provide intuition for this statement, consider working in the basis where H is diagonal and its diagonal entries (and eigenvalues) are σ_i. Then

$$\begin{aligned}
\operatorname{tr}((B_2 q(-\beta H|B_1) - B_1 B_2)\rho) &= \operatorname{tr}((B_2(B_1 \circ \{q(-\beta(\sigma_i - \sigma_j))\}_{ij}) - B_1 B_2)\rho) \\
&\approx \operatorname{tr}((B_2(B_1 \circ \{e^{-\beta(\sigma_i - \sigma_j)}\}_{ij}) - B_1 B_2)\rho) \\
&= \operatorname{tr}((B_2 \rho B_1 \rho^{-1} - B_1 B_2)\rho) \\
&= \operatorname{tr}(B_2 \rho B_1 - B_1 B_2 \rho) \\
&= 0,
\end{aligned}$$

where the second equality uses that q is a good approximation to e^x. In the following lemma, we make the above intuition precise, showing that when q is a sufficiently good flat exponential approximation, this derivation indeed holds up to some small error.

Lemma 3.8.3 (Feasibility of the polynomial approximation)**.** *Let H be a $\mathfrak{K}$-local Hamiltonian with interaction degree $\mathfrak{d}$ and let ρ be a Gibbs state of H at inverse temperature β. For $A, B \in \mathcal{P}_{4\mathfrak{K}}$ and a polynomial q that is a $(\kappa, \eta, \varepsilon)$-flat exponential approximation where*

(a) *$\kappa \geqslant C_1 \beta \log(1/\varepsilon)$ for some sufficiently large constant C_1 depending only on $\mathfrak{K}, \mathfrak{d}$,*

(b) *$\eta < 1/(C_2 \beta)$ for some sufficiently large constant C_2 depending only on $\mathfrak{K}, \mathfrak{d}$,*

we have

$$|\operatorname{tr}(Bq(-\beta H|A)\rho) - \operatorname{tr}(AB\rho)| \leqslant \tfrac{1}{2}\varepsilon \|A\| \|B\|.$$

Proof. By a change of basis, we take H to be diagonal with $H_{ii} = \sigma_i$. Consequently, ρ is also diagonal with $\rho_{ii} = e^{-\beta\sigma_i} / \operatorname{tr} e^{-\beta H}$, and

$$\begin{aligned}
\operatorname{tr}(Bq(-\beta H|A)\rho) - \operatorname{tr}(AB\rho) &= \operatorname{tr}(B \cdot q(-\beta H|A) \cdot \rho) - \operatorname{tr}(B\rho A) \\
&= \frac{1}{\operatorname{tr} e^{-\beta H}} \sum_{i,j=1}^{N} \Big(B_{ji}[q(-\beta H|A)]_{ij} e^{-\beta\sigma_j} - B_{ji}A_{ij}e^{-\beta\sigma_i} \Big) \\
&= \frac{1}{\operatorname{tr} e^{-\beta H}} \sum_{i,j=1}^{N} \Big(B_{ji}A_{ij}q(\beta(\sigma_j - \sigma_i))e^{-\beta\sigma_j} - B_{ji}A_{ij}e^{-\beta\sigma_i} \Big) \\
&= \frac{1}{\operatorname{tr} e^{-\beta H}} \sum_{i,j=1}^{N} B_{ji}A_{ij} \Big(q(\beta(\sigma_j - \sigma_i)) - e^{\beta(\sigma_j - \sigma_i)} \Big) e^{-\beta\sigma_j},
\end{aligned} \tag{3.19}$$

where the first equality follows from cyclicity of trace, the second uses its definition, the third uses Definition 3.4.2 and Lemma 3.4.1, and the fourth groups relevant terms together. Now, we use that q is a $(\kappa, \eta, \varepsilon)$-flat exponential approximation to bound this quantity. Let $L = \kappa/(3\beta)$ and let $S_a = \{i \in [N] : \sigma_i \in [aL, (a+1)L)\}$, so that $[N] = \bigsqcup_a S_a$ (where the index goes from $-\infty$ to ∞.) Continuing from Eq. (3.19),

$$\begin{aligned}
&\frac{1}{\operatorname{tr} e^{-\beta H}} \sum_{i,j=1}^{N} B_{ji}A_{ij} \Big(q(\beta(\sigma_j - \sigma_i)) - e^{\beta(\sigma_j - \sigma_i)} \Big) e^{-\beta\sigma_j} \\
&= \frac{1}{\operatorname{tr} e^{-\beta H}} \sum_{a} \sum_{b} \sum_{i \in S_a} \sum_{j \in S_b} B_{ji}A_{ij} \Big(q(\beta(\sigma_j - \sigma_i)) - e^{\beta(\sigma_j - \sigma_i)} \Big) e^{-\beta\sigma_j} \\
&= \sum_{\alpha} \frac{1}{\operatorname{tr} e^{-\beta H}} \sum_{a} \sum_{i \in S_a} \sum_{j \in S_{a+\alpha}} B_{ji}A_{ij} \Big(q(\beta(\sigma_j - \sigma_i)) - e^{\beta(\sigma_j - \sigma_i)} \Big) e^{-\beta\sigma_j}
\end{aligned} \tag{3.20}$$

Now, we bound (3.20) by breaking into cases depending on the value of α. First, suppose $-2 \leqslant \alpha \leqslant 2$. For $i \in S_a$ and $j \in S_{a+\alpha}$, the input to q is bounded by κ: $\beta|\sigma_j - \sigma_i| < \beta(\alpha + 1)L \leqslant \kappa$. So it follows from the approximation guarantee (Definition 3.5.1) that

$$|q(\beta(\sigma_j - \sigma_i)) - e^{\beta(\sigma_j - \sigma_i)}| \leqslant \varepsilon, \tag{3.21}$$

and thus

$$\begin{aligned}
&\Big|\sum_{i\in S_a}\sum_{j\in S_{a+\alpha}} B_{ji}A_{ij}\Big(q(\beta(\sigma_j-\sigma_i))-e^{\beta(\sigma_j-\sigma_i)}\Big)e^{-\beta\sigma_j}\Big| \\
&\leqslant \varepsilon\sum_{i\in S_a}\sum_{j\in S_{a+\alpha}}|B_{ji}A_{ij}|e^{-\beta\sigma_j} \\
&\leqslant \varepsilon\Big(\max_{j\in S_{a+\alpha}}\sum_{i\in S_a}|B_{ji}A_{ij}|\Big)\sum_{j\in S_{a+\alpha}}e^{-\beta\sigma_j} \\
&\leqslant \varepsilon\Big(\max_{j\in S_{a+\alpha}}\sum_{i\in S_a}|B_{ji}|^2\sum_{i\in S_a}|A_{ij}|^2\Big)^{1/2}\sum_{j\in S_{a+\alpha}}e^{-\beta\sigma_j} \\
&\leqslant \varepsilon\|A_{S_a,S_{a+\alpha}}\|\|B_{S_a,S_{a+\alpha}}\|\sum_{j\in S_{a+\alpha}}e^{-\beta\sigma_j}, \qquad (3.22)
\end{aligned}$$

where A_{S_a,S_b} and B_{S_a,S_b} are the submatrices of A and B indexed by $i\in S_a$ and $j\in S_b$ (in the eigenbasis of H). Next, suppose $\alpha\geqslant 3$. For $i\in S_a$ and $j\in S_{a+\alpha}$, $\sigma_j>\sigma_i$, so by the guarantee on q (Definition 3.5.1), $|q(\beta(\sigma_j-\sigma_i))|\leqslant\max(1,e^{\beta(\sigma_j-\sigma_i)})e^{\eta|\beta(\sigma_j-\sigma_i)|}=e^{(1+\eta)\beta(\sigma_j-\sigma_i)}$. Thus,

$$\begin{aligned}
&\Big|\sum_{i\in S_a}\sum_{j\in S_{a+\alpha}} B_{ji}A_{ij}\Big(q(\beta(\sigma_j-\sigma_i))-e^{\beta(\sigma_j-\sigma_i)}\Big)e^{-\beta\sigma_j}\Big| \\
&\leqslant \sum_{i\in S_a}\sum_{j\in S_{a+\alpha}}|B_{ji}A_{ij}|\Big(e^{(1+\eta)\beta(\sigma_j-\sigma_i)}+e^{\beta(\sigma_j-\sigma_i)}\Big)e^{-\beta\sigma_j} \\
&= \sum_{i\in S_a}\sum_{j\in S_{a+\alpha}}|B_{ji}A_{ij}|\Big(e^{\eta\beta(\sigma_j-\sigma_i)}+1\Big)e^{-\beta\sigma_i} \\
&\leqslant (e^{\eta\beta(\alpha+1)L}+1)\Big(\max_{i\in S_a}\sum_{j\in S_{a+\alpha}}|B_{ji}A_{ij}|\Big)\sum_{i\in S_a}e^{-\beta\sigma_i} \\
&\leqslant (e^{\eta\beta(\alpha+1)L}+1)\|A_{S_a,S_{a+\alpha}}\|\|B_{S_a,S_{a+\alpha}}\|\sum_{i\in S_a}e^{-\beta\sigma_i}\,. \qquad (3.23)
\end{aligned}$$

Finally, suppose $\alpha\leqslant -3$. For $i\in S_a$ and $j\in S_{a+\alpha}$, we have $\sigma_j<\sigma_i$, so by the same guarantee on q (Definition 3.5.1), $|q(\beta(\sigma_j-\sigma_i))|\leqslant e^{\eta\beta(\sigma_i-\sigma_j)}$. Thus,

$$\begin{aligned}
&\Big|\sum_{i\in S_a}\sum_{j\in S_{a+\alpha}} B_{ji}A_{ij}\Big(q(\beta(\sigma_j-\sigma_i))-e^{\beta(\sigma_j-\sigma_i)}\Big)e^{-\beta\sigma_j}\Big| \\
&\leqslant \sum_{i\in S_a}\sum_{j\in S_{a+\alpha}}|B_{ji}A_{ij}|(e^{-\beta\sigma_j}e^{\eta\beta(|\alpha|+1)L}+e^{-\beta\sigma_i}) \\
&\leqslant \|B_{S_a,S_{a+\alpha}}\|\|A_{S_a,S_{a+\alpha}}\|\Big(\sum_{i\in S_a}e^{-\beta\sigma_i}+e^{\eta\beta(|\alpha|+1)L}\sum_{j\in S_{a+\alpha}}e^{-\beta\sigma_j}\Big)\,. \qquad (3.24)
\end{aligned}$$

By Corollary 3.3.10, for an absolute constant $c > 1$ depending only on $\mathfrak{K}, \mathfrak{d}$,

$$\begin{aligned} \|A_{S_a,S_{a+\alpha}}\| &\leqslant \|A\| e^{c-(|\alpha|-1)L/c} \\ \|B_{S_a,S_{a+\alpha}}\| &\leqslant \|B\| e^{c-(|\alpha|-1)L/c} . \end{aligned} \tag{3.25}$$

We choose $\eta = 1/(C_2\beta)$ small enough to ensure that $\eta\beta(|\alpha|+1) \leqslant (|\alpha|-1)/c$ whenever $|\alpha| \geqslant 3$. Combining Eqs. (3.22) to (3.25) with the above, we get the following bound on Eq. (3.20).

$$\begin{aligned}
&\left| \sum_{\alpha} \frac{1}{\operatorname{tr} e^{-\beta H}} \sum_{a} \sum_{i \in S_a} \sum_{j \in S_{a+\alpha}} B_{ji} A_{ij} \Big(q(\beta(\sigma_j - \sigma_i)) - e^{\beta(\sigma_j - \sigma_i)} \Big) e^{-\beta \sigma_j} \right| && \text{from Eq. (3.20)} \\
&\leqslant \frac{1}{\operatorname{tr} e^{-\beta H}} \sum_{a} \Big(\sum_{\alpha=-2}^{2} \varepsilon \|A_{S_a,S_{a+\alpha}}\| \|B_{S_a,S_{a+\alpha}}\| \sum_{j \in S_{a+\alpha}} e^{-\beta \sigma_j} && \text{by Eq. (3.22)} \\
&\qquad + \sum_{\alpha=3}^{\infty} (e^{\eta\beta(\alpha+1)L} + 1) \|A_{S_a,S_{a+\alpha}}\| \|B_{S_a,S_{a+\alpha}}\| \sum_{i \in S_a} e^{-\beta \sigma_i} && \text{by Eq. (3.23)} \\
&\qquad + \sum_{\alpha=-3}^{-\infty} \|B_{S_a,S_{a+\alpha}}\| \|A_{S_a,S_{a+\alpha}}\| \Big(\sum_{i \in S_a} e^{-\beta \sigma_i} + e^{\eta\beta(|\alpha|+1)L} \sum_{j \in S_{a+\alpha}} e^{-\beta \sigma_j} \Big) \Big) \\
& && \text{by Eq. (3.24)} \\
&\leqslant \frac{\|A\| \|B\|}{\operatorname{tr} e^{-\beta H}} \sum_{a} \Big(\sum_{\alpha=-2}^{2} \varepsilon \sum_{j \in S_{a+\alpha}} e^{-\beta \sigma_j} \\
&\qquad + \sum_{\alpha=3}^{\infty} (e^{\eta\beta(\alpha+1)L} + 1) e^{2(c-(|\alpha|-1)L/c)} \sum_{i \in S_a} e^{-\beta \sigma_i} \\
&\qquad + \sum_{\alpha=-3}^{-\infty} e^{2(c-(|\alpha|-1)L/c)} \Big(\sum_{i \in S_a} e^{-\beta \sigma_i} + e^{\eta\beta(|\alpha|+1)L} \sum_{j \in S_{a+\alpha}} e^{-\beta \sigma_j} \Big) \Big) && \text{by Eq. (3.25)} \\
&= \|A\| \|B\| \Big(\sum_{\alpha=-2}^{2} \varepsilon \\
&\qquad + \sum_{\alpha=3}^{\infty} (e^{\eta\beta(\alpha+1)L} + 1) e^{2(c-(|\alpha|-1)L/c)} \\
&\qquad + \sum_{\alpha=-3}^{-\infty} e^{2(c-(|\alpha|-1)L/c)} (1 + e^{\eta\beta(|\alpha|+1)L}) \Big) && \text{by } \textstyle\sum_a \sum_{i \in S_{a+\alpha}} e^{-\beta \sigma_i} = \operatorname{tr}(e^{-\beta H}) \\
&\leqslant (5 + e^{2c}) \|A\| \|B\| \Big(\varepsilon + \sum_{\alpha, |\alpha| \geqslant 3} (1 + e^{\eta\beta(|\alpha|+1)L}) e^{-2(|\alpha|-1)L/c} \Big)
\end{aligned}$$

$$\leqslant (5+e^{2c})\operatorname{tr}(e^{-\beta H})\|A\|\|B\|\left(\varepsilon+\sum_{\alpha,|\alpha|\geqslant 3}2e^{-(|\alpha|-1)L/c}\right)$$

$$\leqslant \tfrac{1}{2}\varepsilon\|A\|\|B\| \qquad \text{taking } L=\kappa/(3\beta)\geqslant C_1\log(1/\varepsilon)$$

The last step is where we need C_1 to be sufficiently large in terms of $\mathfrak{K},\mathfrak{d}$. Plugging this back into the relation at the beginning, we conclude

$$|\operatorname{tr}(Bq(-\beta H|A)\rho)-\operatorname{tr}(AB\rho)|\leqslant \tfrac{1}{2}\varepsilon\|A\|\|B\|$$

as desired. □

Now we can complete the proof of Lemma 3.7.4.

Proof of Lemma 3.7.4. The first two constraints are satisfied by Lemmas 3.8.1 and 3.8.2:

$$|\widetilde{\operatorname{tr}}(A_1A_2(H\rho-\rho H))|^2\leqslant 2|\widetilde{\operatorname{tr}}(A_1A_2(H\rho-\rho H))|^2+2|\operatorname{tr}(A_1A_2(H\rho-\rho H))|^2\leqslant \tfrac{1}{2}\varepsilon^2. \tag{3.26}$$

As for the final constraint, by Theorem 3.5.5, and as stated in (3.18), the polynomial $q_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}$ is a flat exponential approximation with parameters

$$\left(0.001\cdot 2^{\gamma\beta}\log(1/\varepsilon),\frac{5}{\gamma\beta},0.001\varepsilon\right).$$

Thus, by Lemma 3.8.3, we have

$$|\operatorname{tr}(B_2q_{\gamma\beta,\ell_0}(-\beta H|B_1)\rho-\operatorname{tr}(B_1B_2\rho))|\leqslant 0.1\varepsilon$$

for all $B_1,B_2\in\mathcal{B}$. Combining the above with Lemma 3.8.1 immediately implies that all constraints are satisfied. □

3.9 A Sum-of-Squares Proof That the Commutator Is Small (Lemma 3.7.6)

In this section, we prove Lemma 3.7.6. We will only use a subset of the constraints, namely the commutator constraints involving $\widetilde{\mathrm{tr}}(A(H'\rho - \rho H'))$, to prove this. The proof will involve two parts. First, we will show that the constraints imply that $\mathrm{tr}([H, H'](H'\rho - \rho H'))$ is small. Then, we show that $\mathrm{tr}([H, H'](H'\rho - \rho H'))$ being small actually implies that $[H, H']$ must be small.

Lemma 3.9.1 (Trace inner product is bounded)**.** *Under Assumption 3.7.3, we have*

$$\{ \mathcal{C}_{\lambda'} \} \overset{\lambda'}{\vdash} \{ \, |\mathrm{tr}([H, H'](H'\rho - \rho H'))| \leqslant m^2 (10^{\mathfrak{K}}\mathfrak{d})^5 \varepsilon_0 \, \} \, .$$

Proof. By Lemma 3.3.4, we can write $[H, H'] = \sum_{F \in \mathcal{P}_{2\mathfrak{K}}} f_a(\lambda') F$ where each f_a is a linear functions of $\lambda'_1, \ldots, \lambda'_m$. Furthermore we have the following properties

- Each f_a has nonzero coefficients on at most $\mathfrak{d}$ terms
- All of the coefficients in each of the f_a have magnitude at most 1

The total number of terms $F \in \mathcal{P}_{2\mathfrak{K}}$ is at most $m(10^{\mathfrak{K}}\mathfrak{d})^2$ (recall Corollary 3.3.14). We have the constraint $-\varepsilon_0 \leqslant \widetilde{\mathrm{tr}}(F(H'\rho - \rho H')) \leqslant \varepsilon_0$ for each $F \in \mathcal{P}_{2\mathfrak{K}}$, and thus summing over the F's

$$\{ \mathcal{C}_{\lambda'} \} \overset{\lambda'}{\vdash} \Bigg\{ \widetilde{\mathrm{tr}}([H, H'](H'\rho - \rho H')) = \widetilde{\mathrm{tr}}\Big(\sum_{F \in \mathcal{P}_{2\mathfrak{K}}} f_a(\lambda') F (H'\rho - \rho H') \Big) \\ \leqslant \mathfrak{d} \sum_{F \in \mathcal{P}_{2\mathfrak{K}}} \widetilde{\mathrm{tr}}(F(H'\rho - \rho H')) \leqslant m 10^{2\mathfrak{K}} \mathfrak{d}^3 \varepsilon_0 \Bigg\}, \tag{3.27}$$

where the first inequality follows from the constraint that $-1 \leqslant \lambda'_i \leqslant 1$ and that each f_a has at most $\mathfrak{d}$ non-zeros, and the second inequality follows from there being at most $m(10^{\mathfrak{K}}\mathfrak{d})^2$ terms $F \in \mathcal{P}_{2\mathfrak{K}}$ and the corresponding trace being at most ε_0 due to the

constraints. Similarly, we can derive

$$\{\, \mathcal{C}_{\lambda'} \,\} \overset{\lambda'}{\vdash} \left\{ \widetilde{\mathrm{tr}}([H, H'](H'\rho - \rho H')) \geqslant -m 10^{2\mathfrak{K}} \mathfrak{d}^3 \varepsilon_0 \right\}. \tag{3.28}$$

Next, again expanding H and H' in the Pauli basis, we have

$$\overset{\lambda'}{\vdash} \left\{ \mathrm{tr}([H, H'](H'\rho - \rho H')) = \sum_{F_1, F_2 \in \mathcal{P}_{2\mathfrak{K}}} g_{F_1,F_2}(\lambda') \,\mathrm{tr}(F_1 F_2 \rho) \right\}, \tag{3.29}$$

where $g_{F_1,F_2}(\lambda')$ are quadratic functions of the λ' each with at most $\mathfrak{d}$ nonzero monomials and coefficients between ± 2. Similarly,

$$\overset{\lambda'}{\vdash} \left\{ \widetilde{\mathrm{tr}}([H, H'](H'\rho - \rho H')) = \sum_{F_1, F_2 \in \mathcal{P}_{2\mathfrak{K}}} g_{F_1,F_2}(\lambda') \widetilde{\mathrm{tr}}(F_1 F_2 \rho) \right\}. \tag{3.30}$$

Using Assumption 3.7.3 and the constraint that $-1 \leqslant \lambda'_i \leqslant 1$, we have

$$\begin{aligned} \{\, \mathcal{C}_{\lambda'} \,\} \overset{\lambda'}{\vdash} \Bigg\{ & \mathrm{tr}([H, H'](H'\rho - \rho H')) \\ & \leqslant \sum_{F_1, F_2 \in \mathcal{P}_{2\mathfrak{K}}} g_{F_1,F_2}(\lambda') \widetilde{\mathrm{tr}}(F_1 F_2 \rho) + \varepsilon_0 \cdot \sum_{F_1, F_2 \in \mathcal{P}_{2\mathfrak{K}}} g_{F_1,F_2}(\lambda') \\ & = \widetilde{\mathrm{tr}}([H, H'](H'\rho - \rho H')) + \varepsilon_0 \sum_{F_1, F_2 \in \mathcal{P}_{2\mathfrak{K}}} g_{F_1,F_2}(\lambda') \\ & \leqslant m^2 (10^{\mathfrak{K}} \mathfrak{d})^5 \varepsilon_0 \Bigg\}, \end{aligned} \tag{3.31}$$

where the first inequality follows from (3.29), (3.30) and Assumption 3.7.3 and the second follows from (3.27), the fact that g has coefficients between $[-2, 2]$ and there are at most $m^2(10^{\mathfrak{K}}\mathfrak{d})^4$ terms $F_1, F_2 \in \mathcal{P}_{2\mathfrak{K}}$. Similarly, using the lower bound estimate in (3.28) we have

$$\{\, \mathcal{C}_{\lambda'} \,\} \overset{\lambda'}{\vdash} \left\{ \mathrm{tr}([H, H'](H'\rho - \rho H')) \geqslant -m^2 (10^{\mathfrak{K}} \mathfrak{d})^5 \varepsilon_0 \right\}, \tag{3.32}$$

which completes the proof. □

Next, we move on to the second part of the proof. First, we prove the following inequality relating $\mathrm{tr}([H, H'](H'\rho - \rho H'))$ to $\mathrm{tr}((\mathrm{i}[H', H])^2\rho)$. We note that we consider the expression $\mathrm{i}[H', H]$ to ensure that the matrix is Hermitian and the coefficients of the λ_i' are real.

Lemma 3.9.2 (Lower bounding the commutator at arbitrary temperature)**.** *Given* $0 < \beta$, $H' = \sum_{i\in[m]} \lambda_i' E_i$, $H = \sum_{i\in[m]} \lambda_i E_i$ *and* $\rho = e^{-\beta H}$, *we have*

$$\vdash^{\lambda'} \left\{ \frac{\beta}{1 + 2\beta\|H\|} \mathrm{tr}\big((\mathrm{i}[H', H])^2\rho\big) \leqslant \mathrm{tr}([H', H](H'\rho - \rho H')) \right\}.$$

Proof. Consider the basis where H is diagonal and let its eigenvalues be σ_i. Let $Z = \mathrm{tr}(e^{-\beta H})$. Then,

$$\begin{aligned}
\vdash^{\lambda'} \Big\{ \mathrm{tr}([H', H](H'\rho - \rho H')) &= \mathrm{tr}((H'H - HH')(H'\rho - \rho H')) \\
&= \frac{1}{Z} \mathrm{tr}((H' \circ \{\sigma_j - \sigma_i\}_{ij})(H' \circ \{e^{-\beta\sigma_j} - e^{-\beta\sigma_i}\}_{ij})) \\
&= \frac{1}{Z} \sum_{i,j} H'_{ij} H'_{ji} (\sigma_j - \sigma_i)(e^{-\beta\sigma_i} - e^{-\beta\sigma_j}) \\
&= \frac{1}{Z} \sum_{i,j} |H'_{ij}|^2 (\sigma_j - \sigma_i)(e^{-\beta\sigma_i} - e^{-\beta\sigma_j}) \\
&= \frac{1}{Z} \sum_{i,j} |H'_{ij}|^2 (\sigma_j - \sigma_i)(1 - e^{-\beta(\sigma_j - \sigma_i)}) e^{-\beta\sigma_i} \Big\}
\end{aligned} \tag{3.33}$$

By a similar argument,

$$\begin{aligned}
\vdash^{\lambda'} \Big\{ \mathrm{tr}\big((\mathrm{i}[H', H])^2\rho\big) &= \frac{1}{Z} \mathrm{tr}((H' \circ \{\sigma_j - \sigma_i\}_{ij})(H' \circ \{(\sigma_i - \sigma_j) e^{-\beta\sigma_j}\}_{ij})) \\
&= \frac{1}{Z} \sum_{i,j} |H'_{ij}|^2 (\sigma_j - \sigma_i)^2 e^{-\beta\sigma_i} \Big\}.
\end{aligned} \tag{3.34}$$

We observe that, since $e^x \geqslant 1+x$, $e^{-x} \leqslant \frac{1}{1+x}$, and so

$$x(1-e^{-\beta x}) \geqslant |x|(1-e^{-\beta|x|}) \geqslant \frac{\beta x^2}{1+|\beta x|}.$$

Applying this inequality with $x = \sigma_j - \sigma_i$, which are constants in the sum-of-squares proof system, and substituting back into (3.33), we have,

$$\begin{aligned} \vdash^{\lambda'} \Big\{ \operatorname{tr}([H',H](H'\rho - \rho H')) &= \frac{1}{Z}\sum_{i,j}|H'_{ij}|^2(\sigma_j-\sigma_i)(1-e^{-\beta(\sigma_j-\sigma_i)})e^{-\beta\sigma_i} \\ &\geqslant \frac{1}{Z}\sum_{i,j}|H'_{ij}|^2(\sigma_j-\sigma_i)^2\frac{\beta}{1+|\beta||\sigma_j-\sigma_i|}e^{-\beta\sigma_i} \\ &\geqslant \frac{\beta}{1+2\beta\|H\|}\frac{1}{Z}\sum_{i,j}|H'_{ij}|^2(\sigma_j-\sigma_i)^2e^{-\beta\sigma_i} \\ &= \frac{\beta}{1+2\beta\|H\|}\operatorname{tr}\big((\mathrm{i}[H',H])^2\rho\big) \Big\}, \end{aligned} \tag{3.35}$$

where the second inequality follows from $|\sigma_j - \sigma_i| \leqslant 2\|H\|$ and the last equality follows from (3.34).

□

Lemma 3.9.3. *Let $H' = \sum_{a=1}^m \lambda'_a E_a$ and write $\mathrm{i}[H,H'] = \sum_{b=1}^{|\mathcal{P}_{2\mathfrak{K}}|}\gamma_b F_b$ as a linear combination of $2\mathfrak{K}$-local Pauli matrices F_b where each of the γ_b are linear expressions in the λ'_a. Then, for each γ_b, we have*

$$\{\, \mathcal{C}_{\lambda'} \,\} \vdash^{\lambda'} \left\{\, \gamma_b^2 \leqslant e^{\mathcal{O}_{\mathfrak{K},\mathfrak{d}}(\beta)} \operatorname{tr}((\mathrm{i}[H',H])^2\rho) \,\right\}.$$

Proof. Note that $\mathrm{i}[H',H]$ is Hermitian and is $2\mathfrak{K}$-local. Thus, by Corollary 3.3.8, we have that for any real values for the λ'_i, the inequality

$$\gamma_b^2 \leqslant e^{\mathcal{O}_{\mathfrak{K},\mathfrak{d}}(\beta)} \operatorname{tr}((\mathrm{i}[H',H])^2\rho)$$

holds. Now both sides of the above are quadratic expressions in the λ'_i so by Fact 3.3.29, the difference between the two sides can be written as a sum of squares of linear functions of the λ'_i. Thus,

$$\{ \mathcal{C}_{\lambda'} \} \overset{\lambda'}{\vdash} \{ \gamma_b^2 \leqslant e^{\mathcal{O}_{\mathfrak{K},\mathfrak{d}}(\beta)} \operatorname{tr}((\mathrm{i}[H', H])^2 \rho) \}$$

as desired. □

Now we can finish the proof of Lemma 3.7.6 by combining the previous lemmas.

Proof of Lemma 3.7.6. Combining Lemma 3.9.1, Lemma 3.9.2, and Lemma 3.9.3, we get

$$\{ \mathcal{C}_{\lambda'} \} \overset{\lambda'}{\vdash} \{ \gamma_b^2 \leqslant e^{\mathcal{O}_{\mathfrak{K},\mathfrak{d}}(\beta)} m^3 \varepsilon_0 \}$$

since $\|H\| \leqslant m$ (and we can adjust the $O_{\mathfrak{K},\mathfrak{d}}(1)$ in the exponent to absorb the other factors). It follows from Fact 3.3.33 that

$$\{ \gamma_b^2 \leqslant e^{\mathcal{O}_{\mathfrak{K},\mathfrak{d}}(\beta)} m^3 \varepsilon_0 \} \overset{\lambda'}{\vdash} \{ -e^{\mathcal{O}_{\mathfrak{K},\mathfrak{d}}(\beta)} m^{1.5} \sqrt{\varepsilon_0} \leqslant \gamma_b \leqslant e^{\mathcal{O}_{\mathfrak{K},\mathfrak{d}}(\beta)} m^{1.5} \sqrt{\varepsilon_0} \}$$

and this is exactly the desired statement. □

3.10 A Sum-of-Squares Proof of Identifiability (Lemma 3.7.7)

In this section, we prove Lemma 3.7.7. At a high level, we will rely on properties of $q_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}$ proven in Section 3.5. However, since we are working with commutator polynomials, we will need to invoke the translation between polynomials and commutators in Section 3.4 at each step. First, it is critical that H' and H almost commute so that the "error" terms that appear in the translation (those on the RHS of Theorem 3.4.9) are small. We make this precise in the following subsection.

3.10.1 Bounding Error Terms: Polynomials to Nested Commutators

We begin by showing the following lemma:

Lemma 3.10.1 (Bounding switched commutators of type 3)**.** *Let $S, T \in \{0,1\}^*$ be arbitrary sequences of length at most ℓ. Let $A \in \mathcal{P}_{\mathfrak{K}}$. Then we can write*

$$\mathrm{i}^{|S|+|T|}\left[(H,H')_S,[[H',H],[(H,H')_T,A]]\right] = \sum_{G_c \in \mathcal{P}_{4(\ell+1)\mathfrak{K}}} \zeta_c G_c$$

where ζ_c are polynomials in the λ' and the terms G_c are in $\mathcal{P}_{4(\ell+1)\mathfrak{K}}$ and have distance at most $4(\ell+1)\mathfrak{K}$ from the support of A. If Assumption 3.7.3 holds then

$$\{\,\mathcal{C}_{\lambda'}\,\} \vdash^{\lambda'} \Big\{ -(|S|+|T|+1)!(2\mathfrak{d})^{4\ell} e^{\mathcal{O}_{\mathfrak{K},\mathfrak{d}}(\beta)} m^{1.5}\sqrt{\varepsilon_0} \leqslant \zeta_c \leqslant (|S|+|T|+1)!(2\mathfrak{d})^{4\ell} e^{\mathcal{O}_{\mathfrak{K},\mathfrak{d}}(\beta)} m^{1.5}\sqrt{\varepsilon_0} \Big\}.$$

Note that the $\mathrm{i}^{|S|+|T|}$ is to make the expression Hermitian so that the ζ_c are real polynomials.

Proof. We can write $\mathrm{i}[H,H'] = \sum_b \gamma_b F_b$ for $F_b \in \mathcal{P}_{2\mathfrak{K}}$ and by Lemma 3.7.6

$$\{\,\mathcal{C}_{\lambda'}\,\} \vdash^{\lambda'} \{\, -e^{\mathcal{O}_{\mathfrak{K},\mathfrak{d}}(\beta)} m^{1.5}\sqrt{\varepsilon_0} \leqslant \gamma_b \leqslant e^{\mathcal{O}_{\mathfrak{K},\mathfrak{d}}(\beta)} m^{1.5}\sqrt{\varepsilon_0} \,\}\,. \tag{3.36}$$

For notational convenience, for an index $j \in \{1,2,\ldots,|S|+|T|\}$, let $\lambda_{a,[j]}$ be equal to λ_a if the jth entry of the sequence ST (concatenated) is 0 and equal to λ'_a otherwise. Now we can apply Lemma 3.3.12 (with $\mathfrak{K} \leftarrow 2\mathfrak{K}$ and $\mathfrak{d} \leftarrow 10\mathfrak{d}^2$, since $[H,H']$ is $2\mathfrak{K}$-local) to write

$$\begin{aligned}
&\mathrm{i}^{|S|+|T|}\left[(H,H')_S,[[H',H],[(H,H')_T,A]]\right] \\
&= 2^{|S|+|T|+1} \sum_{a_1,\ldots,a_{|S|+|T|},b} c_{a_1,\ldots,a_{|S|+|T|},b} \cdot \left(\prod_{j\in[|S|+|T|]} \lambda_{a_j,[j]}\right) \cdot \gamma_b \cdot A_{a_1,\ldots,a_{|S|+|T|},b}
\end{aligned}$$

where $c_{a_1,\ldots,a_{|S|+|T|},b} \in \pm 1, \pm \mathrm{i}$ and the sum has at most $(|S|+|T|+1)!\mathfrak{d}^{2(|S|+|T|+1)}$ terms and each of the terms $A_{a_1,\ldots,a_{|S|+|T|},b} \in \mathcal{P}_{4(\ell+1)\mathfrak{K}}$ and has distance at most $4(\ell+1)\mathfrak{K}$ from the support of A. Thus, we can rewrite the original commutator in the form $\sum_{G_c \in \mathcal{P}_{4(\ell+1)\mathfrak{K}}} \zeta_c G_c$ where each of the ζ_c is a polynomial with real coefficients (because the original commutator is Hermitian) in the variables λ' of degree at most $2\ell+2$.

Since we have the constraint $-1 \leqslant \lambda'_a \leqslant 1$ and (3.36) and also we know that $-1 \leqslant \lambda_a \leqslant 1$, combining over all of the terms in the above sum,

$$\{\, \mathcal{C}_{\lambda'} \,\} \big|_{2\ell+2}^{\lambda'} \left\{ \zeta_C \leqslant (|S|+|T|+1)!(2\mathfrak{d})^{2(|S|+|T|+1)} e^{\mathcal{O}_{\mathfrak{K},\mathfrak{d}}(\beta)} m^{1.5} \sqrt{\varepsilon_0} \right\} .$$

The lower bound can be obtained in a similar manner. □

Lemma 3.10.2 (Bounding switched commutators of type 1 and 2)**.** *Let $S \in \{0,1\}^*$ be an arbitrary sequence of length at most ℓ. Let $A \in \mathcal{P}_{\mathfrak{K}}$. Then we can write*

$$\mathrm{i}^{|S|} [(H, H')_S, A] = \sum \zeta_c G_c$$

where the terms $G_c \in \mathcal{P}_{(\ell+1)\mathfrak{K}}$ and have distance at most $(\ell+1)\mathfrak{K}$ from the support of A. Furthermore,

$$\{\, \mathcal{C}_{\lambda'} \,\} \big|_{2\ell+2}^{\lambda'} \left\{ |\zeta_C| \leqslant (|S|+1)!(4\mathfrak{d})^{\ell} \right\} .$$

Note that the $\mathrm{i}^{|S|}$ ensures that the expression is Hermitian so that the ζ_c are real.

Proof. The proof is the same as Lemma 3.10.1. We don't need Lemma 3.7.6 at all and just need to use the constraint that $-1 \leqslant \lambda'_a \leqslant 1$ and the fact that $-1 \leqslant \lambda_a \leqslant 1$ to bound the coefficients. □

We will also need the following lemma that bounds the effect of the sampling error. This is analogous to Lemma 3.8.1 but we now need that it is enforced for all potential choices of λ'.

Lemma 3.10.3 (Estimated expectations are close to true expectations)**.** *Under*

Assumption 3.7.3, for all $B_1, B_2 \in \mathcal{B}$,

$$\{ \mathcal{C}_{\lambda'} \} \overset{\lambda'}{\vdash} \{ |\widetilde{\mathrm{tr}}(B_2 q_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}(-\beta H'|B_1)\rho) - \mathrm{tr}(B_2 q_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}(-\beta H'|B_1)\rho)|^2 \leqslant 0.01\varepsilon^2 \}$$

Proof. By Lemma 3.3.12 and Claim 3.5.11, we can write $q_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}(-\beta H'|B_1)$ as a linear combination of the form $\sum_{A\in\mathcal{A}} c_A A$ where $\mathcal{A} \subset \mathcal{P}_{2^{C_{\mathfrak{K},\mathfrak{d}}\beta+1}\ell_0\mathfrak{K}}$ and the coefficients c_A are polynomials in the λ' of degree at most $q_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}$. Furthermore, the sum of the magnitudes of all of the coefficients of all of the c_A is at most $((1+\beta)\mathfrak{d})^{2^{C_{\mathfrak{K},\mathfrak{d}}\beta+5}\ell_0}$. We can then write

$$\vdash \left\{ \widetilde{\mathrm{tr}}(B_2 q_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}(-\beta H' \mid B_1)\rho) - \mathrm{tr}(B_2 q_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}(-\beta H' \mid B_1)\rho) = \sum_{A\in\mathcal{A}} c_A(\widetilde{\mathrm{tr}}(B_2 A\rho) - \mathrm{tr}(B_2 A\rho)) \right\}.$$

Thus, conditioned on the event that Assumption 3.7.3 holds and using the constraints that $-1 \leqslant \lambda' \leqslant 1$, we get the desired bound. □

3.10.2 No Small Local Marginals in Sum-of-Squares

Now we move onto the main proof. At a high level, we will show that if H, H' are far, then there must be certain choices for B_1, B_2 in the constraint system $\{ \mathcal{C}_{\lambda'} \}$ that "witness" this and thus the constraints will be violated. We break up the desired statement into a series of inequalities.

First, we can take a suitable linear combination of the constraints to derive that the following quantities must be small. Note that in the expressions below, B and $[H-H', B] + 0.25[H-H', B]_3$ are playing the role of our "witness".

Lemma 3.10.4 (Nested commutator polynomials are bounded in SoS)**.** *Let q denote the polynomial $q_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}$. Under Assumption 3.7.3, we have for any $B \in \mathcal{P}_{\mathfrak{K}}$,*

$$\{ \mathcal{C}_{\lambda'} \} \overset{\lambda'}{\vdash} \{ |\mathrm{tr}(([H-H', B] + 0.25[H-H', B]_3)(q(-\beta H'|B) - q(-\beta H|B))\rho)| \leqslant (2\mathfrak{d})^{12}\varepsilon^2 \}$$

Proof. Fix any $B \in \mathcal{P}_{\mathfrak{K}}$. By Lemma 3.3.4, we have

$$\left|\frac{\lambda'}{}\right. \left\{ \mathrm{i}([H - H', B] + 0.25[H - H', B]_3) = \sum_{F_b \in \mathcal{P}_{4\mathfrak{K}}} \gamma_b F_b \right\},$$

for some γ_b that are degree-3 polynomials in the indeterminates (λ_i')'s with real coefficients (because the expression is Hermitian) in the λ'. Since $-1 \leqslant \lambda_i \leqslant 1$ and we also have the constraint that $-1 \leqslant \lambda_i' \leqslant 1$, we get

$$\{ \mathcal{C}_{\lambda'} \} \left|\frac{\lambda'}{}\right. \{ -(2\mathfrak{d})^3 \leqslant \gamma_b \leqslant (2\mathfrak{d})^3 \} .$$

Also, at most $\mathfrak{d}^3$ of the γ_b are nonzero. We recall the following statements: for all $B_1, B_2 \in \mathcal{B}$,

$$\{ \mathcal{C}_{\lambda'} \} \left|\frac{\lambda'}{}\right. \left\{ \left|\widetilde{\mathrm{tr}}(B_2 q(-\beta H' \mid B_1)\rho) - \mathrm{tr}(B_2 q(-\beta H' \mid B_1)\rho)\right|^2 \leqslant 0.01\varepsilon^2 \right\}$$

(Lemma 3.10.3)

$$\left|\frac{\lambda'}{}\right. \left\{ \left|\widetilde{\mathrm{tr}}(B_2 q(-\beta H \mid B_1)\rho) - \mathrm{tr}(B_2 q(-\beta H \mid B_1)\rho)\right|^2 \leqslant 0.01\varepsilon^2 \right\},$$

(Lemma 3.8.1)

$$\left|\frac{\lambda'}{}\right. \left\{ \left|\widetilde{\mathrm{tr}}(B_1 B_2 \rho) - \widetilde{\mathrm{tr}}(B_2 q(-\beta H \mid B_1)\rho)\right|^2 \leqslant \varepsilon^2 \right\} \qquad \text{(Lemma 3.7.4)}$$

Using the (last) constraint in the system $\mathcal{C}_{\lambda'}$ and the above, we deduce

$$\{ \mathcal{C}_{\lambda'} \} \left|\frac{\lambda'}{}\right. \left\{ |\mathrm{tr}(B_2 q_{C_{\mathfrak{K},\mathfrak{d}},\beta,\ell_0}(-\beta H'|B_1)\rho) - \mathrm{tr}(B_2 q_{C_{\mathfrak{K},\mathfrak{d}},\beta,\ell_0}(-\beta H|B_1)\rho)|^2 \leqslant 1.44\varepsilon^2 \right\} .$$

Now we can plug in $B_1 = B$ and $B_2 = F_b$ into the above and then take a linear combination with coefficients equal to the γ_b. Using the properties of the γ_b at the beginning, we conclude

$$\{ \mathcal{C}_{\lambda'} \} \left|\frac{\lambda'}{}\right. \left\{ |\mathrm{tr}(([H - H', B] + 0.25[H - H', B]_3)(q(-\beta H'|B) - q(\beta H|B))\rho)| \leqslant (2\mathfrak{d})^6 \varepsilon \right\}$$

as desired. □

On the other hand, we will show using the properties of $q_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}$ in Theorem 3.5.6 that the expression on the LHS in the above is actually lower bounded by some function of $[H - H', B]$. The first step of this is the following lemma.

Lemma 3.10.5 (Key polynomial identity for nested commutators in sum-of-squares)**.** *Let*

$$X(B) = \operatorname{tr}\left(([H - H', B] + 0.25[H - H', B]_3)(q_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}(-\beta H|B) - q_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}(-\beta H'|B))\rho\right) - \frac{1}{2000}\left(\operatorname{tr}([H - H', B]_2 p_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}(-\beta H|B)\rho)\right) .$$

Under Assumption 3.7.3, we have for any $B \in \mathcal{P}_{\mathfrak{K}}$,

$$\{\, \mathcal{C}_{\lambda'} \,\} \overset{\lambda'}{\vdash} \{\mathsf{Re}(X(B)) \geqslant -\varepsilon\}$$
$$\{\, \mathcal{C}_{\lambda'} \,\} \overset{\lambda'}{\vdash} \{|\mathsf{Im}(X(B))| \leqslant \varepsilon\}$$

Proof. By Theorem 3.5.6, we can write the following polynomial equality in two formal variables x, y

$$((x-y)+0.25(x-y)^3)(q_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}(x)-q_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}(y))-\frac{1}{2000}(x-y)^2 p_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}(x) = \sum_{j=1}^{m} r_j(x,y)^2 \tag{3.37}$$

where $m \leqslant 10^{2^{C_{\mathfrak{K},\mathfrak{d}}\beta}\ell_0}$ and each of the polynomials r_j is $(2^{C_{\mathfrak{K},\mathfrak{d}}\beta}\ell_0 + 10, 200^{2^{C_{\mathfrak{K},\mathfrak{d}}\beta}\ell_0})$-bounded. Also note that since H, H', B are Hermitian, for any polynomial q

$$\overset{\lambda'}{\vdash} \left\{\, q(H, H' \mid B)^\dagger = q(-H, -H' \mid B) \,\right\} . \tag{3.38}$$

Thus, we can apply Theorem 3.4.9 to write the following formal identity:

$$\begin{aligned}\vdash^{\lambda'} \Bigg\{ & \operatorname{tr}\left(([H-H',B]+0.25[H-H',B]_3)(q_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}(-\beta H|B)-q_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}(-\beta H'|B))\rho\right) \\ & -\frac{1}{2000}\operatorname{tr}([H-H',B]_2 p_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}(-\beta H|B)\rho) \\ & =\sum_{j=1}^{m}\operatorname{tr}(r_j(H,H'|B)r_j(H,H'|B)^\dagger\rho)+D\Bigg\},\end{aligned} \tag{3.39}$$

where D is a sum of four types of terms given on the RHS of Theorem 3.4.9. By Claim 3.5.11, all of the polynomials in the original identity are $(2^{C_{\mathfrak{K},\mathfrak{d}}\beta+1}\ell_0, 200^{2^{C_{\mathfrak{K},\mathfrak{d}}\beta}\ell_0})$-bounded and thus Theorem 3.4.9 combined with applying Lemma 3.10.1 and Lemma 3.10.2 to the individual terms in D gives us that

$$\begin{aligned}\{\,\mathcal{C}_{\lambda'}\,\} &\vdash^{\lambda'} \left\{|\mathsf{Re}(D)| \leqslant 2^{4^{C_{\mathfrak{K},\mathfrak{d}}\beta}\ell_0} m^{1.5}\sqrt{\varepsilon_0}\right\} \\ \{\,\mathcal{C}_{\lambda'}\,\} &\vdash^{\lambda'} \left\{|\mathsf{Im}(D)| \leqslant 2^{4^{C_{\mathfrak{K},\mathfrak{d}}\beta}\ell_0} m^{1.5}\sqrt{\varepsilon_0}\right\}.\end{aligned}$$

Now by definition, $2^{4^{C_{\mathfrak{K},\mathfrak{d}}\beta}\ell_0} m^{1.5}\sqrt{\varepsilon_0} \leqslant 0.1\varepsilon$. Finally, it remains to note that

$$\{\,\mathcal{C}_{\lambda'}\,\} \vdash^{\lambda'} \left\{\operatorname{tr}(r_j(H,H'|B)r_j(H,H'|B)^\dagger\rho) \geqslant 0\right\}.$$

This is because we can rewrite the LHS above as $\|\rho^{1/2} r_j(H,H'|B)\|_F^2$ which is a sum of squares in the real and imaginary parts of the entries of the matrix $\rho^{1/2} r_j(H,H'|B)$. This matrix has entries that are polynomials with complex coefficients in the λ'. Separating out the real and imaginary parts, the real and imaginary parts of the entries are thus polynomials in the λ' with real coefficients. So overall, $\|\rho^{1/2} r_j(H,H'|B)\|_F^2$ is a sum of squares of polynomials in the λ'. Combining everything with (3.39) completes the proof. □

Next, we analyze the subtracted term $\operatorname{tr}([H-H',B]_2 p_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}(-\beta H|B)\rho)$ in Lemma 3.10.5. We use the property that $p_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}$ is a good approximation to the exponential to relate it to a much simpler expression.

Lemma 3.10.6 (Derivative is uniformly lower bounded)**.** *Define*

$$Y(B) = \operatorname{tr}([H - H', B]_2 p_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}(-\beta H|B)\rho) - \operatorname{tr}((\mathrm{i}[H - H', B])^2\rho)\,.$$

Then for any $B \in \mathcal{P}_{\mathfrak{K}}$

$$\{\,\mathcal{C}_{\lambda'}\,\} \Big|\!\!\frac{\lambda'}{} \{\mathsf{Re}(Y(B)) \geqslant -(2\mathfrak{d})^4\varepsilon\}$$
$$\{\,\mathcal{C}_{\lambda'}\,\} \Big|\!\!\frac{\lambda'}{} \{|\mathsf{Im}(Y(B))| \leqslant (2\mathfrak{d})^4\varepsilon\}\,.$$

Proof. Let

$$Z(B) = \operatorname{tr}([H - H', B]_2 p_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}(-\beta H|B)\rho) - \operatorname{tr}(B[H - H', B]_2\rho)\,.$$

We can write $[H - H', B]_2 = \sum_{F_b \in \mathcal{P}_{3\mathfrak{K}}} \gamma_b F_b$ for some γ_b that are degree-2 polynomials with real coefficients (because the expression is Hermitian) in the λ'. Since $-1 \leqslant \lambda_i \leqslant 1$ and we have the constraint that $-1 \leqslant \lambda'_i \leqslant 1$, we get

$$\{\,\mathcal{C}_{\lambda'}\,\} \Big|\!\!\frac{\lambda'}{} \{-(2\mathfrak{d})^2 \leqslant \gamma_b \leqslant (2\mathfrak{d})^2\}\,. \tag{3.40}$$

Also the number of nonzero γ_b is at most $2\mathfrak{d}^2$. Now by Theorem 3.5.5 the polynomial $p_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}$ is a weak exponential approximation with parameters

$$\left(0.001 \cdot 2^{C_{\mathfrak{K},\mathfrak{d}}\beta}\log(1/\varepsilon), \frac{5}{C_{\mathfrak{K},\mathfrak{d}}\beta}, 0.001\varepsilon\right)\,.$$

and thus by Lemma 3.8.3, we have

$$|\operatorname{tr}(F p_{C_{\mathfrak{K},\mathfrak{d}}\beta,\ell_0}(-\beta H|B)\rho - \operatorname{tr}(BF\rho)| \leqslant 0.1\varepsilon$$

for all $F \in \mathcal{P}_{3\mathfrak{K}}$. Note that the above is simply a numerical inequality and does not involve any variables of the SoS system. Now taking a linear combination of the above over all $F \in \mathcal{P}_{3\mathfrak{K}}$ given by the decomposition $[H - H', B]_2 = \sum_{F_b \in \mathcal{P}_{3\mathfrak{K}}} \gamma_b F_b$ and using

(3.40), we conclude

$$\begin{aligned} \{\, \mathcal{C}_{\lambda'} \,\} &\overset{\lambda'}{\vdash} \{\mathsf{Re}(Z(B)) \geqslant -\mathfrak{d}^4\varepsilon\} \\ \{\, \mathcal{C}_{\lambda'} \,\} &\overset{\lambda'}{\vdash} \{|\mathsf{Im}(Z(B))| \leqslant \mathfrak{d}^4\varepsilon\} \end{aligned} \tag{3.41}$$

Next, note that

$$Y(B) - Z(B) = \operatorname{tr}(B[H - H', B][\rho, H'])$$

and writing $B[H - H', B]$ as a linear combination of elements of $\mathcal{P}_{2\mathfrak{K}}$ and using that $-1 \leqslant \lambda_a \leqslant 1$ and the constraint $-1 \leqslant \lambda'_a \leqslant 1$, we get (from the commutator constraints in $\{\, \mathcal{C}_{\lambda'} \,\}$)

$$\begin{aligned} \{\, \mathcal{C}_{\lambda'} \,\} &\overset{\lambda'}{\vdash} \{|\mathsf{Re}(Y(B) - Z(B))| \leqslant \varepsilon\} \\ \{\, \mathcal{C}_{\lambda'} \,\} &\overset{\lambda'}{\vdash} \{|\mathsf{Im}(Y(B) - Z(B))| \leqslant \varepsilon\} \end{aligned}$$

and combining this with (3.41) completes the proof. □

Lemma 3.10.7 (No small local marginals for $H - H'$)**.** *Fix a $B \in \mathcal{P}_{\mathfrak{K}}$. Write $\mathrm{i}[H - H', B] = \sum_{F_b \in \mathcal{P}_{2\mathfrak{K}}} \gamma_b F_b$ as a linear combination of $2\mathfrak{K}$-local Pauli matrices F_b where each of the γ_b are linear expressions in the λ'. Then, for each γ_b, we have*

$$\{\, \mathcal{C}_{\lambda'} \,\} \overset{\lambda'}{\vdash} \left\{\, \gamma_b^2 \leqslant e^{\mathcal{O}_{\mathfrak{K},\mathfrak{d}}(\beta)} \operatorname{tr}((\mathrm{i}[H - H', B])^2 \rho) \,\right\}.$$

Proof. The proof is exactly the same as the proof of Lemma 3.9.3. □

Lemma 3.10.8 (Selecting each coefficient in Hamiltonian)**.** *Let $X = \sum_{E_a \in \mathcal{P}_{\mathfrak{K}}} x_a E_a$ be written as a linear combination of $\mathfrak{K}$-local Pauli matrices. Then for any $E_a \neq I$, there exist $B \in \mathcal{P}_{\mathfrak{K}}$ and $P \in \mathcal{P}_{2\mathfrak{K}}$ whose supports intersect with E_a such that*

$$\vdash \left\{ \left(\frac{1}{2} \operatorname{tr}(\mathrm{i}[X, B]P) \right)^2 = x_a^2 \right\},$$

Proof. Note that for any $A_1, A_2, B \in \mathcal{P}_{\mathfrak{K}}$ such that $[A_1, B], [A_2, B] \neq 0$ and $A_1 \neq A_2$,

then $[A_1, B] \neq [A_2, B]$. In other words, taking the commutator with a fixed Pauli matrix is injective as long as the commutator is nonzero. Now since $E_a \neq I$, clearly we can choose $B \in \mathcal{P}_{\mathfrak{K}}$ with the same support as E_a such that $[B, E_a] \neq 0$. By Lemma 3.3.4, we can write $\mathrm{i}[X, B]$ as a linear combination of $2\mathfrak{K}$-local Pauli matrices and the coefficient of $\mathrm{i}[B, E_a]/2$ (which is a $2\mathfrak{K}$-local Pauli matrix itself) is $\pm 2x_a$. Thus, taking P to be $\mathrm{i}[B, E_a]/2$ gives the desired equality. □

Now we can complete the proof of Lemma 3.7.7.

Proof of Lemma 3.7.7. Combining Lemma 3.10.4, Lemma 3.10.5, Lemma 3.10.6 gives

$$\{ \mathcal{C}_{\lambda'} \} \overset{\lambda'}{\vdash} \{\mathrm{tr}((\mathrm{i}[H - H', B])^2 \rho) \leqslant (10\mathfrak{d})^6 \varepsilon\} .$$

Note that $\mathrm{i}[H - H', B]$ is Hermitian so the expression on the LHS is a real polynomial in the λ'. Now fix an index a and we will analyze $\lambda_a - \lambda'_a$. By Lemma 3.10.8, we can find $B \in \mathcal{P}_{\mathfrak{K}}$ and $F_b \in \mathcal{P}_{2\mathfrak{K}}$ such that the coefficient of F_b in the representation $\mathrm{i}[H - H', B] = \sum_{F_b \in \mathcal{P}_{2\mathfrak{K}}} \gamma_b F_b$ is $\pm 2(\lambda_a - \lambda'_a)$. Then Lemma 3.10.7 implies that

$$\{ \mathcal{C}_{\lambda'} \} \overset{\lambda'}{\vdash} \{(\lambda_a - \lambda'_a)^2 \leqslant 2^{C_{\mathfrak{K},\mathfrak{d}} \beta} \varepsilon\}$$

as long as $C_{\mathfrak{K},\mathfrak{d}}$ is sufficiently large in terms of $\mathfrak{K}, \mathfrak{d}$ and this completes the proof. □

3.11 A Faster Algorithm

We now complete the proof of Theorem 3.7.1. The key to obtaining the faster runtime is observing that actually only a specific family of monomials appear in all of the sum-of-squares proofs in the previous sections. The key observation is formally stated below.

Definition 3.11.1 (Relevant monomials via cluster expansion)**.** We say a monomial in the variables λ', say $\lambda'_{a_1} \lambda_{a_2} \cdots \lambda'_{a_c}$, is relevant if there exist three clusters $(E^{[i]}_{a_1}, \ldots, E^{[i]}_{a_C})$ for $i \in \{1, 2, 3\}$, $E^{[i]}_{a_j} \in \mathcal{P}_{\mathfrak{K}}$ such that

- $C \leqslant 10 \cdot 4^{C_{\mathfrak{K},\mathfrak{d}}\beta} \log(1/\varepsilon)$
- $\{E_{a_1}, \ldots, E_{a_c}\} \subseteq \bigcup_i \{E_{a_1}^{[i]}, \ldots, E_{a_C}^{[i]}\}$ as unordered multisets

We use $\mathcal{L}$ to denote the set of all relevant monomials in the λ'.

Lemma 3.11.2. *Fix a $E_a \in \mathcal{P}_{\mathfrak{K}}$. Then we can write*

$$2^{C_{\mathfrak{K},\mathfrak{d}}\beta}\varepsilon - (\lambda_a - \lambda'_a)^2 = \sum_i r_i(\lambda')^2 \prod_{g \in S_i} g(\lambda')$$

where the r_i are polynomials in the λ' and S_i are subsets of constraints in $\{\mathcal{C}_{\lambda'}\}$ (written in the form $g(\lambda') \geqslant 0$). Furthermore, each product of constraints $\prod_{g \in S_i} g(\lambda')$ involves

- *At most one commutator constraint*
- *At most one polynomial approximation constraint*
- *A product of the constraints $\lambda_a + 1 \geqslant 0, \lambda_a - 1 \leqslant 0$ where the λ_a that appear form a relevant monomial*

Proof. This follows from examining the proofs in Section 3.9 and Section 3.10.2, using Lemma 3.3.12 to characterize all of the monomials that appear whenever we expand a nested commutator in H and H'. Note we can ensure that all of the constraints being multiplied in the sets S_i are distinct because if any constraint is multiplied twice, we get a term of the form $g(\lambda')^2$ and can instead fold it into the $r_i(\lambda')^2$. □

Now using Lemma 3.3.13, we can count the total number of relevant monomials.

Lemma 3.11.3. *The total number of distinct relevant monomials is at most $m \cdot (1/\varepsilon)^{10^{C_{\mathfrak{K},\mathfrak{d}}\beta}}$.*

Proof. This follows immediately from Lemma 3.3.13. □

Finally, we have the following theorem from Steurer and Tiegel [198] that allows us to solve sum-of-squares systems with running time depending only on the number of monomials that appear in the proofs.

Theorem 3.11.4 (Degree reduction via linearization[198])**.** *Let $p : \mathbb{R}^n \to \mathbb{R}$ be a multivariate polynomial of degree at most t. Suppose that there exists a system of polynomial inequalities $\mathcal{A} = \{q_1 \geqslant 0, \ldots, q_m \geqslant 0\}$ such that $\mathcal{A} \left|\frac{t}{x}\right. \{p(x) \geqslant 0\}$ and further assume that this proof can be written in the form*

$$\sum_i r_i(x)^2 \prod_{j \in S_i} q_j(x)$$

where $S_i \subseteq [m]$ and there are at most M distinct sets S_i. Also assume that the number of distinct monomials that appear in $p(x)$ is at most N. Then we can write a polynomial system $\mathcal{A}'$ in x and some additional auxiliary variables such that

- *$\mathcal{A}'$ is feasible whenever $\mathcal{A}$ is feasible*
- *$\mathcal{A}' \left|\vdash \{p(x) \geqslant 0\}\right.$*

and we can compute a pseudoexpectation satisfying $\mathcal{A}'$ in time $O(m + M + (tN)^3)$.

Proof of Theorem 3.7.1. The proof is exactly the same as the proof of Theorem 3.7.8 except using Lemma 3.11.3 and Lemma 3.11.2 to bound the complexity of the final sum-of-squares proof and using Theorem 3.11.4 to solve the sum-of-squares system in running time $\mathrm{poly}(m, \log(1/\delta), (1/\varepsilon)^{2^{f(\mathfrak{K},\mathfrak{d})\beta}})$. □

Part II

Learning General Quantum States

Chapter 4

Learning from Single-Copy Measurements

Now we move onto the second thread of this thesis. In the first part of this thesis, we focused on local Hamiltonians and their associated Gibbs states, which define systems of particles that interact locally. In the next part of this thesis, we will study learning general quantum states, which may have arbitrary non-local interactions.

We study *quantum state tomography*, which is perhaps the most basic inverse problem for quantum systems. The learner is given n copies of an unknown d-dimensional quantum mixed state ρ, and the goal is to output $\widehat{\rho}$ which is close to ρ in some measure of distance *; here we will consider probably the two most ubiquitous and well-studied notions — trace distance and infidelity.

In the *coherent* setting, that is, when the learner is allowed to make arbitrary measurements to the product state $\rho^{\otimes n}$, the situation is generally well-understood. The works [170, 102] established the optimal rate for this problem, that $n = \Theta(d^2/\varepsilon^2)$ copies are necessary and sufficient. The aforementioned paper of [102] also showed that, up to a logarithmic factor, $\Omega(d^2/\gamma)$ copies are sufficient and necessary to learn a state to

*Generally it will suffice to consider the setting where the success probability is some fixed constant.

infidelity γ.

However, in virtually all cases, the upper bounds that achieve the optimal rates require heavily entangled measurements over the joint state $\rho^{\otimes n}$. These measurements require all n copies of ρ to be prepared simultaneously, which renders such approaches currently impractical, at least in the near term. In light of this, there has been a recent surge of interest in the *incoherent* a.k.a. *single-copy* setting. Here, the algorithm is restricted to only making measurements of a single copy of ρ at a time. While such measurements are strictly weaker than general entangled measurements, they are much more practical and can be performed on real world quantum computers [123].

In this chapter, we will essentially resolve the complexity of state tomography with incoherent measurements both for trace distance and infidelity. En route, we will develop a new learning-theoretic framework for dealing with adaptivity, one of the most significant challenges for reasoning about incoherent measurements.

4.1 Results

Throughout, let $\rho \in \mathbb{C}^{d \times d}$ be the unknown state. Our goal will be to estimate ρ given measurement access to copies of ρ.

Incoherent Measurements. Next, we formally define what we mean by an algorithm that uses incoherent measurements. Intuitively, such an algorithm operates as follows: given n copies of ρ, it iteratively measures the i-th copy using a POVM (which could depend on the results of previous measurements), records the outcome, and then repeats this process on the $(i+1)$-th copy. After having performed all n measurements, it must output an answer based on the (classical) sequence of outcomes it has received.

One of the main difficulties that arises when dealing with incoherent measurements is understanding the power of *adaptivity*. In general, an algorithm that uses incoherent measurements can sequentially measure each copy of ρ, and moreover can choose how

to measure the i-th copy of ρ based on the results of the previous $i-1$ measurements. We say such an algorithm is *adaptive*. In contrast, a *nonadaptive* algorithm must specify all of its measurements ahead of time.

Our Results We fully resolve the copy complexity of quantum state tomography with (potentially adaptively chosen) incoherent measurements, for both learning in trace distance and learning in infidelity. For trace distance, our main result is the following:

Theorem 4.1.1 (Informal, see Theorem 4.2.11)**.** *With incoherent measurements, $\Omega(d^3/\varepsilon^2)$ copies are necessary for tomography of d-dimensional states to ε error in trace distance.*

This matches the upper bound given by nonadaptive algorithms, and therefore for this problem, adaptivity provably does not help. In contrast, for learning in infidelity, our main result is the following:

Theorem 4.1.2 (Informal, see Theorem 4.5.1)**.** *With adaptive, incoherent measurements, $\widetilde{O}(d^3/\gamma)$ copies[†] are sufficient for tomography of d-dimensional states to γ infidelity.*

Because trace distance and fidelity are related via $1 - F(\rho,\sigma) \leqslant \|\rho-\sigma\|_1 \leqslant \sqrt{2(1-F(\rho,\sigma))}$ [89], Theorem 4.1.1 implies that $\Omega(d^3/\gamma)$ copies are *necessary* for tomography in infidelity, so up to polylogarithmic factors, our upper bound in Theorem 4.1.2 settles the complexity of tomography in infidelity with incoherent measurements. This is an improvement upon the lower bound of $\Omega(d^3/\gamma^2)$ for nonadaptive measurements proven in [102], giving a provable separation between adaptive and nonadaptive protocols.

Our Techniques Here, we give a very high level discussion of our proof techniques for Theorems 4.1.1 and 4.1.2.

[†]We use $f = \widetilde{O}(g)$ to denote that $f = C \cdot g \cdot \log^c(g))$ for some absolute constants $c, C > 0$.

We begin with the lower bound. The main difficulty with proving lower bounds for state tomography with incoherent measurements is that essentially all the existing lower bound frameworks in the literature were fundamentally for testing problems. In such settings, it suffices to demonstrate hardness for a point-versus-mixture distinguishing task, where the goal is to distinguish between the case where the unknown mixed state is a single point versus the case where it is drawn from a mixture over alternate hypotheses. Such a setup is mathematically nice because the resulting likelihood ratios have a (relatively) simple multilinear form. However, for learning tasks, no such reduction exists; indeed, it is more natural to demonstrate hardness for a mixture-versus-mixture distinguishing task, but here the resulting likelihood ratios are much more complicated. Indeed, this phenomena seems to appear more generally in a variety of other (classical) learning settings, see e.g. [189].

We avoid this by directly bounding how much information the algorithm can learn from incoherent measurements, a technique which we believe may be of independent interest. We demonstrate that, for a carefully chosen prior on mixed states, the posterior distribution of the algorithm after $o(d^3/\varepsilon^2)$ incoherent measurements is anti-concentrated around the true mixed state. It is perhaps surprising that we are able to directly bound the behavior of the posterior distribution, but it turns out that the "tilt" caused by the measurements can be upper bounded "by hand", and then the required anti-concentration follows from classical results in random matrix theory.

We now turn our attention to the upper bound. First, let us briefly discuss why adaptivity may help for learning in infidelity. At a very high level, learning in infidelity seems to correspond to learning the eigenvalues and eigenvectors of the density matrix to some degree of relative accuracy. However, nonadaptive algorithms are unable to do this easily: the small eigenspaces are hidden by the "noise" caused by the large eigenvalues. One the other hand, adaptive algorithms can learn the large eigenspaces, then project them away to reveal the information about the smaller eigenvalues.

The main challenge with this approach is dealing with the error accumulation in

the infidelity as we iterate this process. In particular, we cannot exactly learn the large eigenspaces, and so we must control the error the projections incur. In constant dimensions, one can directly brute force the calculations, but in high dimensions, the calculations become intractable. Our main technical contribution for this part of the paper is a new technique to bound the fidelity between two noncommuting states, by carefully guessing a matrix square root for their symmetrized product.

Concurrent work. Our proof of Theorem 4.1.2 directly generalizes to give that $\widetilde{O}(dr^2/\gamma)$ measurements are sufficient when the unknown state has rank-r. In concurrent and independent work, Flammia and O'Donnell [84] also obtain, up to polylogarithmic factors, an upper bound of $O(dr^2/\gamma)$ for tomography with adaptive single-copy measurements. Their guarantee is somewhat stronger as their error is measured in quantum relative entropy, which is an upper bound on infidelity. They also obtain a bound of $\widetilde{O}(d^{3/2}r^{3/2}/\gamma)$ for the same problem where γ is given by Bures chi-squared divergence, which also upper bounds infidelity

4.1.1 Additional Related Work

Apart from the aforementioned bounds, there have also been lower bounds in the incoherent setting when the measurements are partially adaptive or come from a set of bounded size [156], and when the measurements are Pauli [86].

We also note there is a large literature on understanding the power that adaptivity affords for tomography in infidelity in the *asymptotic* setting [28, 113, 159, 81]. They obtain rates that are linear in $1/\gamma$, but unlike our Theorem 4.1.2, these results get some unspecified dependence on d and/or only apply to the regime of $d = O(1)$.

[205] gives upper and lower bounds for *pure state* tomography in a different setting where instead of getting copies of the unknown state, one has access to a unitary which prepares the state.

Finally, the recent work of [80] gives a constant-factor separation between non-adaptive

and adaptive algorithms for quantum hypothesis selection, as well as a polynomial separation for a problem in which one is promised that an unknown state can be diagonalized in one of m known bases and would like to approximate the state in trace distance by measuring copies. Their separation is with respect to the parameters d and m, rather than ε as in our work.

4.2 Proof of Lower Bound

In this section we prove Theorem 4.2.11, our hardness result for tomography in trace distance.

4.2.1 Lower Bound Framework

We begin by reviewing a standard framework for representing an adaptive algorithm as a tree.

Definition 4.2.1 (Tree representation, see e.g. [66])**.** Fix an unknown d-dimensional mixed state ρ. An algorithm for state tomography that only uses n incoherent, possibly adaptive, measurements of ρ can be expressed as a pair $(\mathcal{T}, \mathcal{A})$, where $\mathcal{T}$ is a rooted tree $\mathcal{T}$ of depth n satisfying the following properties:

- Each node is labeled by a string of vectors $\boldsymbol{x} = (x_1, \ldots, x_t)$, where each x_i corresponds to measurement outcome observed in the i-th step.

- Each node $\boldsymbol{x}$ is associated with a probability $p^\rho(\boldsymbol{x})$ corresponding to the probability of observing $\boldsymbol{x}$ over the course of the algorithm. The probability for the root is 1.

- At each non-leaf node, we measure ρ using a rank-1 POVM $\left\{\omega_x d \cdot xx^\dagger\right\}_x$ to obtain classical outcome $x \in \mathbb{S}^{d-1}$. The children of $\boldsymbol{x}$ consist of all strings $\boldsymbol{x}' = (x_1, \ldots, x_t, x)$ for which x is a possible POVM outcome.

- If $\boldsymbol{x}' = (x_1, \ldots, x_t, x)$ is a child of $\boldsymbol{x}$, then

$$p^\rho(\boldsymbol{x}') = p^\rho(\boldsymbol{x}) \cdot \omega_x d \cdot x^\dagger \rho x \,. \tag{4.1}$$

- Every root-to-leaf path is length-n. Note that $\mathcal{T}$ and ρ induce a distribution over the leaves of $\mathcal{T}$.

$\mathcal{A}$ is a randomized algorithm that takes as input any leaf $\boldsymbol{x}$ of $\mathcal{T}$ and outputs a state $\mathcal{A}(\boldsymbol{x})$. The output of $(\mathcal{T}, \mathcal{A})$ upon measuring n copies of a state ρ is the random variable $\mathcal{A}(\boldsymbol{x})$, where $\boldsymbol{x}$ is sampled from the aforementioned distribution over the leaves of $\mathcal{T}$.

We briefly note that in this definition, we assume that the POVMs are always rank-1. It is a standard fact that this is without loss of generality (see e.g. [66, Lemma 4.8]).

We also recall the definition of the Gaussian Orthogonal Ensemble (GOE) and define a trace-centered variant, which will be the basis of our hard distribution.

Definition 4.2.2 (GOE, Trace-centered GOE)**.** A sample $G \sim \mathrm{GOE}(d)$ is a symmetric matrix with independent Gaussians on and above the diagonal, with $G_{i,i} \sim \mathcal{N}(0, 2/d)$ and $G_{i,j} \sim \mathcal{N}(0, 1/d)$ for $i < j$. A sample $G' \sim \mathrm{GOE}^*(d)$ is sampled by $G' = G - \mathrm{tr}(G)$ where $G \sim \mathrm{GOE}(d)$.

We recall the following standard fact about extremal eigenvalues of the GOE matrix.

Lemma 4.2.3 ([66, Lemma 6.2])**.** *If $G \sim \mathrm{GOE}^*(d)$, then $\|G\|_{\mathrm{op}} \leqslant 3$ with probability $1 - e^{-\Omega(d)}$.*

4.2.2 Construction of Hard Distribution

We construct the following hard distribution μ of quantum states. Let $U \subseteq \mathbb{R}^{d \times d}$ be the affine subspace of symmetric matrices with trace 1 and $U_0 \subseteq \mathbb{R}^{d \times d}$ be the linear subspace of symmetric matrices with trace 0. These spaces inherit the inner product

of $\mathbb{R}^{d\times d}$, which defines Lebesgue measures Leb_U and Leb_{U_0} on them. Let $\sigma = \frac{1}{100}$ be a small constant. A sample $\rho \sim \mu$ is generated by

$$\rho = \frac{1}{d}(I_d + \sigma G),$$

where G is a sample from $\mathrm{GOE}^*(d)$ conditioned on $\|G\|_{\mathrm{op}} \leqslant 4$. Note that such matrices are clearly psd and thus valid quantum states. Concretely, μ has density (with respect to Leb_U)

$$\mu(\rho) = \frac{1}{Z}\exp\left(-\frac{d^3}{4\sigma^2}\|\rho - \frac{1}{d}I_d\|_F^2\right)\mathbb{1}\{\rho \in S_{\mathrm{supp}}\}, \qquad S_{\mathrm{supp}} = \left\{\rho \in U : \|\rho - \frac{1}{d}I_d\|_{\mathrm{op}} \leqslant \frac{4\sigma}{d}\right\}.$$

where Z is a normalizing constant. Further define a set of "good" states

$$S_{\mathsf{good}} = \left\{\rho \in U : \|\rho - \frac{1}{d}I_d\|_{\mathrm{op}} \leqslant \frac{3\sigma}{d}\right\},$$

which corresponds to the event $\|G\|_{\mathrm{op}} \leqslant 3$. Due to Lemma 4.2.3, $\mu(S_{\mathsf{good}}) \geqslant 1 - e^{-\Omega(d)}$.

In the below proof, we will show that all $\rho_0 \in S_{\mathsf{good}}$ are hard to learn. The important property of S_{good} is that it is far from the boundary of $\mathrm{supp}(\mu) = S_{\mathrm{supp}}$; this ensures that we can choose a suitable sub-sampling of μ in a neighborhood of ρ_0, which is isotropic around ρ_0.

Finally we record the following straightforward fact.

Lemma 4.2.4. *For all $\rho, \rho' \in S_{\mathrm{supp}}$, $\exp(-4d^2) \leqslant \mu(\rho)/\mu(\rho') \leqslant \exp(4d^2)$.*

Proof. For all $\rho \in S_{\mathrm{supp}}$,

$$0 \leqslant \frac{d^3}{4\sigma^2}\|\rho - \frac{1}{d}I\|_F^2 \leqslant \frac{d^4}{4\sigma^2}\|\rho - \frac{1}{d}I\|_{\mathrm{op}}^2 \leqslant 4d^2. \qquad \square$$

4.2.3 Anticoncentration of Posterior Distribution

Fix a tomography algorithm $(\mathcal{T}, \mathcal{A})$ as in Definition 4.2.1, and let $\mathcal{T}_\rho$ denote the distribution over observation sequences $\boldsymbol{x} = (x_1, \ldots, x_n)$ when $\mathcal{T}$ is run on state ρ. Note that for any states ρ, ρ' in the support of μ, the likelihood ratio

$$\frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho'}}(\boldsymbol{x}) = \prod_{i=1}^{n} \frac{x_i^\dagger \rho x_i}{x_i^\dagger \rho' x_i}$$

is well defined, since μ is supported on full-rank matrices. Let $\nu_{\boldsymbol{x}}$ denote the posterior distribution of ρ given observations $\boldsymbol{x}$. The density ratio of any $\rho, \rho' \in S_{\text{supp}}$ under $\nu_{\boldsymbol{x}}$ is given by Bayes' rule, and equals

$$\frac{\nu_{\boldsymbol{x}}(\rho)}{\nu_{\boldsymbol{x}}(\rho')} = \frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho'}}(\boldsymbol{x}) \cdot \frac{\mu(\rho)}{\mu(\rho')}.$$

So, for an arbitrary reference state $\rho' \in S_{\text{supp}}$ (below we take $\rho' = \rho_0$, the unknown true state) the density of $\nu_{\boldsymbol{x}}$ is

$$\nu_{\boldsymbol{x}}(\rho) = \frac{1}{Z_{\boldsymbol{x}}} \frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho'}}(\boldsymbol{x}) \mu(\rho), \qquad Z_{\boldsymbol{x}} = \int_U \frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho'}}(\boldsymbol{x}) \mu(\rho) \, \mathrm{d}\mathsf{Leb}_U(\rho).$$

The main technical component of the proof is the following anti-concentration result for $\nu_{\boldsymbol{x}}$.

Definition 4.2.5. For $\rho \in U$, let $B(\rho, \varepsilon)$ denote the ball $\{\rho' \in U : \|\rho' - \rho\|_1 \leqslant \varepsilon\}$. Similarly for $\rho \in U_0$, let $B(\rho, \varepsilon) = \{\rho' \in U_0 : \|\rho' - \rho\|_1 \leqslant \varepsilon\}$.

Theorem 4.2.6. *Suppose $d \gg 1$, $\varepsilon \leqslant \varepsilon_0$ for an absolute constant ε_0, and $n \ll d^3/\varepsilon^2$. If $\rho \in S_{\mathsf{good}}$ and $\boldsymbol{x} \sim \mathcal{T}_{\rho_0}$, there is an event $S_{\rho_0} \in \sigma(\boldsymbol{x})$, with $\mathbb{P}(\boldsymbol{x} \in S_{\rho_0}) \geqslant 1 - \exp(-d^2)$, on which $\nu_{\boldsymbol{x}}(B(\rho_0, \varepsilon)) \ll 1$.*

Let C be a large constant we will set later and $\varepsilon_0 = \sigma/C$. The starting point of the

proof of Theorem 4.2.6 is the estimate

$$\frac{\nu_{\boldsymbol{x}}(B(\rho_0, C\varepsilon))}{\nu_{\boldsymbol{x}}(B(\rho_0, \varepsilon))} = \frac{\int_{B(\rho_0, C\varepsilon)} \frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x})\mu(\rho)\ \mathrm{d}\mathsf{Leb}_U(\rho)}{\int_{B(\rho_0, \varepsilon)} \frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x})\mu(\rho)\ \mathrm{d}\mathsf{Leb}_U(\rho)} \tag{4.2}$$

$$\geqslant \exp(-4d^2)\frac{\int_{B(\rho_0, C\varepsilon)} \frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x})\mathbb{1}\{\rho \in S_{\mathrm{supp}}\}\ \mathrm{d}\mathsf{Leb}_U(\rho)}{\int_{B(\rho_0, \varepsilon)} \frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x})\mathbb{1}\{\rho \in S_{\mathrm{supp}}\}\ \mathrm{d}\mathsf{Leb}_U(\rho)}, \tag{4.3}$$

where the second line uses Lemma 4.2.4. Applying Lemma 4.2.4 in this way amounts to replacing μ in the numerator of (4.2) with a measure that sub-samples it, and in the denominator with a measure that upper bounds it. Define the volumes

$$V_1 = \int_{B(0,1)} \mathrm{d}\mathsf{Leb}_{U_0}(\rho), \qquad V_2 = \int_{B(0,1)} \mathbb{1}\left\{\|\rho\|_{\mathrm{op}} \leqslant 1/d\right\}\ \mathrm{d}\mathsf{Leb}_{U_0}(\rho).$$

We now separately bound the numerator and denominator of (4.3) in terms of these volumes, beginning with the denominator.

Lemma 4.2.7. *If $\rho_0 \in S_{\mathsf{good}}$, there is an event $S_{\rho_0} \in \sigma(\boldsymbol{x})$, with $\mathbf{Pr}(\boldsymbol{x} \in S_{\rho_0}) \geqslant 1 - \exp(-d^2)$, on which*

$$\int_{B(\rho_0, \varepsilon)} \frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x})\mathbb{1}\{\rho \in S_{\mathrm{supp}}\}\ \mathrm{d}\mathsf{Leb}_U(\rho) \leqslant \exp(d^2)\varepsilon^{(d+2)(d-1)/2}V_1\,.$$

Proof. Note that $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{T}_{\rho_0}} \frac{\mathrm{d}T_\rho}{\mathrm{d}T_{\rho_0}}(\boldsymbol{x}) = 1$. So

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{T}_{\rho_0}} \int_{B(\rho_0, \varepsilon)} \frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x})\mathbb{1}\{\rho \in S_{\mathrm{supp}}\}\ \mathrm{d}\mathsf{Leb}_U(\rho) &= \int_{B(\rho_0, \varepsilon)} \mathbb{1}\{\rho \in S_{\mathrm{supp}}\}\ \mathrm{d}\mathsf{Leb}_U(\rho) \\ &\leqslant \int_{B(\rho_0, \varepsilon)} \mathrm{d}\mathsf{Leb}_U(\rho) = \varepsilon^{(d+2)(d-1)/2}V_1.\end{aligned}$$

The exponent $(d+2)(d-1)/2$ comes from the fact that the space of symmetric matrices has dimension $d(d+1)/2$, so U has dimension $d(d+1)/2 - 1 = (d+2)(d-1)/2$. The result follows from Markov's inequality. □

Before bounding the numerator of (4.3), we define the set

$$N_*(\rho_0) = \left\{\rho \in U : \|\rho - \rho_0\|_{\mathrm{op}} \leqslant \frac{C\varepsilon}{d}, \|\rho - \rho_0\|_1 \leqslant C\varepsilon\right\},$$

which is an isotropic neighborhood of ρ_0. Let $\gamma(\rho_0)$ denote the uniform distribution on $N_*(\rho_0)$ (w.r.t. Leb_U). That is, for bounded measurable test function $f : U \to \mathbb{R}$,

$$\mathop{\mathbb{E}}_{\rho\sim\gamma(\rho_0)} f(\rho) = \frac{\int_U f(\rho)\mathbb{1}\{\rho \in N_*(\rho_0)\}\mathrm{d}\mathsf{Leb}_U(\rho)}{\int_U \mathbb{1}\{\rho \in N_*(\rho_0)\}\mathrm{d}\mathsf{Leb}_U(\rho)}.$$

Lemma 4.2.8. *If $\rho_0 \in S_{\mathsf{good}}$ and $\varepsilon \leqslant \varepsilon_0$, then*

$$\int_{B(\rho_0,C\varepsilon)} \frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x})\mathbb{1}\{\rho \in S_{\mathrm{supp}}\}\ \mathrm{d}\mathsf{Leb}_U(\rho) \geqslant (C\varepsilon)^{(d+2)(d-1)/2} \mathop{\mathbb{E}}_{\rho\sim\gamma(\rho_0)}\left[\frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x})\right] V_2. \quad (4.4)$$

Proof. Because $\rho_0 \in S_{\mathsf{good}}$, we have $\|\rho_0 - \frac{1}{d}I_d\|_{\mathrm{op}} \leqslant \frac{3\sigma}{d}$. Because $\varepsilon \leqslant \varepsilon_0 = \sigma/C$, if ρ satisfies $\|\rho - \rho_0\|_{\mathrm{op}} \leqslant \frac{C\varepsilon}{d}$ we have the implication chain

$$\|\rho - \rho_0\|_{\mathrm{op}} \leqslant \frac{C\varepsilon}{d} \Rightarrow \|\rho - \rho_0\|_{\mathrm{op}} \leqslant \frac{\sigma}{d} \Rightarrow \|\rho - \frac{1}{d}I_d\|_{\mathrm{op}} \leqslant \frac{4\sigma}{d} \Rightarrow \rho \in S_{\mathrm{supp}}.$$

So, letting X denote the left-hand side of (4.4), we have

$$\begin{aligned} X &\geqslant \int_{B(\rho_0,C\varepsilon)} \frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x})\mathbb{1}\left\{\|\rho - \rho_0\|_{\mathrm{op}} \leqslant \frac{C\varepsilon}{d}\right\}\ \mathrm{d}\mathsf{Leb}_U(\rho) = \int_U \frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x})\mathbb{1}\{\rho \in N_*(\rho_0)\}\ \mathrm{d}\mathsf{Leb}_U(\rho) \\ &= \mathop{\mathbb{E}}_{\rho\sim\gamma(\rho_0)}\left[\frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x})\right] \int_U \mathbb{1}\{\rho \in N_*(\rho_0)\}\ \mathrm{d}\mathsf{Leb}_U(\rho). \end{aligned}$$

Finally, note that

$$\int_U \mathbb{1}\{\rho \in N_*(\rho_0)\}\ \mathrm{d}\mathsf{Leb}_U(\rho) = \int_{B(\rho_0,C\varepsilon)} \mathbb{1}\left\{\|\rho - \rho_0\|_{\mathrm{op}} \leqslant \frac{C\varepsilon}{d}\right\}\ \mathrm{d}\mathsf{Leb}_U(\rho) = (C\varepsilon)^{(d+2)(d-1)/2} V_2,$$

which conludes the proof. □

It remains to control $\mathbb{E}_{\rho\sim\gamma(\rho_0)}\left[\frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x})\right]$ and the volume ratio V_2/V_1. The former

measures the information gained from the observations $\boldsymbol{x}$. It is bounded by the following lemma, which formalizes the intuition that each observation improves the log likelihood ratio by at most $O(\varepsilon^2/d)$. This is the step that uses the hypothesis $n \ll d^3/\varepsilon^2$.

Lemma 4.2.9. *If $\rho_0 \in S_{\mathsf{good}}$, then for any sequence of unit vectors $\boldsymbol{x} = (x_1, \ldots, x_n)$,*

$$\mathop{\mathbb{E}}_{\rho\sim\gamma(\rho_0)} \left[\frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x})\right] \geqslant \exp(-d^2).$$

Proof. For all $\rho \in S_{\mathrm{supp}}$, the eigenvalues of ρ lie within $[0.96/d, 1.04/d]$. Thus, for any unit vector x, $\frac{x^\dagger \rho x}{x^\dagger \rho_0 x} \in [0.96/1.04, 1.04/0.96] \subseteq [0.9, 1.1]$. Using the fact that $\log(1+a) \geqslant a - \frac{2}{3}a^2$ for $|a| \leqslant 0.1$, we have

$$\log \frac{x^\dagger \rho x}{x^\dagger \rho_0 x} \geqslant \frac{x^\dagger(\rho-\rho_0)x}{x^\dagger\rho_0 x} - \frac{2}{3}\left(\frac{x^\dagger(\rho-\rho_0)x}{x^\dagger\rho_0 x}\right)^2 \geqslant \frac{x^\dagger(\rho-\rho_0)x}{x^\dagger\rho_0 x} - d^2(x^\dagger(\rho-\rho_0)x)^2.$$

By symmetry, $\mathbb{E}_{\rho\sim\gamma(\rho_0)}(\rho-\rho_0) = 0$, and by rotational invariance,

$$d^2 \mathop{\mathbb{E}}_{\rho\sim\gamma(\rho_0)} \left[(x^\dagger(\rho-\rho_0)x)^2\right] = \mathop{\mathbb{E}}_{\rho\sim\gamma(\rho_0)} \left[\|\rho-\rho_0\|_F^2\right] \leqslant \mathop{\mathbb{E}}_{\rho\sim\gamma(\rho_0)} \left[\|\rho-\rho_0\|_1 \|\rho-\rho_0\|_{\mathrm{op}}\right] \leqslant \frac{C^2\varepsilon^2}{d}.$$

By Jensen's inequality and the above estimates,

$$\begin{aligned}
\log \mathop{\mathbb{E}}_{\rho\sim\gamma(\rho_0)} \left[\frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x})\right] &\geqslant \sum_{i=1}^n \mathop{\mathbb{E}}_{\rho\sim\gamma(\rho_0)} \log \frac{x_i^\dagger \rho x_i}{x_i^\dagger \rho_0 x_i} \\
&\geqslant \sum_{i=1}^n \mathop{\mathbb{E}}_{\rho\sim\gamma(\rho_0)} \left[\frac{x_i^\dagger(\rho-\rho_0)x_i}{x_i^\dagger\rho_0 x_i} - d^2(x_i^\dagger(\rho-\rho_0)x_i)^2\right] \\
&\geqslant -\frac{C^2 n\varepsilon^2}{d}.
\end{aligned}$$

Finally, because $n \ll d^3/\varepsilon^2$, this is lower bounded by $-d^2$. □

The volume ratio V_2/V_1 is bounded by the following lemma, whose proof we defer to Section 4.3. The proof uses tools from random matrix theory.

Lemma 4.2.10. *We have that $V_2/V_1 \geqslant \exp(-3d^2)$.*

We now put the above claims together to prove Theorem 4.2.6.

Proof of Theorem 4.2.6. Let S_{ρ_0} be the event from Lemma 4.2.7. By the calculation (4.3) and Lemmas 4.2.7 and 4.2.8, for all $\boldsymbol{x} \in S_{\rho_0}$,

$$\frac{\nu_{\boldsymbol{x}}(B(\rho_0, C\varepsilon))}{\nu_{\boldsymbol{x}}(B(\rho_0, \varepsilon))} \geqslant \exp(-5d^2) C^{(d+2)(d-1)/2} \mathop{\mathbb{E}}_{\rho \sim \gamma(\rho_0)} \left[\frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x}) \right] \cdot \frac{V_2}{V_1}.$$

Lemmas 4.2.9 and 4.2.10 bound the remaining factors, giving

$$\frac{\nu_{\boldsymbol{x}}(B(\rho_0, C\varepsilon))}{\nu_{\boldsymbol{x}}(B(\rho_0, \varepsilon))} \geqslant \exp(-9d^2) C^{(d+2)(d-1)/2}.$$

Taking $C = e^{20}$ gives $\frac{\nu_{\boldsymbol{x}}(B(\rho_0, C\varepsilon))}{\nu_{\boldsymbol{x}}(B(\rho_0, \varepsilon))} \gg 1$. Since $\nu_{\boldsymbol{x}}(B(\rho_0, C\varepsilon)) \leqslant 1$, this implies $\nu_{\boldsymbol{x}}(B(\rho_0, \varepsilon)) \ll 1$. □

4.2.4 Main Lower Bound

We can now prove our main lower bound for tomography with incoherent measurements, which we state formally below:

Theorem 4.2.11. *There exist absolute constants $\varepsilon_0 > 0$ and $d_0 \in \mathbb{N}$ such that for any $0 < \varepsilon < \varepsilon_0$ and any integer $d \geqslant d_0$, the following holds. If $n = o(d^3/\varepsilon^2)$, then for any algorithm for state tomography $(\mathcal{T}, \mathcal{A})$ that uses n incoherent, possibly adaptive, measurements, its output $\widehat{\rho}$ upon measuring n copies of ρ satisfies $\|\rho - \widehat{\rho}\|_1 > \varepsilon$ with probability $1 - o(1)$.*

Proof. Let $S \in \sigma(\rho, \boldsymbol{x})$ be the event that $\rho \sim \mu$ lies in S_{good} and $\boldsymbol{x} \sim \mathcal{T}_\rho$ lies in S_ρ. In this proof we will abuse notation and use $\mathcal{A}$ to also denote the internal randomness used by $\mathcal{A}$. It suffices to show $\mathbf{Pr}_{\mathcal{A}, \rho \sim \mu, \boldsymbol{x} \sim \mathcal{T}_\rho}[\|\mathcal{A}(\boldsymbol{x}) - \rho\|_1 \leqslant \varepsilon] = o(1)$.

First note that

$$\mathbf{Pr}_{\mathcal{A}, \rho, \boldsymbol{x}} \left[\|\mathcal{A}(\boldsymbol{x}) - \rho\|_1 \leqslant \varepsilon \right] = \mathop{\mathbb{E}}_{\mathcal{A}, \boldsymbol{x}} \mathop{\mathbb{E}}_{\rho \sim \nu_{\boldsymbol{x}}} \left[\mathbf{1} \|\mathcal{A}(\boldsymbol{x}) - \rho\|_1 \leqslant \varepsilon \right] \tag{4.5}$$

$$\leqslant \mathop{\mathbb{E}}_{\mathcal{A},\boldsymbol{x}} \mathop{\mathbb{E}}_{\rho\sim\nu_{\boldsymbol{x}}} [\mathbf{1}\|\mathcal{A}(\boldsymbol{x}) - \rho\|_1 \leqslant \varepsilon \text{ and } (\rho,\boldsymbol{x}) \in S] + o(1) \quad (4.6)$$

where the second step follows by a union bound and the fact that $\mathbf{Pr}[(\rho,\boldsymbol{x}) \notin S] = e^{-\Omega(d)} + e^{-\Omega(d^2)} = o(1)$ by Theorem 4.2.6.

For any choice of internal randomness for $\mathcal{A}$ and any transcript $\boldsymbol{x}$, let $\rho_{\boldsymbol{x}}^{\mathcal{A}}$ denote an arbitrary state for which $(\rho_{\boldsymbol{x}}^{\mathcal{A}},\boldsymbol{x}) \in S$ and $\|\mathcal{A}(\boldsymbol{x}) - \rho_{\boldsymbol{x}}^{\mathcal{A}}\|_1 \leqslant \varepsilon$, if such a state exists. Denote by $\mathcal{E}$ the event that such a state exists. Then under $\mathcal{E}$, for any state ρ for which $\|\mathcal{A}(\boldsymbol{x}) - \rho\|_1 \leqslant \varepsilon$, we have $\|\rho_{\boldsymbol{x}}^{\mathcal{A}} - \rho\|_1 \leqslant 2\varepsilon$. If $\mathcal{E}$ does not occur for some choice of internal randomness for $\mathcal{A}$ and some $\boldsymbol{x}$, note that the corresponding inner expectation in (4.6) is zero. We can thus upper bound the double expectation in (4.6) by

$$\mathop{\mathbb{E}}_{\mathcal{A},\boldsymbol{x}|\mathcal{E}} \mathop{\mathbb{E}}_{\rho\sim\nu_{\boldsymbol{x}}} \left[\mathbf{1}\|\rho_{\boldsymbol{x}}^{\mathcal{A}} - \rho\|_1 \leqslant 2\varepsilon \text{ and } (\rho,\boldsymbol{x}) \in S\right] \leqslant \mathop{\mathbb{E}}_{\mathcal{A},\boldsymbol{x}|\mathcal{E}} \mathbf{Pr}_{\rho\sim\nu_{\boldsymbol{x}}} \left[\|\rho_{\boldsymbol{x}}^{\mathcal{A}} - \rho\|_1 \leqslant 2\varepsilon\right] = o(1), \quad (4.7)$$

where in the last step we used the fact that under $\mathcal{E}$ we have $(\rho_{\boldsymbol{x}}^{\mathcal{A}},\boldsymbol{x}) \in S$, so by Theorem 4.2.6 the posterior measure $\nu_{\boldsymbol{x}}$ places $o(1)$ mass on the trace norm ε-ball around $\rho_{\boldsymbol{x}}^{\mathcal{A}}$. □

4.3 Lower Bounding the Volume Ratio: Proof of Lemma 4.2.10

In this section, we will prove Lemma 4.2.10, which lower bounds the volume ratio V_2/V_1.

Let

$$V = \left\{ \boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d) \in \mathbb{R}^d : \sum_{i=1}^{d} \lambda_i = 0 \right\}.$$

This is a subspace of $\mathbb{R}^d$ of codimension 1. It inherits the inner product of $\mathbb{R}^d$, which defines a Lebesgue measure Leb_V. Define

$$\Delta = \left\{ \boldsymbol{\lambda} \in V : \lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_d, \sum_{i=1}^{d} |\lambda_i| \leqslant 1 \right\}, \qquad \Delta' = \left\{ \boldsymbol{\lambda} \in \Delta : \max_{i \in [d]} |\lambda_i| \leqslant 1/d \right\},$$

and

$$\Gamma = \left\{ \boldsymbol{\lambda} \in V : \left| \lambda_i - \frac{d - 2i + 1}{d^2} \right| \leqslant \frac{1}{d^4} \right\}.$$

Lemma 4.3.1. *We have* $\Gamma \subseteq \Delta'$.

Proof. If $\boldsymbol{\lambda} \in \Gamma$, then

$$\lambda_i - \lambda_{i+1} \geqslant \frac{2}{d^2} - \frac{2}{d^4} > 0$$

and

$$\sum_{i=1}^{d} |\lambda_i| \leqslant d \cdot \frac{1}{d} = 1,$$

so $\boldsymbol{\lambda} \in \Delta$. Moreover

$$|\lambda_i| \leqslant \frac{d-1}{d^2} + \frac{1}{d^4} \leqslant \frac{1}{d}.$$ □

The volume ratio V_2/V_1 is the probability that if ρ is drawn from the uniform measure on $B(0,1)$ (w.r.t. Leb_{U_0}), then $\|\rho\|_{\mathrm{op}} \leqslant 1/d$. The main random matrix theory fact we will use is [12, Theorem 2.5.2], which implies that if $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$ are the eigenvalues of ρ drawn from this distribution, then $\boldsymbol{\lambda}$ has density (w.r.t. Leb_V) $\frac{1}{Z} f(\boldsymbol{\lambda})$, where Z is a normalizing constant and

$$f(\boldsymbol{\lambda}) = \mathbb{1}\{\boldsymbol{\lambda} \in \Delta\} \prod_{1 \leqslant i < j \leqslant d} |\lambda_i - \lambda_j|.$$

For measurable $A \subseteq V$, let $\mathrm{Vol}(A) = \int_A 1 \mathrm{d}\mathsf{Leb}_V$ denote the volume of A. Then,

$$\frac{V_2}{V_1} = \frac{\int_{\Delta'} f(\boldsymbol{\lambda}) \, \mathrm{d}\mathsf{Leb}_V(\boldsymbol{\lambda})}{\int_{\Delta} f(\boldsymbol{\lambda}) \, \mathrm{d}\mathsf{Leb}_V(\boldsymbol{\lambda})} \geqslant \frac{\int_{\Gamma} f(\boldsymbol{\lambda}) \, \mathrm{d}\mathsf{Leb}_V(\boldsymbol{\lambda})}{\int_{\Delta} f(\boldsymbol{\lambda}) \, \mathrm{d}\mathsf{Leb}_V(\boldsymbol{\lambda})} \geqslant \frac{\mathrm{Vol}(\Gamma)}{\mathrm{Vol}(\Delta)} \cdot \frac{\inf_{\boldsymbol{\lambda} \in \Gamma} f(\boldsymbol{\lambda})}{\sup_{\boldsymbol{\lambda} \in \Delta} f(\boldsymbol{\lambda})}. \tag{4.8}$$

The following three propositions bound the quantities in the right-hand side of (4.8).

Proposition 4.3.2. *For any* $\boldsymbol{\lambda} \in \Gamma$, $f(\boldsymbol{\lambda}) \geqslant 1/((2e)^{d^2/2} d^{d(d-1)/2})$.

Proof. For each $i \in [d]$,

$$\prod_{j \in [d] \setminus \{i\}} |\lambda_i - \lambda_j| \geqslant \prod_{j=1}^{i-1} \left(\frac{2(i-j)}{d^2} - \frac{2}{d^4} \right) \prod_{j=i+1}^{d} \left(\frac{2(j-i)}{d^2} - \frac{2}{d^4} \right) \geqslant \prod_{j=1}^{i-1} \frac{i-j}{d^2} \prod_{j=i+1}^{d} \frac{j-i}{d^2} \geqslant \prod_{j=1}^{d-1} \frac{j}{2d^2}$$
$$\geqslant \frac{(d-1)!}{2^d d^{2(d-1)}} = \frac{d!}{2^d d^{2d-1}} \geqslant \frac{(d/e)^d}{2^d d^{2d-1}} \geqslant \frac{1}{(2e)^d d^{d-1}}.$$

Thus

$$f(\boldsymbol{\lambda}) = \left(\prod_{i=1}^{d} \prod_{j \in [d] \setminus \{i\}} |\lambda_i - \lambda_j| \right)^{1/2} \geqslant \frac{1}{(2e)^{d^2/2} d^{d(d-1)/2}}. \qquad \square$$

Proposition 4.3.3. *For any* $\boldsymbol{\lambda} \in \Delta$, $f(\boldsymbol{\lambda}) \leqslant e^{2d^2}/d^{d(d-1)/2}$.

Proof. Let $(\bar{\lambda}_1, \ldots, \bar{\lambda}_d)$ be the permutation of $(\lambda_1, \ldots, \lambda_d)$ with $|\bar{\lambda}_1| \leqslant \cdots \leqslant |\bar{\lambda}_d|$. For each $i \in [d]$,

$$\prod_{j<i} |\bar{\lambda}_i - \bar{\lambda}_j| \leqslant (2|\bar{\lambda}_i|)^{i-1} \leqslant \left(\frac{e^{2d|\bar{\lambda}_i|}}{d} \right)^{i-1} \leqslant \frac{e^{2d^2|\bar{\lambda}_i|}}{d^{i-1}},$$

so (since $\sum_{i=1}^{d} |\lambda_i| \leqslant 1$)

$$f(\boldsymbol{\lambda}) = \prod_{i=1}^{d} \prod_{j<i} |\bar{\lambda}_i - \bar{\lambda}_j| \leqslant \frac{e^{2d^2 \sum_{i=1}^{d} |\bar{\lambda}_i|}}{d^{d(d-1)/2}} \leqslant \frac{e^{2d^2}}{d^{d(d-1)/2}}. \qquad \square$$

Proposition 4.3.4. *We have* $\mathrm{Vol}(\Delta) \leqslant e^{O(d)}$ *and* $\mathrm{Vol}(\Gamma) \geqslant d^{-O(d)}$.

Proof. The projection of V onto its first $d-1$ coordinates has Jacobian $\Theta(1)$. Because this projection maps Δ injectively to $\mathbb{R}^{d-1}$ and each resulting coordinate is in $[-1, 1]$,

$$\mathrm{Vol}(\Delta) \leqslant \Theta(1) \cdot 2^{d-1} \leqslant e^{O(d)},$$

proving the first conclusion. Note that Γ is the set of $\boldsymbol{\lambda}$ satisfying $\lambda_i = \frac{d-2i+1}{d^2} + \frac{1}{d^4}\mu_i$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ ranges over

$$\Gamma' = \left\{ \boldsymbol{\mu} \in V : \max_{i \in [d]} |\mu_i| \leqslant 1 \right\}.$$

The set Γ' certainly contains the set of $\boldsymbol{\mu}$ where $|\mu_1|, \ldots, |\mu_{d-1}| \leqslant 1/d$ and $\mu_d = -\sum_{i=1}^{d-1} \mu_i$. So,

$$\mathrm{Vol}(\Gamma') \geqslant \Theta(1)(2/d)^{d-1} = d^{-O(d)}.$$

Finally,

$$\mathrm{Vol}(\Gamma) = \mathrm{Vol}(\Gamma') \cdot (d^{-4})^{d-1} = d^{-O(d)},$$

proving the second conclusion. □

Proof of Lemma 4.2.10. By equation (4.8) and the last three propositions,

$$\frac{V_2}{V_1} \geqslant \frac{d^{-O(d)}}{e^{O(d)}} \cdot \frac{(2e)^{-d^2/2}}{e^{2d^2}} \geqslant e^{-3d^2}.$$ □

4.4 Basic Learning Results

Now we focus on the upper bound for learning in fidelity. Recall that for two quantum states ρ, σ, the fidelity between them is $F(\rho, \sigma) = \mathrm{tr}(\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}})^2$. This is the quantum analogue of the *Bhattacharyya coefficient*, which for two classical probability distributions p, q over $[d]$, is defined to be $BC(p, q) = \sum_{i=1}^{d} \sqrt{p_i}\sqrt{q_i}$. Note that if ρ and σ commute, then $F(\rho, \sigma) = BC(p, q)^2$, where p, q are the classical distributions given by the eigenvalues of ρ and σ, respectively. The following inequality is well known:

Fact 4.4.1.

$$BC(p, q) \geqslant 1 - O\left(\chi^2(p\|q)\right).$$

As an immediate corollary of this, we have the following:

Corollary 4.4.2. *Let $\rho \in :^{d \times d}$ be an arbitrary mixed state, and let $\tilde{\rho} = (1-\gamma)\rho + \gamma \cdot \frac{1}{d} I$. Then $F(\rho, \tilde{\rho}) \geqslant 1 - O(\gamma)$.*

Proof. Let p and $\tilde{p}$ denote the the distributions given by the eigenvalues of ρ and $\tilde{\rho}$, respectively. Since ρ and $\tilde{\rho}$ commute, it suffices to show that $BC(p, \tilde{p}) \geqslant 1 - O(\gamma)$. However, we have:

$$\chi^2(\tilde{p} \| p) = \sum_{i=1}^{d} \frac{(\gamma p_i + \gamma/d)^2}{\tilde{p}_i} \leqslant 2 \sum_{i=1}^{d} \frac{\gamma^2 p_i^2}{\tilde{p}_i} + 2 \sum_{i=1}^{d} \frac{\gamma^2}{d^2 \tilde{p}_i} \leqslant O(\gamma) \ ,$$

since $\tilde{p}_i \geqslant \gamma/d$ for all i. □

It is well-known that infidelity is not a metric on mixed states. However, the associated Bures metric, defined as

$$D_B(\rho, \sigma)^2 = 2 \left(1 - \sqrt{F(\rho, \sigma)} \right) \ ,$$

does satisfy the triangle inequality and is hence a valid metric. As an immediate consequence of this, we obtain that infidelity still satisfies a weak version of the triangle inequality:

Corollary 4.4.3. *Let k be a positive integer, and let $\gamma \ll 1/k^2$ be sufficiently small. Let $\rho_1, \ldots, \rho_k$ be a sequence of mixed states satisfying $1 - F(\rho_i, \rho_{i+1}) \leqslant \gamma$ for all $i = 1, \ldots, k-1$. Then*

$$1 - F(\rho_1, \rho_k) \leqslant O(k^2 \gamma) \ .$$

Proof. By repeated application of the Taylor series expansion of the square root function around 1, we know that $D_B(\rho_i, \rho_{i+1}) \leqslant O(\sqrt{\gamma})$ for all $i = 1, \ldots, k-1$. Therefore by the triangle inequality, we have that $D_B(\rho_1, \rho_k) \leqslant O(k\sqrt{\gamma})$, from which the claim immediately follows. □

4.4.1 Learning a State in Spectral Norm

Finally, we require the following guarantee for one of the standard estimators for a mixed state based on unentangled measurements. Given n copies of a mixed state ρ, we measure each one with the uniform POVM over the sphere $\{d\,|v\rangle\,\langle v|\,dv\}$ where v ranges over the unit sphere, and let $|v_i\rangle$ denote the outcome of the ith measurement. We consider the estimator $H_n(\rho) = H_n(\rho, v_1, \ldots, v_n)$, defined as

$$H_n(\rho) = \frac{1}{n}\sum_{i=1}^{n}\left((d+1)\,|v_i\rangle\,\langle v_i| - I\right)\ . \tag{4.9}$$

We show the following rate for this estimator. The same rate is claimed in the proof sketch of Theorem 2 in [101], but to our knowledge no full proof of this exists in the literature. We include a full proof for completeness:

Theorem 4.4.4. *There exists a universal constant C so that for all n, we have that*

$$\|H_n(\rho) - \rho\|_{\mathrm{op}} \leqslant C \cdot \max\left(\frac{d+\log 1/\delta}{n}, \sqrt{\frac{d+\log 1/\delta}{n}}\right)\ ,$$

with probability $1-\delta$.

The key concentration lemma we require is the following:

Lemma 4.4.5. *Let $|v\rangle$ be the outcome of measuring ρ using the uniform POVM. Then, for any fixed pure state $|u\rangle$, and for all $k \geqslant 1$, we have*

$$\mathbb{E}\Big[(d+1)^k\,\langle u|v\rangle^{2k}\Big] \leqslant (k+1)^{k+1}\ .$$

Proof. First consider the case where $\rho = |w\rangle\,\langle w|$ is a pure state. For any t, let Π_t denote projection onto the t-fold symmetric subspace. Then, we have:

$$\mathbb{E}\Big[\langle u|v\rangle^{2k}\Big] = d\cdot\int \langle u|v\rangle^{2k}\,\langle v|w\rangle^2\,dv \tag{4.10}$$

$$= d\cdot\langle u|^{\otimes k}\otimes\langle w|\left(\int |v\rangle^{\otimes(k+1)}\,\langle v|^{\otimes(k+1)}\,dv\right)|u\rangle^{\otimes k}\otimes|w\rangle \tag{4.11}$$

$$= d \cdot \binom{k+d}{k+1}^{-1} \cdot \left(\langle u|^{\otimes k} \otimes \langle w| \right) \Pi_{k+1} \left(|u\rangle^{\otimes k} \otimes |w\rangle \right) \leqslant d \cdot \binom{k+d}{k+1}^{-1} , \tag{4.12}$$

where the third step follows by the standard Haar integral formulation of Π_t [107]. Therefore, we have that

$$\mathbb{E}\Big[(d+1)^k \langle u|v\rangle^{2k} \Big] \leqslant (d+1)^{k+1} \cdot \binom{k+d}{k+1}^{-1} \leqslant (d+1)^{k+1} \cdot \frac{(k+1)^{k+1}}{(d+k)^{k+1}} \leqslant (k+1)^{k+1} , \tag{4.13}$$

as claimed. The claim for general ρ directly follows by convexity. □

Proof of Theorem 4.4.4. The proof proceeds via the same general strategy as in [101]. Let $\mathcal{N}$ be a 1/3-net of all pure states in $:^d$. For any $u \in \mathcal{N}$, Lemma 4.4.5 implies that the random variable

$$\langle u| \left(\rho - H_n(\rho) \right) |u\rangle = \frac{1}{n} \sum_{i=1}^{n} \left((d+1) \langle u, v_i\rangle^2 - 1 - \langle u| \rho |u\rangle \right)$$

is a sum of n independent $O(1)$-subexponential random variables. Therefore, by standard net arguments, we have that for all $\gamma > 0$,

$$\mathbf{Pr}[\|\rho - H_n(\rho)\|_{\text{op}} > \gamma] \leqslant \exp\left(c_1 d - c_2 n \max(\gamma, \gamma^2) \right) ,$$

for some universal constants c_1, c_2, which is equivalent to what we wanted to show. □

Finally, we also require the following generalization of the estimator we considered above:

Definition 4.4.6. Given a projection matrix $\Pi \in :^{d \times d}$ onto an r-dimensonal subspace, the *projected estimator* on the subspace Π, denoted $H_n(\rho, \Pi)$ is defined as follows. Let $\mathcal{M}_\Pi = \{ r |v\rangle \langle v| \, dv \}$ denote the uniform POVM over the subspace spanned by Π, and consider the POVM over $:^{d \times d}$ defined by $\mathcal{M} = \{ I - \Pi \} \cup \mathcal{M}_\Pi$. Given n copies of ρ, let

$|v_i\rangle$ denote the outcomes of measuring each copy of ρ independently with $\mathcal{M}$, where we say $v_i = \perp$ if the outcome is $I - \Pi$. Then, we define

$$H_n(\rho, \Pi) = \frac{1}{n} \sum_{v_i \neq \perp} \left((r+1) |v_i\rangle \langle v_i| - I \right) \ .$$

For projected estimators we have the following rate, which follows immediately from Theorem 4.4.4 and standard Chernoff bounds:

Lemma 4.4.7. *Let $\rho \in :^{d \times d}$ and let Π be a projection onto an r-dimensional subspace. Let $\alpha = \mathrm{tr}(\Pi \rho \Pi)$. Then, there exists a universal constant C so that*

$$\|H_n(\rho, \Pi) - \Pi \rho \Pi\|_{\mathrm{op}} \leqslant C \cdot \max \left(\frac{d + \log 1/\delta}{n}, \sqrt{\frac{\alpha(d + \log 1/\delta)}{n}} \right) \ .$$

with probability $1 - \delta$.

4.5 Learning a State in Fidelity

In this section, we present our algorithm for tomography in fidelity. Our main theorem is stated below.

Theorem 4.5.1. *Let $\rho \in :^{d \times d}$ be an unknown rank-r mixed state. Given $n = \widetilde{O}(dr^2 \log(1/\delta)/\gamma)$ copies of ρ, there is an algorithm that uses incoherent measurements and with probability $1 - \delta$, outputs a state $\widehat{\rho}$ such that $F(\rho, \widehat{\rho}) \geqslant 1 - \gamma$.*

Let $\eta > 0$ be a parameter to be determined later, and let t be an integer satisfying $t \leqslant \log_2 r/\gamma + 4$. We first describe our algorithm. Divide the samples into t groups of n/t samples each.

We will compute a sequence of orthogonal projections $\Pi_0 = 0, \Pi_1, \ldots, \Pi_t$. To find Π_j given $\Pi_0, \ldots, \Pi_{j-1}$, the algorithm proceed as follows: it forms $\Gamma_j = I - \sum_{i=0}^{j-1} \Pi_i$, and using a fresh batch of samples, it computes $\sigma_j = H_{n/t}(\rho, \Gamma_j)$, and it sets Π_j to be the span of the nonzero eigenvectors of σ_j with corresponding eigenvalue at least 2^{-j}.

When it finishes, it outputs the state

$$M_n(\rho) \stackrel{\text{def}}{=} \frac{\widehat{\sigma}}{\operatorname{tr}(\widehat{\sigma})} \ , \text{ where } \widehat{\sigma} \stackrel{\text{def}}{=} \sum_{i=1}^{t} \Pi_i \sigma_i \Pi_i \ , \tag{4.14}$$

To analyze the algorithm, we make the following definitions. Let $\Pi_{t+1} = \Gamma_{t+1}$, and define the state

$$\rho_{\text{diag}} = \sum_{i=1}^{t+1} \Pi_i \rho \Pi_i \ . \tag{4.15}$$

For all $i = 1, \ldots, t+1$, let $B_i \in :^{d \times \operatorname{rank}(\Pi_i)}$ be any matrix with orthonormal columns so that $B_i B_i^\top = \Pi_i$. Similarly, for $i = 1, \ldots, t+1$, let $C_i \in :^{d \times \operatorname{rank}(\Gamma_i)}$ be any matrix with orthonormal columns so that $C_i C_i^\top = \Gamma_i$.

Notice that by construction $\sum_{i=1}^{t+1} \Pi_i = I$, so this is indeed a valid mixed state. Furthermore, note that ρ_{diag} is block-diagonal with respect to the matrices $\Pi_1, \ldots, \Pi_{t+1}$, that is, after a suitable rotation which sends each B_i to itself, ρ_{diag} has the form

$$\rho_{\text{diag}} = \begin{pmatrix} \rho_1 & & & \\ & \rho_2 & & \\ & & \ddots & \\ & & & \rho_{t+1} \end{pmatrix} \ ,$$

where we let ρ_i denote the projection of $\Pi_i \rho \Pi_i$ onto its nonzero eigenvectors. Note that in this basis, ρ is not block diagonal, and indeed, much of the work in the proof is to bound the contribution of the error because of these off-diagonal blocks. To this end, for all $i = 1, \ldots, t$, let $E_i = B_i^\top \rho C_{i+1}$, so that

$$\rho = \begin{pmatrix} \rho_1 & E_1 & & \\ E_1^\top & \rho_2 & E_2 & \\ & E_2^\top & \ddots & \\ & & & \rho_t \end{pmatrix} \ ,$$

i.e., the E_i are the off-diagonal components of ρ.

Next, since each σ_j is computed with fresh samples, by Lemma 4.4.7 and a union bound, we have that

$$\|\sigma_j - \Gamma_j \rho \Gamma_j\|_{\mathrm{op}} \leqslant \gamma_j \stackrel{\mathrm{def}}{=} C \cdot \max\left(\frac{d + \log t/\delta}{n}, \sqrt{\frac{\alpha_j(d + \log t/\delta)}{n}}\right), \tag{4.16}$$

for all $j = 1, \ldots, t$ simultaneously, with probability at least $1-\delta$, where $\alpha_j = \mathrm{tr}(\Pi_j \rho \Pi_j)$.

For the remainder of the section, let us take $n \gtrsim \frac{(d+\log t/\delta)r^2}{\gamma}$, and we will assume that (4.16) holds. We first show a few basic inequalities. We have the following estimate on the RHS of (4.16).

Lemma 4.5.2. *Let $n \gtrsim \frac{(d+\log t/\delta)r^2}{\gamma}$. Then, we have that $\gamma_j \lesssim 2^{-(j+t)/2}$.*

Proof. By our choice of n, we have that:

$$\frac{d + \log t/\delta}{n} \lesssim \frac{\gamma}{r^2} \leqslant 2^{-(j+t)/2},$$

and

$$\begin{aligned}\sqrt{\frac{\alpha_{j-1}(d + \log t/\delta)}{n}} &= \sqrt{\mathrm{tr}(\Gamma_{j-1}\rho\Gamma_{j-1}) \cdot \frac{d + \log t/\delta}{n}} \\ &\lesssim \sqrt{2^{-(j-1)} \cdot \frac{\gamma}{r}} \leqslant 2^{-(j+t)/2},\end{aligned}$$

where in both cases we use that $t = \log_2 r/\gamma + 4$. □

With this, we can show the following set of useful inequalities:

Lemma 4.5.3. *For all $j = 1, \ldots, t$, we have that*

(i) $\|\Gamma_j \rho \Gamma_j\|_{\mathrm{op}} \leqslant 2^{-(j-1)}$,

(ii) $\|\sigma_j\|_{\mathrm{op}} \leqslant 2^{-(j-2)}$, *and*

(iii) $B_i^\top \rho B_i \succeq 2^{-j-1} I$, *and*

(iv) $\operatorname{tr}(\Gamma_{t+1}\rho\Gamma_{t+1}) \leqslant \gamma/2$.

Proof. To prove the first bullet point, we proceed by induction. The case where $j = 1$ is trivial. Now, suppose the claim holds for some $j - 1$. By (4.16), we have that

$$\|\sigma_{j-1} - \Gamma_{j-1}\rho\Gamma_{j-1}\|_{\mathrm{op}} \leqslant \gamma_{j-1} \ .$$

Now Γ_j is defined to be the projection onto the eigenvectors of σ_{j-1} with eigenvalue less than 2^{-j}. Therefore, $\|\Gamma_j\rho\Gamma_j\|_{\mathrm{op}} \leqslant \gamma_{j-1} + 2^{-j}$. The second and third claims then both follow from Lemma 4.5.2, and by the definition of M_j. Finally, the last claim follows because ρ has at most r nonzero eigenvalues, and therefore so does $\operatorname{tr}(\Gamma_{t+1}\rho\Gamma_{t+1})$; moreover by the above, each one is at most $2^{-t+2} \leqslant \gamma/(2r)$. □

One simple but important consequence of these inequalities is that the subspaces Π_i are low dimensional:

Corollary 4.5.4. *For all i, we have* $\operatorname{rank}(\Pi_i) \leqslant r$.

Proof. Note that ρ has r nonzero eigenvalues by assumption, so by Lemma 4.5.3, σ_i can only have r eigenvalues which are at least 2^{-i-1}. □

We will show the following two key estimates. Roughly, Lemma 4.5.5 bounds the contribution to infidelity form the error on the diagonal blocks of ρ and ρ_{diag}. Lemma 4.5.6 bounds the contribution to infidelity from the off-diagonal blocks in ρ. Putting them together with the (weak) triangle inequality in Corollary 4.4.3 will immediately imply Theorem 4.5.1.

Lemma 4.5.5. *Let $\rho, M_n(\rho)$, and ρ_{diag} be as above, and assume that* (4.16) *holds. Then, we have that*

$$F(M_n(\rho), \rho_{\mathrm{diag}}) \geqslant 1 - \gamma t$$

Lemma 4.5.6. *Let $\rho, M_n(\rho)$, and ρ_{diag} be as above, and assume that* (4.16) *holds. Then, we have that*

$$F(\rho, \rho_{\text{diag}}) \geqslant 1 - \gamma \cdot \min(t, d/r) \ .$$

Proof of Theorem 4.5.1. Combining Lemma 4.5.5, Lemma 4.5.6 and Corollary 4.4.3, we get

$$F(M_n(\rho).\rho) \geqslant 1 - O(\gamma \min(t, d/r))$$

and since $t \leqslant \log_2(r/\gamma) + 4$, we can redefine γ appropriately (scaling down by a logarithmic factor) to complete the proof. □

4.5.1 Proof of Lemma 4.5.5

We first require the following fact, which is a direct consequence of Proposition 3.1 in [131], and direct calculation:

Theorem 4.5.7. *Let $f(X) ::_{\geqslant 0}^{d\times d} \to \mathbb{R}$ be defined by $f(X) = \operatorname{tr}(\sqrt{X})$. Let $A \in :^{d\times d}$ be a non-degenerate Hermitian matrix. Then, for all symmetric matrices $B \in :^{d\times d}$, we have that*

$$\nabla^2 f(A)[B, B] \geqslant -\frac{1}{4} \operatorname{tr}\left(BA^{-3/2}B\right) \tag{4.17}$$

where $\nabla^2 f(A)$ denotes the Hessian of $f(A)$ (where we treat the entries of A as variables) and $\nabla^2 f(A)[B, B]$ denotes taking the quadratic form of this Hessian at B.

Proof of Lemma 4.5.5. For conciseness, throughout this proof we will let $M = M_n(\rho)$. Let $\Delta = M - \rho_{\text{diag}}$. By Taylor's theorem, we know that there is some $\beta \in [0, 1]$ so that

$$\begin{aligned}
&F(M_n(\rho), \rho_{\text{diag}}) \\
&= \operatorname{tr}\left(\sqrt{M^2 - M^{1/2}\Delta M^{1/2}}\right) \\
&= 1 + \frac{1}{2}\operatorname{tr}\left(M^{-1}M^{1/2}\Delta M^{1/2}\right) + D^2 f\left(M^2 - \beta M^{1/2}\Delta M^{1/2}\right)\left[M^{1/2}\Delta M^{1/2}, M^{1/2}\Delta M^{1/2}\right] \\
&\geqslant 1 - \frac{1}{4}\operatorname{tr}\left((M^{1/2}\Delta M^{-1/4}(M + \beta\Delta)^{-3/2}M^{-1/4}\Delta M^{1/2}\right) \ ,
\end{aligned}$$

where in the last line we used Theorem 4.5.7.

Notice that both M and Δ are block diagonal, and moreover, by construction, they share the same block structure. Thus, if we let $M_i = \Pi_i M \Pi_i$ and $\Delta_i = \Pi_i \Delta \Pi_i$, then we have that

$$\begin{aligned} &\operatorname{tr}\left((M^{1/2}\Delta M^{-1/4}(M+\beta\Delta)^{-3/2}M^{-1/4}\Delta M^{1/2}\right) \\ &= \sum_{i=1}^{t} \operatorname{tr}\left(M_i^{1/2}\Delta_i M_i^{-1/4}(M_i+\beta\Delta_i)^{-3/2}M_i^{-1/4}\Delta_i M_i^{1/2}\right) , \end{aligned}$$

where in a slight abuse of notation, we let M_i^{-1} and $(M_i + \beta\Delta)^{-1}$ denote the pseudoinverses of the associated matrices. For any i, we have that

$$\begin{aligned} &\operatorname{tr}\left(M_i^{1/2}\Delta_i M_i^{-1/4}(M_i+\beta\Delta_i)^{-3/2}M_i^{-1/4}\Delta_i M_i^{1/2}\right) \\ &\leqslant \operatorname{tr}(M_i)\cdot\|\Delta_i M_i^{-1/4}(M_i+\beta\Delta_i)^{-3/2}M_i^{-1/4}\Delta_i\|_{\mathrm{op}} \\ &\leqslant C\cdot\operatorname{tr}(M_i)\cdot\gamma_i^2\cdot\|M_i^{-1/4}(M_i+\beta\Delta_i)^{-3/2}M_i^{-1/4}\|_{\mathrm{op}} , \end{aligned}$$

where the last step follows because $\|\Delta_i\|_{\mathrm{op}} \leqslant \gamma_i$ by Lemma 4.4.7 and (4.16). By construction, we have that all nonzero eigenvalues of M_i are at least 2^{-i}, so all nonzero eigenvalues of $M_i + \beta\Delta_i$ are at least $2^{-(i+1)}$. Putting it all together, we obtain that

$$\operatorname{tr}\left(M_i^{1/2}\Delta_i M_i^{-1/4}(M_i+\beta\Delta_i)^{-3/2}M_i^{-1/4}\Delta_i M_i^{1/2}\right) \lesssim \operatorname{tr}(M_i)\cdot\gamma_i^2\cdot 2^{2i} \leqslant r\cdot 2^i\gamma_i^2 \leqslant \gamma ,$$

as claimed. □

4.5.2 Proof of Lemma 4.5.6

First, observe that we may disregard the subspace spanned by Γ_{t+1}:

Lemma 4.5.8. *Let $\overline{\Gamma} = I - \Gamma_{t+1}$. We have that $\operatorname{tr}(\overline{\Gamma}\rho\overline{\Gamma}) = \operatorname{tr}(\overline{\Gamma}\rho_{\mathrm{diag}}\overline{\Gamma}) \stackrel{def}{=} c$, and moreover, if we let $\tilde{\rho} = \frac{1}{c}\overline{\Gamma}\rho\overline{\Gamma}$ and $\tilde{\rho}_{\mathrm{diag}} = \frac{1}{c}\overline{\Gamma}\rho_{\mathrm{diag}}\overline{\Gamma}$, then $F(\tilde{\rho},\rho) \leqslant 1-\gamma/2$ and $F(\tilde{\rho}_{\mathrm{diag}},\rho_{\mathrm{diag}}) \leqslant 1-\gamma/2$.*

Proof. The first claim follows since Π_i and $\overline{\Gamma}$ all commute. The fact that $F(\tilde{\rho}, \rho) \leqslant 1 - \gamma/2$ immediately follows from (iv) in Lemma 4.5.3, and the last claim follows since the same lemma implies that $\mathrm{tr}(\Gamma_{t+1}\rho_{\mathrm{diag}}\Gamma_{t+1}) \leqslant \gamma/2$ as well. □

In light of this, for the rest of the proof, we will assume that $\overline{\Gamma} = I$; in particular, this implies that ρ and ρ_{diag} both have minimum eigenvalue at least 2^{-j-1} by Lemma 4.5.3. The above lemma implies that this incurs at most an additional additive γ to the overall fidelity calculation. We now establish the following bound on the matrices E_i:

Lemma 4.5.9. *Let $n \gtrsim \frac{(d+\log T/\delta)r^2}{\gamma}$, and assume that (4.16) holds. Then, for all $i = 1, \ldots, t$, we have that $\|E_i\|_{\mathrm{op}} \leqslant 2\gamma_j$.*

Proof. For any i, we have that

$$\begin{aligned}\|E_i\|_{\mathrm{op}} &= \|\Gamma_j \left(\rho - \rho_{\mathrm{diag}}\right) \Gamma_j\|_{\mathrm{op}} \\ &\leqslant \|\Gamma_j \rho \Gamma_j - \sigma_j\|_{\mathrm{op}} + \|\Gamma_j \rho_{\mathrm{diag}} \Gamma_j - \sigma_j\|_{\mathrm{op}} \\ &\leqslant 2\|\Gamma_j \rho \Gamma_j - \sigma_j\|_{\mathrm{op}} = 2\gamma_j \ ,\end{aligned}$$

where the third inequality follows because $\Gamma_j \rho_{\mathrm{diag}} \Gamma_j$ is the projection of $\Gamma_j \rho \Gamma_j$ onto a basis which commutes with σ_j. □

To prove our overall claim we will proceed via a hybrid argument. For all $j = 0, \ldots, t$, let

$$\rho^{(j)} = \sum_{i=1}^{j} \Pi_i \rho \Pi_i + \Gamma_{j+1} \rho \Gamma_{j+1} \ .$$

Note that by design, we have that $\rho^{(0)} = \rho$ and $\rho^{(t)} = \rho_{\mathrm{diag}}$. For these matrices, we show:

Lemma 4.5.10. *For all $j = 1, \ldots, t$, we have that*

$$F(\rho^{(j-1)}, \rho^{(j)}) \geqslant 1 - 2^{-t-1} \min(rt, d) \ .$$

To prove Lemma 4.5.10, we rely on the following fact that bounds the contribution from off-diagonal perturbations to infidelity between two states.

Lemma 4.5.11. *Let $0 < c_2, c_1 < 1$ satisfy $c_2 \leqslant c_1/10$, and let $M \in :^{d_1 \times d_1}$ and $N \in :^{d_2 \times d_2}$ be PSD matrices satisfying $c_1 I \preceq M \preceq 4c_1 I$ and $c_2 I \preceq N \preceq 2c_1 I$. Let $E \in :^{d_1 \times d_2}$ satisfy $\|E\|_{\mathrm{op}} \leqslant \eta$, and define the matrices*

$$A = \begin{pmatrix} M & E \\ E^\top & N \end{pmatrix} \text{ , and } A_{\mathrm{diag}} = \begin{pmatrix} M & 0 \\ 0 & N \end{pmatrix} . \tag{4.18}$$

Suppose further that $A \succeq 0$, and that $\eta \leqslant \sqrt{c_1 c_2}$. Then

$$\mathrm{tr}\left(\sqrt{A_{\mathrm{diag}}^{1/2} A A_{\mathrm{diag}}^{1/2}}\right) \geqslant \mathrm{tr}(A) - c_2(d_1 + d_2) . \tag{4.19}$$

Proof. For any δ, let $M_\delta = M - \delta I$ and similarly let $N_\delta = N - \delta I$. we will explicitly construct a matrix B which will be a PSD lower bound for the matrix $A_{\mathrm{diag}}^{1/2} A A_{\mathrm{diag}}^{1/2}$. Our guess will have the form

$$B = \begin{pmatrix} M_{c_2} & X \\ X^\top & N_{c_2} \end{pmatrix} , \tag{4.20}$$

for some matrix X we define shortly. Note that if we can show that $A_{\mathrm{diag}}^{1/2} A A_{\mathrm{diag}}^{1/2} \succeq B^2$, we are done, since by operator monotonicity of the matrix square root, we have that

$$\mathrm{tr}\left(\sqrt{A_{\mathrm{diag}}^{1/2} A A_{\mathrm{diag}}^{1/2}}\right) \geqslant \mathrm{tr}(B) \geqslant \mathrm{tr}(A) + \mathrm{tr}(M) - c_2(d_1 + d_2) = \mathrm{tr}(A) - c_2(d_1 + d_2) ,$$

as claimed.

It remains to demonstrate how to construct such an X. We choose the following matrix:

$$X \stackrel{\mathrm{def}}{=} \sum_{i=1}^{\infty} M_{(c_2 - c_1)}^{-i} \left(M^{1/2} E N^{1/2}\right) N_{(c_2 + c_1)}^{i-1} . \tag{4.21}$$

We first note that this sum is convergent, since $\|M_{c_2 - c_1}^{-i}\| \leqslant \frac{1}{(1.9 c_1)^i}$ and $\|N_{c_2 + c_1}^{i}\|_{\mathrm{op}} \leqslant$

$(1.1c_1)^i$. Next, we note that

$$M_{c_2}X + XN_{c_2} = M_{(c_2-c_1)}X + XN_{(c_2+c_1)} = M^{1/2}EN^{1/2} \, .$$

Therefore, we have that

$$B^2 = \begin{pmatrix} M_{c_2}^2 + XX^\top & M_{c_2}X + XN_{c_2} \\ (M_{c_2}X + XN_{c_2})^\top & X^\top X + N_{c_2}^2 \end{pmatrix} = \begin{pmatrix} M_{c_2}^2 + XX^\top & M^{1/2}EN^{1/2} \\ \left(M^{1/2}EN^{1/2}\right)^\top & X^\top X + N_{c_2}^2 \end{pmatrix} , \tag{4.22}$$

At the same time, we have that

$$A_{\text{diag}}^{1/2} A A_{\text{diag}}^{1/2} = \begin{pmatrix} M^2 & M^{1/2}EN \\ \left(M^{1/2}EN^{1/2}\right)^\top & N^2 \end{pmatrix} , \tag{4.23}$$

so in particular the off-diagonal block of B^2 exactly matches that of $A_{\text{diag}}^{1/2} A A_{\text{diag}}^{1/2}$. Therefore,

$$A_{\text{diag}}^{1/2} A A_{\text{diag}}^{1/2} - B^2 = \begin{pmatrix} M^2 - M_{c_2}^2 - XX^\top & \\ & N^2 - N_{c_2}^2 - X^\top X \end{pmatrix} \tag{4.24}$$

$$= \begin{pmatrix} 2c_2 M - c_2^2 I - XX^\top & \\ & 2c_2 N - c_2^2 I - X^\top X \end{pmatrix} \tag{4.25}$$

Therefore for our candidate to be a valid PSD lower bound, it suffices to show that $c_2^2 I + XX^\top \preceq 2c_2 M$ and similarly $c_2^2 I + X^\top X \preceq 2c_2 N$. Note that for both of these inequalities to be satisfied, it suffices to show that $X^\top X \preceq c_2 N$.

Define the matrix

$$Q = \sum_{i=1}^{\infty} M_{(c_2-c_1)}^{-i} \left(M^{1/2}E\right) N_{(c_2+c_1)}^{i-1} \, . \tag{4.26}$$

Since N and N_δ commute for all δ, we have that

$$X^\top X = N^{1/2} Q^\top Q N^{1/2} \, .$$

Additionally, we have that

$$\|Q\|_{\text{op}} = \left\| \sum_{i=1}^{\infty} M_{(c_2-c_1)}^{-i} \left(M^{1/2}E\right) N_{(c_2+c_1)}^{i-1} \right\|_{\text{op}} \leqslant \sum_{i=1}^{\infty} \left\| M_{(c_2-c_1)}^{-i} \left(M^{1/2}E\right) N_{(c_2+c_1)}^{i-1} \right\|_{\text{op}} \tag{4.27}$$

$$\leqslant \frac{1.9}{1.1c_1} \cdot \|M^{1/2}E\|_{\text{op}} \sum_{i=1}^{\infty} \left(\frac{1.1}{1.9}\right)^i \tag{4.28}$$

$$\lesssim \frac{\eta}{\sqrt{c_1}} \leqslant \sqrt{c_2}\,, \tag{4.29}$$

by assumption. Therefore, we have that $X^\top X \preceq c_2 N$, so B is indeed a valid PSD lower bound on $A_{\text{diag}}^{1/2} A A_{\text{diag}}^{1/2}$. □

Proof of Lemma 4.5.10. Fix some $j \in \{1, \ldots, t\}$. Let $B = C_{j+1}^\top \rho C_{j+1}$, and define the matrices

$$A \stackrel{\text{def}}{=} \begin{pmatrix} \rho_j & E_j \\ E_j^\top & B \end{pmatrix}, \text{ and } A_{\text{diag}} = \begin{pmatrix} \rho_j & \\ & B \end{pmatrix}.$$

Then

$$F(\rho^{(j-1)}, \rho^{(j)}) = F\left(\left(\begin{pmatrix} \tau & \\ & A \end{pmatrix}, \begin{pmatrix} \tau & \\ & A_{\text{diag}} \end{pmatrix} \right) \right),$$

for some shared, unnormalized state τ. Recall that by Lemma 4.5.2, we have that $\rho_j \succeq 2^{-j-1} I$, and that $2^{-t-1} I \preceq B \preceq 2^{-j+1} I$. Additionally, we have that $\|E_j\|_{\text{op}} \leqslant 2\gamma_j$. Therefore the matrices A and A_{diag} satisfy the conditions of Lemma 4.5.11 with parameters $c_1 = 2^{-j-1}$, $c_2 = 2^{-t-1}$, and $\eta = 2\gamma_j$. Note that Lemma 4.5.2 immediately implies that $\eta \leqslant \sqrt{c_1 c_2}$, as necessary. Therefore we have that

$$F\left(\left(\begin{pmatrix} \tau & \\ & A \end{pmatrix}, \begin{pmatrix} \tau & \\ & A_{\text{diag}} \end{pmatrix} \right) \right) = \text{tr}(\tau) + \text{tr}\left(\sqrt{A_{\text{diag}}^{1/2} A A_{\text{diag}}^{1/2}} \right) \tag{4.30}$$

$$\geqslant \text{tr}(\tau) + \text{tr}(A) - 2^{-t-1} \left(\sum_{i=1}^{t} \text{rank}(\Pi_i) \right) \tag{4.31}$$

$$\geqslant 1 - 2^{-t-1} \min(rt, d)\,, \tag{4.32}$$

as claimed. □

We can now finish the proof of Lemma 4.5.6.

Proof of Lemma 4.5.6. The desired statement follows by combining Lemma 4.5.8, Lemma 4.5.10 and Corollary 4.4.3 and redefining γ appropriately (scaling down by a logarithmic factor). □

Chapter 5

State Certification

In the previous chapter, we studied quantum state tomography. In this chapter, we will study another fundamental, and closely related task, *state certification*. In quantum state certification, the learner is given n copies of a mixed state $\rho \in \mathbb{C}^{d \times d}$, and an explicit description of a mixed state $\sigma \in \mathbb{C}^{d \times d}$, and the objective is to distinguish with, say at least $2/3$ probability, between the case where $\rho = \sigma$ or if it is ε-far from σ in trace distance. An important special case of state certification is the case when $\sigma = I_d/d$, which is called *mixedness testing*.

In addition to being a core primitive in the verification of quantum experiments, state certification is also fundamental from a mathematical perspective. While state tomography is the most basic quantum learning problem, state certification is the natural property testing variant. Mixedness testing and state certification are the natural quantum analogues of uniformity testing and identity testing, respectively, two of the most well-studied problems in classical distribution testing.

Similar to quantum state tomography, it was not until relatively recently that the copy complexity of state certification and mixedness testing were first understood in the coherent setting. The seminal paper of [169] first demonstrated that $n = \Theta(d/\varepsilon^2)$ copies were necessary and sufficient to solve mixedness testing. Follow-up work of [27]

later demonstrated that $n = O(d/\varepsilon^2)$ is also sufficient for the more general problem of state certification. Combined with the lower bound for mixedness testing, this resolved the copy complexity of state certification, in the worst case over σ.

However, as was the case for tomography, due to the difficulty of implementing coherent measurements over many copies, there has been a considerable amount of attention in recent years devoted to understanding the statistical power of algorithms that only use incoherent measurements, which was also posed as an open problem in Wright's thesis [215]. A recent work of [54] demonstrated that if the measurements are additionally chosen non-adaptively, then $n = \Theta(d^{3/2}/\varepsilon^2)$ copies are necessary and sufficient to solve mixedness testing. They also demonstrated that *any* algorithm using incoherent measurements—even those chosen adaptively—must use at least $n = \Omega(d^{4/3}/\varepsilon^2)$ copies. Nevertheless, this still leaves a polynomial gap between the upper and lower bounds, even for the special case of mixedness testing.

In this chapter, we will essentially resolve the complexity of state certification with incoherent measurements, building upon the framework developed in the previous chapter for dealing with the challenge of adaptivity.

5.1 Results

Our first result is to fully resolve the copy complexity for mixedness testing: we prove that adaptivity does not improve the sample complexity at all, except possibly up to constant factors.

Theorem 5.1.1 (Informal, see Theorem 5.3.1)**.** *The copy complexity of mixedness testing using incoherent measurements is* $n = \Theta(d^{3/2}/\varepsilon^2)$.

By completely pinning down the copy complexity of mixedness testing with incoherent measurements, this answers open questions of [215] and [54].

Instance-optimal lower bounds for state certification. We next turn to state certification. Because mixedness testing is a special case of state certification, Theorem 5.1.1 immediately implies that $n = \Omega(d^{3/2}/\varepsilon^2)$ copies are necessary for state certification, in the worst case over all choices of the reference state σ. This, coupled with a matching upper bound from [65, Lemma 6.2], resolves the copy complexity of state certification with incoherent measurements for worst-case σ.

However, it should be clear that this bound is not the correct bound for all possible σ. For instance, when σ is pure, it is not hard to see that $\Theta(1/\varepsilon^2)$ copies are sufficient and necessary. This raises the natural question: what is the copy complexity of state certification with incoherent measurements, as a function of the reference state σ? This is the quantum analogue of the (classical) distribution testing problem of obtaining instance optimal bounds for identity testing against a known distribution over d elements [2, 3, 204, 75, 48, 130]. In the classical version of the problem, there is a known distribution p over $\{1, \dots, d\}$, and we are given samples from a distribution q. We are asked to distinguish between the case where $p = q$, and the case when $\|p - q\|_1 > \varepsilon$. A landmark result of [204] states that the sample complexity of this question is (essentially) characterized by the $\ell_{2/3}$-quasinorm of p.

Here, we ask whether or not a similar characterization can be obtained for the quantum version of the question. Prior work of [65] demonstrated such a characterization, but under the caveat that the measurements are chosen non-adaptively. At a high level, they showed that the copy complexity of the problem is governed by the *fidelity* between σ and the maximally mixed state (recall Definition 1.4.7). More precisely, they showed that if $\overline{\sigma}$ and $\underline{\sigma}$ are states given by zeroing out eigenvalues of σ that have total mass at most $\Theta(\varepsilon^2)$ and $\Theta(\varepsilon)$ respectively and normalizing, then the copy complexity with non-adaptive measurements, denoted n, satisfies

$$\widetilde{\Omega}\left(\frac{d \cdot \underline{d}_{\mathsf{eff}}^{1/2}}{\varepsilon^2} \cdot F(\underline{\sigma}, \frac{1}{d} I_d)\right) \leqslant n \leqslant \widetilde{O}\left(\frac{d \cdot \overline{d}_{\mathsf{eff}}^{1/2}}{\varepsilon^2} \cdot F(\overline{\sigma}, \frac{1}{d} I_d)\right), \tag{5.1}$$

where $\underline{d}_{\mathsf{eff}}$ (resp. $\overline{d}_{\mathsf{eff}}$) is the "effective dimension" of the problem, namely, the rank of

$\underline{\sigma}$ (resp. $\overline{\sigma}$). In the same work, they also gave lower bounds for arbitrary (possibly adaptive) incoherent measurements, but, like with mixedness testing, these lower bounds were looser and did not match the corresponding upper bound. In light of this, we ask whether we can give an instance-optimal characterization of the complexity of state certification with incoherent measurements. Our second result here is to give such a characterization:

Theorem 5.1.2 (Informal, see Theorem 5.5.1)**.** *For any σ, and ε sufficiently small, the copy complexity of state certification w.r.t. σ using incoherent measurements is upper and lower bounded by* (5.1)*.*

We regard this as strong evidence that, as with mixedness testing, adaptivity does not help for state certification. It is not always a tight bound, as there are states for which the upper and lower bounds in (5.1) can differ by polynomial factors for some choices of ε, and so this bound can be loose, even in the non-adaptive setting. Still, we conjecture that for all σ, the copy complexity of state certification to σ with incoherent and non-adaptive measurements is the same as that with arbitrary incoherent measurements. Indeed, when ε is sufficiently small compared to the smallest nonzero eigenvalue of σ, our bounds are tight up to logarithmic factors.

Our techniques. We achieve our new lower bounds via a new proof technique which we believe may be of independent interest. As with other lower bounds in this area, we reduce to a "one-versus-many" distinguishing problem. To construct this instance, prior work leveraged the natural quantum analogue of Paninski's famous construction in the lower bound for (classical) uniformity testing [174] – namely, an additive perturbation by a multiple of $UZU^\dagger$, where U is a Haar random matrix and $Z = \mathrm{diag}(1, \ldots, -1, \ldots)$ has equally many $+1$s and -1s.

We instead use a different hard instance based on Gaussian perturbations. While this introduces a number of additional technical challenges, the key advantage of this instance is that the likelihood ratio for this instance has a very clean, self-similar form. This allows us to essentially reduce the problem into one of understanding the

concentration of a certain matrix martingale defined by the learning process, as well as an auxiliary matrix balancing question. We can then use classical tools from scalar and matrix concentration to demonstrate that the likelihood ratio is close to 1 with high probability over all possible outcomes of the learning algorithm, which yields our desired lower bound.

Not only does this framework dramatically simplify many of the difficult concentration calculations in prior work such as [54], it also has the conceptual advantage that it never requires a *pointwise* bound on the likelihood ratio. To our knowledge, all prior lower bounds against adaptive algorithms in this literature required some worst-case pointwise bound on the likelihood ratio. For some problems, e.g. shadow tomography [66], this was already sufficient to prove tight lower bounds. However, for mixedness testing, a worst-case bound cannot be sufficient, and from a technical perspective, the fact that [54] had to balance between their (much tighter) average case bound on the likelihood ratio and this (fairly large) worst-case bound to control the contribution of certain tail events was why their overall lower bound was loose. Consequently, we believe that this martingale-based technique may also yield tight lower bounds for a number of other problems in the literature.

5.1.1 Additional Related Work

The questions we consider in this chapter fall under the domain of quantum state property testing. See [165] for a more complete survey on property testing of quantum states. In this literature, roughly speaking, there are two settings considered, the *asymptotic regime*, and the *non-asymptotic regime*, the latter of which is the setting we study.

In the former setting, one considers the regime of parameters where $n \to \infty$ and d, ε are held fixed and relatively small, and the goal is to precisely characterize the exponential rate of convergence as a function of n. In this setting, quantum state certification is usually called *quantum state discrimination*, see e.g. [58, 25, 41] and references within. However, since d and ε are fixed, this allows for rates which could

depend exponentially on the dimensionality of the problem.

Instead, we consider the "non-asymptotic regime," where the goal is to characterize the statistical rate, as a function of d and ε. Similar work in this regime includes the aforementioned works of [169] and [27]. However, as described previously, their algorithms require using fully entangled measurements.

The work here falls into the line of work considering restricted classes of measurements, and specifically, those with without quantum memory. Understanding the power of such algorithms in the context of mixedness testing and, more generally, spectrum testing was posed as an open problem in [215]. Similar questions have also been considered in other settings, such as shadow tomography [1]. However, until recently, lower bounds for algorithms without quantum memory usually only held in the non-adaptive setting, e.g. [102, 65]. Recent work of [54] demonstrated the first lower bound against general (possibly adaptive) incoherent measurements for such a task. Subsequently, there has been a flurry of work demonstrating similar bounds in a variety of settings [122, 5, 124, 123, 66, 17, 64, 156, 63]. It is an interesting question if our techniques can be extended to also improve any of the lower bounds in these works.

Other restricted models of computation have also been considered in the literature. [220] gives algorithms for various quantum property testing problems using local measurements which act on each individual qubit, and in an non-adaptive manner. A number of works considers the special case where the measurements are only Pauli matrices [85, 86, 74, 19]. Overall, these classes of measurements seem to be much more restrictive than general non-adaptive measurements. In particular, the copy complexity of tasks such as mixedness testing under these measurements seem to be asymptotically higher than general incoherent measurements.

5.2 Preliminaries

We will represent an adaptive algorithm that makes incoherent measurements in the same way as in the previous chapter (recall Definition 4.2.1).

Notation. Given $z \in \mathbb{R}$, we use z_- to denote $-\min(z, 0)$. We use $\wedge$ and $\vee$ to denote min and max.

For a string $\boldsymbol{x} = (x_1, \ldots, x_n)$, we let $\boldsymbol{x}_{\sim i}$ and $\boldsymbol{x}_{\sim i,j}$ denote the string with the i-th index removed and the string with the i-th and j-th indices removed. For any set $S \subseteq [n]$, we let $\boldsymbol{x}_S$ denote the string restricted to the entries in S.

We will work with the following random matrix ensembles:

Definition 5.2.1 (Trace-centered Gaussian orthogonal ensemble (GOE))**.** For $d \in \mathbb{N}$, let $G \sim \mathrm{GOE}(d)$, that is, $G \in \mathbb{R}^{d \times d}$ is symmetric with upper diagonal entries sampled independently from $\mathcal{N}(0, 1/d)$ and diagonal entries sampled independently from $\mathcal{N}(0, 2/d)$.

Define $M = G - \frac{\mathrm{tr}(G)}{d} I_d$. We say that M is a *trace-centered GOE matrix* and denote its distribution $\mathrm{GOE}^*(d)$. For $U \subseteq \boldsymbol{R}^{d \times d}$, $\overline{M}$ is a *U-truncated trace-centered GOE matrix* if it is drawn from $\mathrm{GOE}^*(d)$ conditioned on $\overline{M} \in U$. We denote the distribution of $\overline{M}$ by $\mathrm{GOE}^*_U(d)$.

Definition 5.2.2 (Truncated Ginibre)**.** For $d_1, d_2 \in \mathbb{N}$, let $G \sim \mathrm{Gin}(d_1, d_2)$ be the (normalized) $d_1 \times d_2$ *Ginibre matrix*, that is, $G \in \mathbb{R}^{d_1 \times d_2}$ has i.i.d. entries $\mathcal{N}(0, 1/d_1)$. For $U \subseteq \mathbb{R}^{d_1 \times d_2}$, $\overline{G}$ is a *U-truncated $d_1 \times d_2$ Ginibre matrix* if it is drawn from $\mathrm{Gin}(d_1, d_2)$ conditioned on $\overline{G} \in U$. We denote the distribution of $\overline{G}$ by $\mathrm{Gin}_U(d)$.

Our lower bounds are based on Le Cam's two-point method which we briefly review here. The following is an elementary result in binary hypothesis testing:

Fact 5.2.3 (See e.g. Theorem 4.3 from [216])**.** *Given distributions p_0, p_1 over a domain $\mathcal{S}$, if $d_{\mathrm{TV}}(p_0, p_1) < 1/3$, there is no $\mathcal{A} : \mathcal{S} \to \{0, 1\}$ for which $\mathbf{Pr}[x \sim p_i]\mathcal{A}(x) = i \geqslant 2/3$*

for both $i = 0, 1$.

Now consider a state distinguishing task of the form

$$H_0 : \rho = \sigma \qquad \text{and} \qquad H_1 : \rho = \sigma_M, \tag{5.2}$$

where σ_M is a random state sampled from some distribution $\mathcal{D}$ over the set of states satisfying $\|\sigma - \sigma_M\|_1 > \varepsilon$. Recall from Definition 4.2.1 that a learning algorithm that uses n incoherent measurements corresponds to a tree $\mathcal{T}$ of depth n, and $\rho = \sigma$ and $\rho = \sigma_M$ induce distributions p_0 and p_M on the leaves of this tree. We can use Fact 5.2.3 to reduce proving a copy complexity lower bound for state certification with respect to σ, which is a worst-case guarantee over all possible input states ρ, to bounding $d_{\mathrm{TV}}(p_0, \mathbb{E}[M]p_M)$, which is an average-case bound.

Lemma 5.2.4 (Le Cam's two-point method, see e.g. Lemma 1 in [219])**.** *If there is a distribution* $\mathcal{D}$ *over states satisfying* $\|\sigma - \sigma_M\|_1 > \varepsilon$ *for which* $d_{\mathrm{TV}}(p_0, \mathbb{E}[M]p_M) \leqslant 1/3$ *for any tree* $\mathcal{T}$ *of depth* n*, then any algorithm* $\mathcal{A}$ *using incoherent measurements for state certification with respect to* σ *must make more than* n *incoherent measurements to achieve success probability at least* $2/3$.

Proof. Suppose to the contrary there existed such an algorithm $\mathcal{A}$ using at most n incoherent measurements, and let p_0 and p_M denote the distributions over the leaves of the tree corresponding to $\mathcal{A}$ when $\rho = \sigma$ and $\rho = \sigma_M$ respectively. Suppose when it succeeds, $\mathcal{A}$ outputs 0 when $\rho = \sigma$ and 1 when $\|\rho - \sigma\|_1 > \varepsilon$. Let $p_1 \triangleq \mathbb{E}[M \sim \mathcal{D}]p_M$. Because $\mathcal{A}$ successfully outputs 1 with probability $2/3$ when given as input the state σ_M for any M, $2/3 \leqslant \mathbb{E}[M]\mathbf{Pr}[\boldsymbol{x} \sim p_M]\mathcal{A}(\boldsymbol{x}) = 1 = \mathbb{E}[\boldsymbol{x} \sim p_1]\mathcal{A}(\boldsymbol{x}) = 1$. Similarly, $2/3 \leqslant \mathbb{E}[\boldsymbol{x} \sim p_0]\mathcal{A}(\boldsymbol{x}) = 0$. By Fact 5.2.3, this would contradict the bound on $d_{\mathrm{TV}}(p_0, p_1)$. □

5.3 Lower Bound for Mixedness Testing

In this section we prove the following theorem, which is the formal version of Theorem 5.1.1.

Theorem 5.3.1. *Let $d \gg 1$ and $0 < \varepsilon \leqslant 1/12$. Any algorithm that uses incoherent measurements which, given n copies of a mixed state $\rho \in \mathbb{C}^{d \times d}$, can distinguish between the case where $\rho = I_d/d$ and where $\|\rho - I_d/d\|_1 > \varepsilon$ with probability at least $2/3$, must use at least $n = \Omega(d^{3/2}/\varepsilon^2)$ copies.*

By the upper bound in [54], this is tight up to constant factors. Also note that by standard amplification arguments, the choice of constant in the success probability is arbitrary, and can be taken to be any constant which is strictly larger than $1/2$.

In fact, we will prove a slightly stronger theorem, which will be useful later on for our lower bounds against state certification. Namely, we will show that the same bound holds not just when the null hypothesis is the maximally mixed state, but for any state whose smallest and largest eigenvalues are comparable.

More formally, let $A \in \boldsymbol{R}^{d \times d}$ be a diagonal matrix with diagonal entries $a_1 \geqslant \cdots \geqslant a_d > 0$, satisfying $2a_d \geqslant a_1$, and $\operatorname{tr}(A) = d$. We consider the task of distinguishing between the following two alternatives:

$$H_0 : \rho = \frac{1}{d}A \qquad \text{and} \qquad H_1 : \rho = \frac{1}{d}(A + \varepsilon \overline{M}). \tag{5.3}$$

Here, $\overline{M} \sim \mathrm{GOE}_U^*(d)$ for the U given by Lemma 5.3.2 below.

Lemma 5.3.2. *There exists $U \subseteq \boldsymbol{R}^{d \times d}$ such that if $M \sim \mathrm{GOE}^*(d)$, then $\mathbf{Pr}\, M \notin U \leqslant \exp(-\Omega(d))$ and on the event $M \in U$, we have $\|M\|_{\mathrm{op}} \leqslant 3$ and $\|M\|_1 \geqslant d/12$.*

We defer the proof of this lemma to Appendix C.3. Our main result for the distinguishing task (5.3) is the following.

Theorem 5.3.3. *If $d \gg 1$ and $\varepsilon \leqslant 1/12$, then any algorithm using incoherent*

measurements that distinguishes between H_0 and H_1 with success probability at least 2/3 requires $n = \Omega(d^{3/2}/\varepsilon^2)$ copies.

Again, by standard amplification arguments, the choice of constant in the success probability is arbitrary, and can be taken to be any constant greater than 1/2. Note that the bounds in Lemma 5.3.2 ensure that under H_1, ρ is psd (and thus a valid quantum state) and has trace distance $\Omega(\varepsilon)$ to $\frac{1}{d}A$. In particular, since any algorithm for mixedness testing must solve this distinguishing problem as well, setting $A = I_d$ into Theorem 5.3.3 immediately implies Theorem 5.3.1.

Take any learning tree $\mathcal{T}$ corresponding to an algorithm for this task that uses n incoherent measurements. Recalling the terminology from Definition 4.2.1, we let p_0 and p_1 denote the distributions over leaves of $\mathcal{T}$ induced by ρ under H_0 and H_1 respectively. In the rest of this section, we assume $n \ll d^{3/2}/\varepsilon^2$ and will prove $d_{\mathrm{TV}}(p_0, p_1) = o(1)$. It is clear that this immediately implies Theorem 5.3.3.

We let $L^*(\cdot)$ denote the likelihood ratio between p_1 and p_0. That is, for a sequence of vectors $\boldsymbol{x} = (x_1, \ldots, x_n)$, let $L^*(\boldsymbol{x}) \triangleq p_1(\boldsymbol{x})/p_0(\boldsymbol{x})$. Note that

$$L^*(\boldsymbol{x}) = \mathbb{E}[\overline{M} \sim \mathrm{GOE}_U^*(d)] * \prod_{i=1}^{n} \left(1 + \varepsilon \frac{x_i^\dagger \overline{M} x_i}{x_i^\dagger A x_i}\right). \tag{5.4}$$

Define similarly

$$L(\boldsymbol{x}) \triangleq \mathbb{E}[M \sim \mathrm{GOE}^*(d)] * \prod_{i=1}^{n} \left(1 + \varepsilon \frac{x_i^\dagger M x_i}{x_i^\dagger A x_i}\right). \tag{5.5}$$

This is an estimate for the likelihood ratio $L^*(\boldsymbol{x})$ where the conditioned Gaussian integral is replaced by a true Gaussian integral. Most of the computations in this section will be done in terms of $L(\boldsymbol{x})$; the proof of Theorem 5.3.3 below quantifies that $L(\boldsymbol{x})$ is a close approximation of $L^*(\boldsymbol{x})$.

Throughout this section, we will somewhat abuse notation and write $L(\boldsymbol{z})$ for any sequence of unit vectors $\boldsymbol{z} = (z_1, \ldots, z_t)$ of length not necessarily n. This is defined

the same way as in (5.5). We also write $L(\boldsymbol{x}, \boldsymbol{x})$ to denote the value of L on input $(x_1, x_1, x_2, x_2, \ldots, x_n, x_n)$.

The main ingredient in the proof of Theorem 5.3.3 is the following high-probability bound on L evaluated at the leaves of $\mathcal{T}$.

Proposition 5.3.4. *There exists a subset S of the leaves of $\mathcal{T}$ such that $\mathbf{Pr}[p_0] * S = 1 - o(1)$ and for all $\boldsymbol{x} \in S$, $|L(\boldsymbol{x}) - 1| = o(1)$ and $L(\boldsymbol{x}, \boldsymbol{x}) \ll e^{\sqrt{d}}$.*

Let us first prove Theorem 5.3.3 assuming Proposition 5.3.4.

Proof of Theorem 5.3.3. Let U be as in Lemma 5.3.2. Define

$$\overline{L}(\boldsymbol{x}) = \mathbb{E}[M \sim \mathrm{GOE}^*(d)] * I\{M \in U\} \prod_{i=1}^{n} \left(1 + \varepsilon \frac{x_i^\dagger M x_i}{x_i^\dagger A x_i}\right).$$

It is clear that $L^*(\boldsymbol{x}) = \mathbf{Pr} * U^{-1} \overline{L}(\boldsymbol{x})$. For all $\boldsymbol{x} \in S$, by Cauchy-Schwarz

$$\begin{aligned}
|L(\boldsymbol{x}) - \overline{L}(\boldsymbol{x})| &= \left| \mathbb{E}[M \sim \mathrm{GOE}^*(d)] * I\{M \notin U\} \prod_{i=1}^{n} \left(1 + \varepsilon \frac{x_i^\dagger M x_i}{x_i^\dagger A x_i}\right) \right| \\
&\leqslant \mathbf{Pr} * U^{c 1/2} \, \mathbb{E}[M \sim \mathrm{GOE}^*(d)] * \prod_{i=1}^{n} \left(1 + \varepsilon \frac{x_i^\dagger M x_i}{x_i^\dagger A x_i}\right)^{2 1/2} \\
&= \sqrt{\mathbf{Pr} * U^c L(\boldsymbol{x}, \boldsymbol{x})} = o(1).
\end{aligned}$$

Here we use that $\mathbf{Pr} * U^c \leqslant \exp(-\Omega(d))$ and $L(\boldsymbol{x}, \boldsymbol{x}) \ll e^{\sqrt{d}}$. Moreover, we have $|L(\boldsymbol{x}) - 1| = o(1)$. Thus, for all $\boldsymbol{x} \in S$, $\overline{L}(\boldsymbol{x}) = 1 + o(1)$ and

$$\begin{aligned}
|L^*(\boldsymbol{x}) - 1| &\leqslant |L^*(\boldsymbol{x}) - \overline{L}(\boldsymbol{x})| + |\overline{L}(\boldsymbol{x}) - 1| \\
&= \frac{\mathbf{Pr} * U^c}{\mathbf{Pr} * U} \overline{L}(\boldsymbol{x}) + o(1) = o(1).
\end{aligned}$$

Finally,

$$d_{\mathrm{TV}}(p_0, p_1) = 2 \, \mathbb{E}[\boldsymbol{x} \sim p_0] * (L^*(\boldsymbol{x}) - 1)_-$$

$$= 2\,\mathbb{E}[\boldsymbol{x} \sim p_0] * I\{\boldsymbol{x} \in S\}(L^*(\boldsymbol{x}) - 1)_- + 2\,\mathbb{E}[\boldsymbol{x} \sim p_0] * I\{\boldsymbol{x} \notin S\}(L^*(\boldsymbol{x}) - 1)_-$$

$$\leqslant 2 \sup_{\boldsymbol{x} \in S}(L^*(\boldsymbol{x}) - 1)_- + 2\,\mathbf{Pr}[p_0] * S^c = o(1). \qquad \square$$

5.3.1 Recursive Evaluation of Likelihood Ratio

Let $\boldsymbol{z} = (z_1, \ldots, z_t)$ be a sequence of unit vectors. For $1 \leqslant i \leqslant t$, let $\boldsymbol{z}_{\sim i}$ be the sequence $\boldsymbol{z}$ with z_i omitted. Similarly, for $1 \leqslant i < j \leqslant t$, let $\boldsymbol{z}_{\sim i,j}$ be the sequence $\boldsymbol{z}$ with z_i, z_j omitted. The main result of this subsection is the following recursive formula for $L(\boldsymbol{z})$.

Lemma 5.3.5. *The function L satisfies*

$$L(\boldsymbol{z}) = L(\boldsymbol{z}_{\sim t}) + \frac{2\varepsilon^2}{d^2} \sum_{i=1}^{t-1} \left[\frac{d\langle z_i, z_t\rangle^2 - 1}{(z_i^\dagger A z_i)(z_t^\dagger A z_t)} L(\boldsymbol{z}_{\sim i,t}) \right].$$

The proof is based on Isserlis' theorem, which we record below. For k even, let $\mathrm{PMat}(k)$ denote the set of perfect matchings of $\{1, \ldots, k\}$.

Theorem 5.3.6 ([127])**.** *Let $g = (g_1, \ldots, g_k)$ be a jointly Gaussian vector. If k is odd, then $\mathbb{E} \prod_{i=1}^k g_i = 0$. If k is even, then*

$$\mathbb{E} * \prod_{i=1}^{k} g_i = \sum_{\{\{a_1,b_1\},\ldots,\{a_{k/2},b_{k/2}\}\} \in \mathrm{PMat}(k)} \prod_{i=1}^{k/2} \mathbb{E} * g_{a_i} g_{b_i}.$$

Proof of Lemma 5.3.5. For a set $S \subseteq [t]$ with $|S|$ even, let $\mathrm{PMat}(S)$ denote the set of perfect matchings of S. For even $k \leqslant t$, let $\mathrm{Mat}(t, k)$ denote the set of matchings of $[t]$ consisting of $k/2$ pairs. We compute that

$$L(\boldsymbol{z}) = \sum_{S \subseteq [t]} \varepsilon^{|S|}\, \mathbb{E}[M \sim \mathrm{GOE}^*(d)] * \prod_{i \in S} \frac{z_i^\dagger M z_i}{z_i^\dagger A z_i} \qquad \text{(expanding (5.5))}$$

$$= \sum_{\substack{S \subseteq [t] \\ |S| \text{ even}}} \varepsilon^{|S|} \sum_{\{\{a_1,b_1\},\ldots,\{a_{|S|/2},b_{|S|/2}\}\} \in \mathrm{PMat}(S)} \prod_{i=1}^{|S|/2} \mathbb{E}[M \sim \mathrm{GOE}^*(d)] * \frac{z_{a_i}^\dagger M z_{a_i}}{z_{a_i}^\dagger A z_{a_i}} \cdot \frac{z_{b_i}^\dagger M z_{b_i}}{z_{b_i}^\dagger A z_{b_i}} \qquad \text{(Th. 5.3.6)}$$

$$= \sum_{k=0}^{\lfloor t/2 \rfloor} \varepsilon^{2k} \sum_{\{\{a_1,b_1\},\ldots,\{a_k,b_k\}\} \in \mathrm{Mat}(t,2k)} \prod_{i=1}^{k} \mathbb{E}[M \sim \mathrm{GOE}^*(d)] * \frac{z_{a_i}^\dagger M z_{a_i}}{z_{a_i}^\dagger A z_{a_i}} \cdot \frac{z_{b_i}^\dagger M z_{b_i}}{z_{b_i}^\dagger A z_{b_i}}$$

$$= \sum_{k=0}^{\lfloor t/2 \rfloor} \left(\frac{2\varepsilon^2}{d^2}\right)^k \sum_{\{\{a_1,b_1\},\ldots,\{a_k,b_k\}\} \in \mathrm{Mat}(t,2k)} \prod_{i=1}^{k} \frac{d\langle z_{a_i}, z_{b_i}\rangle^2 - 1}{(z_{a_i}^\dagger A z_{a_i})(z_{b_i}^\dagger A z_{b_i})}. \tag{5.6}$$

In the final step we use that for unit vectors $x, y \in \mathbb{C}^d$,

$$\mathbb{E}[M \sim \mathrm{GOE}^*(d)] * (x^\dagger M x)(y^\dagger M y) = \frac{2}{d^2}(d\langle x, y\rangle^2 - 1),$$

which can be verified by direct computation. The lemma follows by partitioning the summands in (5.6) based on whether t appears in the matching, and if so which $i \in \{1, \ldots, t-1\}$ it is paired with. □

5.3.2 High Probability Bound on Likelihood Ratio at Leaves

This subsection gives the main part of the proof of Proposition 5.3.4. For any sequence of unit vectors $\boldsymbol{z} = (z_1, \ldots, z_t)$, define

$$H(\boldsymbol{z}) = \sum_{i=1}^{t} \frac{d z_i z_i^\dagger - I_d}{z_i^\dagger A z_i} \cdot \frac{L(\boldsymbol{z}_{\sim i})}{L(\boldsymbol{z})} \qquad \text{and} \qquad K(\boldsymbol{z}) = \sum_{i=1}^{t} \frac{d z_i z_i^\dagger - I_d}{z_i^\dagger A z_i}.$$

The function H enters our calculations by the following rewriting of Lemma 5.3.5:

$$\frac{L(\boldsymbol{z})}{L(\boldsymbol{z}_{\sim t})} = 1 + \frac{2\varepsilon^2}{d^2} \cdot \frac{z_t^\dagger H(\boldsymbol{z}_{\sim t}) z_t}{z_t^\dagger A z_t}. \tag{5.7}$$

If $\boldsymbol{z} = \boldsymbol{x}_{\leqslant t} \triangleq (x_1, \ldots, x_t)$ is a prefix of $\boldsymbol{x} \sim p_0$, then $\frac{L(\boldsymbol{z})}{L(\boldsymbol{z}_{\sim t})} = \frac{L(\boldsymbol{x}_{\leqslant t})}{L(\boldsymbol{x}_{\leqslant t-1})}$ is one step in the likelihood ratio martingale. As we will see (proof of Claim 5.3.10) below, the multiplicative fluctuation of this step is

$$\mathbb{E}[x_t] * \left(\frac{L(\boldsymbol{x}_{\leqslant t})}{L(\boldsymbol{x}_{\leqslant t-1})}\right)^2 = 1 + O\left(\frac{\varepsilon^4}{d^5}\right) \|H(\boldsymbol{x}_{\leqslant t-1})\|_F^2.$$

Thus, an upper bound on $\|H(\boldsymbol{z})\|_F$ over all prefixes $\boldsymbol{z}$ of $\boldsymbol{x}$ controls the fluctuations

of the likelihood ratio martingale. Because the matrices output by H are hard to control directly, we will use the function K as a proxy for H. The following lemma quantifies this relationship, showing that if $K(\boldsymbol{z})$ is bounded in Frobenius norm, $H(\boldsymbol{z})$ is bounded at the same scale.

Lemma 5.3.7. *Suppose $1 \ll \gamma \ll d/(\varepsilon^2 n^{1/2})$. If $\boldsymbol{z} = (z_1, \dots, z_t)$ is a sequence of unit vectors satisfying $t \leqslant n$ and $\|K(\boldsymbol{z})\|_F \leqslant n^{1/2} d\gamma$, then $\|H(\boldsymbol{z})\|_F \leqslant 3n^{1/2} d\gamma$.*

Note that this lemma is a "deterministic" statement about a sequence of vectors. We will prove this in Subsection 5.3.3 using the bootstrap argument alluded to earlier. The following lemma bounds $K(\boldsymbol{z})$ in Frobenius norm uniformly over all prefixes $\boldsymbol{z}$ of $\boldsymbol{x}$. We will prove this lemma in Subsection 5.3.4 by mimicking the proof of Doob's L^2 maximal inequality for the matrix valued martingale $K(\boldsymbol{x}_{\leqslant t})$.

Lemma 5.3.8. *If $\boldsymbol{x} \sim p_0$, then $\mathbb{E} * \sup_{1 \leqslant t \leqslant n} \|K(\boldsymbol{x}_{\leqslant t})\|_F^2 \lesssim nd^2$.*

We will now prove Proposition 5.3.4 assuming Lemmas 5.3.7 and 5.3.8. We set α, β to be slowly-growing functions such that $1 \ll \alpha \ll \beta \ll d^{3/2}/(\varepsilon^2 n) \wedge d/(\varepsilon^2 n^{1/2})$, and furthermore $\alpha^2 \ll d^{3/2}/(\varepsilon^2 n)$. This is possible because $n \ll d^{3/2}/\varepsilon^2$.

Let $\boldsymbol{x} \sim p_0$. For $1 \leqslant t \leqslant n$, define the filtration $\mathcal{F}_t = \sigma(\boldsymbol{x}_{\leqslant t})$ and the sequences

$$H_t = H(\boldsymbol{x}_{\leqslant t}), \qquad K_t = K(\boldsymbol{x}_{\leqslant t}), \qquad \Phi_t = L(\boldsymbol{x}_{\leqslant t}). \tag{5.8}$$

Consider the time

$$\tau = \inf \left\{ t : \|K_t\|_F > n^{1/2} d\alpha \text{ or } |\Phi_t - 1| > \frac{\varepsilon^2 n}{d^{3/2}} \beta \right\} \cup \{\infty\},$$

which is clearly a stopping time with respect to $\mathcal{F}_t$. Also define the stopped sequence $\Psi_t = \Phi_{t \wedge \tau}$.

Claim 5.3.9. With probability $1 - o(1)$, $\|K_t\|_F \leqslant n^{1/2} d\alpha$ for all $1 \leqslant t \leqslant n$.

Proof. By Lemma 5.3.8,

$$\mathbf{Pr} * \sup_{1 \leqslant t \leqslant n} \|K_t\|_F > n^{1/2} d\alpha \leqslant \frac{\mathbb{E} * \sup_{1 \leqslant t \leqslant n} \|K_t\|_F^2}{nd^2\alpha^2} \lesssim \alpha^{-2} = o(1). \qquad \square$$

Claim 5.3.10. With probability $1 - o(1)$, $|\Psi_n - 1| \leqslant \frac{\varepsilon^2 n}{d^{3/2}} \beta$.

Proof. Note that Ψ_t is a multiplicative martingale: if $\tau \leqslant t-1$ then certainly $\mathbb{E} \frac{\Psi_t}{\Psi_{t-1}} | \mathcal{F}_{t-1} = 1$, and if $\tau > t-1$, (5.7) implies

$$\mathbb{E} * \frac{\Psi_t}{\Psi_{t-1}} | \mathcal{F}_{t-1} = 1 + \frac{2\varepsilon^2}{d^2} \mathbb{E} * \frac{x_t^\dagger H_{t-1} x_t}{x_t^\dagger A x_t} | \mathcal{F}_{t-1} = 1,$$

using that

$$\mathbb{E} * \frac{x_t^\dagger H_{t-1} x_t}{x_t^\dagger A x_t} | \mathcal{F}_{t-1} = \sum_{x_t} \omega_{x_t} (x_t^\dagger H_{t-1} x_t) = \left\langle H_{t-1}, \sum_{x_t} \omega_{x_t} x_t x_t^\dagger \right\rangle = \langle H_{t-1}, I_d/d \rangle = 0. \tag{5.9}$$

We next bound the quadratic increment $\mathbb{E} (\frac{\Psi_t}{\Psi_{t-1}})^2 | \mathcal{F}_{t-1}$. If $\tau \leqslant t-1$ this is 1, and otherwise

$$\mathbb{E} * \left(\frac{\Psi_t}{\Psi_{t-1}} \right)^2 | \mathcal{F}_{t-1} = 1 + \frac{4\varepsilon^2}{d^2} \mathbb{E} * \frac{x_t^\dagger H_{t-1} x_t}{x_t^\dagger A x_t} | \mathcal{F}_{t-1} + \frac{4\varepsilon^4}{d^4} \mathbb{E} * \frac{(x_t^\dagger H_{t-1} x_t)^2}{(x_t^\dagger A x_t)^2} | \mathcal{F}_{t-1}. \tag{5.10}$$

The first expectation is zero by (5.9). To bound the remaining expectation, note that for any unit vector x,

$$x^\dagger A x \geqslant a_d x^\dagger x = a_d \geqslant \frac{1}{2}. \tag{5.11}$$

So,

$$\mathbb{E} * \frac{(x_t^\dagger H_{t-1} x_t)^2}{(x_t^\dagger A x_t)^2} | \mathcal{F}_{t-1} \leqslant 2 \mathbb{E} * \frac{(x_t^\dagger H_{t-1} x_t)^2}{x_t^\dagger A x_t} | \mathcal{F}_{t-1} = 2 \sum_{x_t} \omega_{x_t} x_t^\dagger H_{t-1} (x_t x_t^\dagger) H_{t-1} x_t \tag{5.12}$$

$$\leqslant 2 \sum_{x_t} \omega_{x_t} x_t^\dagger H_{t-1}^2 x_t = 2 \left\langle H_{t-1}^2, \sum_{x_t} \omega_{x_t} x_t x_t^\dagger \right\rangle = 2 \langle H_{t-1}^2, I_d/d \rangle = \frac{2}{d} \|H_{t-1}\|_F^2. \tag{5.13}$$

Moreover, since $\tau > t-1$, $\|K_{t-1}\|_F \leqslant n^{1/2}d\alpha$ and Lemma 5.3.7 implies $\|H_{t-1}\|_F \leqslant 3n^{1/2}d\alpha$. Thus,

$$\mathbb{E}*\left(\frac{\Psi_t}{\Psi_{t-1}}\right)^2 |\mathcal{F}_{t-1} \leqslant 1 + \frac{8\varepsilon^4}{d^5}\|H_{t-1}\|_F^2 \leqslant 1 + \frac{72\varepsilon^4 n}{d^3}\alpha^2.$$

So, for all $1 \leqslant t \leqslant n$,

$$\mathbb{E}\,\Psi_t^2 = \mathbb{E}*\mathbb{E}*\left(\frac{\Psi_t}{\Psi_{t-1}}\right)^2 |\mathcal{F}_{t-1}\Psi_{t-1}^2 \leqslant \left(1 + \frac{72\varepsilon^4 n}{d^3}\alpha^2\right)\mathbb{E}\,\Psi_{t-1}^2,$$

and therefore

$$\mathbb{E}\,\Psi_t^2 \leqslant \left(1 + \frac{72\varepsilon^4 n}{d^3}\alpha^2\right)^n \leqslant \exp\left(\frac{72\varepsilon^4 n^2}{d^3}\alpha^2\right) \leqslant 2$$

since $\frac{\varepsilon^4 n^2}{d^3}\alpha^2 \ll 1$. Moreover,

$$\begin{aligned}\mathbb{E}\,(\Psi_t - 1)^2 &= \mathbb{E}*\mathbb{E}*\left(\frac{\Psi_t}{\Psi_{t-1}}\right)^2 |\mathcal{F}_{t-1}\Psi_{t-1}^2 - 2\,\mathbb{E}*\frac{\Psi_t}{\Psi_{t-1}}|\mathcal{F}_{t-1}\Psi_{t-1} + 1\\ &\leqslant \frac{72\varepsilon^4 n}{d^3}\alpha^2\,\mathbb{E}\,\Psi_{t-1}^2 + \mathbb{E}\,(\Psi_{t-1}-1)^2\\ &\leqslant \frac{144\varepsilon^4 n}{d^3}\alpha^2 + \mathbb{E}\,(\Psi_{t-1}-1)^2,\end{aligned}$$

so by induction

$$\mathbb{E}\,(\Psi_n - 1)^2 \leqslant \frac{144\varepsilon^4 n^2}{d^3}\alpha^2.$$

Thus

$$\mathbf{Pr}*|\Psi_n - 1| > \frac{\varepsilon^2 n}{d^{3/2}}\beta \leqslant \frac{\mathbb{E}*|\Psi_n - 1|^2}{\frac{\varepsilon^4 n^2}{d^3}\beta^2} \leqslant \frac{144\alpha^2}{\beta^2} = o(1).$$

Therefore, $|\Psi_n - 1| \leqslant \frac{\varepsilon^2 n}{d^{3/2}}\beta$ with probability $1 - o(1)$. □

Claim 5.3.11. If $\|K_n\|_F \leqslant n^{1/2}d\alpha$, then $L(\boldsymbol{x}, \boldsymbol{x}) \ll e^{\sqrt{d}}$.

Proof. Using the elementary inequality $e^{2z+z^2} \geqslant (1+z)^2$ and then Cauchy Schwarz,

we can write

$$L(\boldsymbol{x},\boldsymbol{x}) = \mathbb{E}[M \sim \mathrm{GOE}^*(d)] * \prod_{i=1}^{n} \left(1 + \varepsilon \frac{x_i^\dagger M x_i}{x_i^\dagger A x_i}\right)^2 \tag{5.14}$$

$$\leqslant \mathbb{E}[M \sim \mathrm{GOE}^*(d)] * \exp\left(\sum_{i=1}^{n} 2\varepsilon \frac{x_i^\dagger M x_i}{x_i^\dagger A x_i} + \left(\varepsilon \frac{x_i^\dagger M x_i}{x_i^\dagger A x_i}\right)^2\right) \tag{5.15}$$

$$\leqslant \sqrt{\mathbb{E}[M \sim \mathrm{GOE}^*(d)] * \exp\left(4\sum_{i=1}^{n} \varepsilon \frac{x_i^\dagger M x_i}{x_i^\dagger A x_i}\right) \mathbb{E}[M \sim \mathrm{GOE}^*(d)] * \exp\left(2\sum_{i=1}^{n} \left(\varepsilon \frac{x_i^\dagger M x_i}{x_i^\dagger A x_i}\right)^2\right)}. \tag{5.16}$$

Now we bound each of the terms in (5.16). For the first term, we have

$$\mathbb{E}[M \sim \mathrm{GOE}^*(d)] * \exp\left(4\sum_{i=1}^{n} \varepsilon \frac{x_i^\dagger M x_i}{x_i^\dagger A x_i}\right) = \mathbb{E}[M \sim \mathrm{GOE}^*(d)] * \exp\left(\frac{4\varepsilon}{d}\langle M, K_n\rangle\right), \tag{5.17}$$

where we used that $\mathrm{tr}(M) = 0$. As $M = G - \frac{\mathrm{tr}(G)}{d} I_d$ for $G \sim \mathrm{GOE}(d)$, we have that $\langle M, K_n\rangle = \langle G, K_n\rangle$ is distributed as a Gaussian with variance at most $\frac{2}{d}\|K_n\|_F^2 \leqslant 2nd\alpha^2$. So we can bound (5.17) by

$$\mathbb{E}[g \sim \mathcal{N}(0, 32\varepsilon^2 n\alpha^2/d)]\exp(g) = e^{16\varepsilon^2 n\alpha^2/d} \ll e^{\sqrt{d}} \tag{5.18}$$

as $\alpha^2 \ll d^{3/2}/(\varepsilon^2 n)$ by assumption. Next we bound the second term in the product in (5.16). Use $\mathsf{vec}(M)$ to denote rearranging M as a vector in $\mathbb{R}^{d^2}$ (done in a consistent way) and use $\otimes$ to denote the Kronecker product of two matrices. Let $Q \in \mathbb{R}^{d^2 \times d^2}$ be defined as

$$Q = \sum_{i=1}^{n} \frac{x_i x_i^\dagger \otimes x_i x_i^\dagger}{(x_i^\dagger A x_i)^2}.$$

We have

$$\mathbb{E}[M \sim \mathrm{GOE}^*(d)] * \exp\left(2\sum_{i=1}^{n} \left(\varepsilon \frac{x_i^\dagger M x_i}{x_i^\dagger A x_i}\right)^2\right) = \mathbb{E}[M \sim \mathrm{GOE}^*(d)] * \exp\left(2\varepsilon^2 \mathsf{vec}(M)^\dagger Q \mathsf{vec}(M)\right).$$

Now note that

$$Q \preceq 2\sum_{i=1}^{n} \frac{x_i x_i^\dagger \otimes I_d}{x_i^\dagger A x_i} \preceq \left(\frac{2K_n}{d} + \frac{4nI_d}{d}\right) \otimes I_d$$

so we have $\|Q\|_{\mathrm{op}} \leqslant 6n/d$. Also, we have

$$\|Q\|_1 \leqslant 2\sum_{i=1}^{n} \frac{x_i^\dagger x_i}{x_i^\dagger A x_i} \leqslant 4n$$

Let $\lambda_1, \ldots, \lambda_{d^2}$ be the eigenvalues of Q. Next note that we have

$$\begin{aligned}
\mathbb{E}[M \sim \mathrm{GOE}^*(d)] * \exp\left(2\varepsilon^2 \mathsf{vec}(M)^\dagger Q \mathsf{vec}(M)\right) &\leqslant \mathbb{E}[v \sim N(0, I_{d^2})] * \exp\left(10\varepsilon^2/d \cdot v^\dagger Q v\right) \\
&\leqslant \prod_{j=1}^{d^2} \mathbb{E}[g \sim N(0,1)] * \exp(10\varepsilon^2/d \cdot g^2 \lambda_j) \\
&\leqslant e^{20\varepsilon^2(\lambda_1 + \cdots + \lambda_{d^2})/d} \ll e^{\sqrt{d}}.
\end{aligned}$$

In the first step above, we used the convexity of the function inside the expectation to replace the distribution over $M \sim \mathrm{GOE}^*(d)$ with another distribution that can be obtained by adding independent, mean-0 noise to M. Afterwards, we used the rotational invariance of $N(0, I_{d^2})$ and then the bound on $\|Q\|_{\mathrm{op}}$ (together with the fact that $\mathbb{E}[g]e^{cx^2} = (1-2c)^{-1} \leqslant e^{2c}$ for sufficiently small c), and finally the bound on $\|Q\|_1$. Putting everything together, we conclude that $L(\boldsymbol{x}, \boldsymbol{x}) \ll e^{\sqrt{d}}$ as desired. □

Proof of Proposition 5.3.4. Define the event

$$S = \left\{ \sup_{1 \leqslant t \leqslant n} \|K_t\|_F \leqslant n^{1/2} d\alpha \text{ and } |\Psi_n - 1| \leqslant \frac{\varepsilon^2 n}{d^{3/2}}\beta \right\}.$$

By Claims 5.3.9 and 5.3.10, $\mathbf{Pr}[p_0]S = 1 - o(1)$. We will show that if S holds, then $\tau = \infty$. Indeed, if $\tau = t < \infty$, then either $\|K_t\|_F > n^{1/2} d\alpha$ or $|\Phi_t - 1| > \frac{\varepsilon^2 n}{d^{3/2}}\beta$ holds. Since $\Psi_n = \Phi_t$, this contradicts S.

So, $\tau = \infty$ on S. This implies $|L(\boldsymbol{x}) - 1| = |\Phi_n - 1| \leqslant \frac{\varepsilon^2 n}{d^{3/2}}\beta = o(1)$. Moreover $\|K_n\|_F \leqslant n^{1/2} d\alpha$, so by Claim 5.3.11 we have $L(\boldsymbol{x}, \boldsymbol{x}) \ll e^{\sqrt{d}}$. □

5.3.3 Bounding H in Frobenius Norm by Bootstrapping

In this subsection, we prove Lemma 5.3.7. Throughout this subsection, let $\boldsymbol{z} = (z_1, \ldots, z_t)$ be a sequence of unit vectors satisfying $t \leqslant n$ and

$$\|K(\boldsymbol{z})\|_F \leqslant n^{1/2} d\gamma \tag{5.19}$$

for some $1 \ll \gamma \ll d/(\varepsilon^2 n^{1/2})$.

The following lemma bounds a variant of $K(\boldsymbol{z})$ where we multiply each summand by an adversarial $b_i \in [-1, 1]$. This will be used to control the discrepancy $H(\boldsymbol{z}) - K(\boldsymbol{z})$ in the bootstrapping argument.

Lemma 5.3.12. *Uniformly over $b_1, \ldots, b_t \in [-1, 1]$, we have*

$$\|\sum_{i=1}^{t} b_i \frac{dz_i z_i^\dagger - I_d}{z_i^\dagger A z_i}\|_F \leqslant n^{1/2} d\gamma + 2nd^{1/2}.$$

Proof. For any choice of $b_1, \ldots, b_t$,

$$\begin{aligned}
\|\sum_{i=1}^{t} b_i \frac{dz_i z_i^\dagger - I_d}{z_i^\dagger A z_i}\|_F &\leqslant \|\sum_{i=1}^{t} b_i \frac{dz_i z_i^\dagger}{z_i^\dagger A z_i}\|_F + \|\sum_{i=1}^{t} b_i \frac{I_d}{z_i^\dagger A z_i}\|_F \\
&\leqslant \|\sum_{i=1}^{t} \frac{dz_i z_i^\dagger}{z_i^\dagger A z_i}\|_F + \|\sum_{i=1}^{t} \frac{I_d}{z_i^\dagger A z_i}\|_F \\
&\leqslant \|K(\boldsymbol{z})\|_F + 2\|\sum_{i=1}^{t} \frac{I_d}{z_i^\dagger A z_i}\|_F.
\end{aligned}$$

The second inequality holds because the matrices $dz_i z_i^\dagger$ and I_d are both psd. Using (5.11), we have

$$\|\sum_{i=1}^{t} \frac{I_d}{z_i^\dagger A z_i}\|_F \leqslant 2td^{1/2} \leqslant 2nd^{1/2}.$$

The result follows by the assumed bound on $\|K(\boldsymbol{z})\|_F$. □

For $S \subseteq [t]$, let $\boldsymbol{z}_S = (z_i)_{i \in S}$. Further, let

$$H_S = \sum_{i \in S} \frac{dz_i z_i^\dagger - I_d}{z_i^\dagger A z_i} \cdot \frac{L(\boldsymbol{z}_{S \setminus \{i\}})}{L(\boldsymbol{z}_S)} \qquad \text{and} \qquad K_S = \sum_{i \in S} \frac{dz_i z_i^\dagger - I_d}{z_i^\dagger A z_i}.$$

The following lemma gives a preliminary bound on $\|H_S\|_F$. In the proof of Lemma 5.3.7, we will use this bound to control $\|H_S\|_F$ for $|S| = t - O(\log n)$, followed by the bootstrap argument over $O(\log n)$ recursive rounds to contract the bound to $O(n^{1/2}d)$.

Lemma 5.3.13. *For all $S \subseteq [t]$, $\|H_S\|_F \leqslant 2n^{1/2}d\gamma + 4nd^{1/2}$.*

Proof. Note that for any fixed $\overline{M} \in U$, for the U given by Lemma 5.3.2, and any unit vector z,

$$\varepsilon \left| \frac{z^\dagger \overline{M} z}{z^\dagger A z} \right| \leqslant \frac{1}{12} \cdot \frac{3}{1/2} = \frac{1}{2},$$

so $1 + \varepsilon \frac{z^\dagger \overline{M} z}{z^\dagger A z} \in [1/2, 3/2]$. Thus, for all i, $L(\boldsymbol{z}_S)/L(\boldsymbol{z}_{S \setminus \{i\}}) \in [1/2, 3/2]$, which implies

$$\frac{L(\boldsymbol{z}_{S \setminus \{i\}})}{L(\boldsymbol{z}_S)} \in [2/3, 2]. \tag{5.20}$$

Lemma 5.3.12 gives

$$\frac{1}{2}\|H_S\|_F \leqslant n^{1/2}d\gamma + 2nd^{1/2},$$

as desired. □

Proof of Lemma 5.3.7. Let $D = \log \sqrt{n/d}$. If $t < D$, then by equations (5.11) and (5.20),

$$\|H(\boldsymbol{z})\|_F \leqslant \sum_{i=1}^{t} \frac{\|dz_i z_i^\dagger - I_d\|_F}{z_i^\dagger A z_i} \cdot \frac{L(\boldsymbol{z}_{\sim i})}{L(\boldsymbol{z})} \leqslant 4dD \ll n^{1/2}d\gamma$$

as desired. Otherwise $t \geqslant D$. We will prove by induction on $a \geqslant 0$ that if $S \subseteq [t]$ satisfies $|S| = t - D + a$, then

$$\|H_S\|_F \leqslant \xi_a \triangleq 2n^{1/2}d\gamma + 4e^{-a}nd^{1/2}.$$

The base case $a = 0$ holds by Lemma 5.3.13. For the inductive step, assume $a \geqslant 1$.

By the inductive hypothesis and equations (5.7) and (5.11), for all $i \in S$

$$\left| \frac{L(\boldsymbol{z}_S)}{L(\boldsymbol{z}_{S\setminus\{i\}})} - 1 \right| \leqslant \frac{2\varepsilon^2}{d^2} \cdot \| \frac{H_{S\setminus i}}{z_i^\dagger A z_i} \|_{\mathrm{op}} \leqslant \frac{4\varepsilon^2}{d^2} \|H_{S\setminus i}\|_F \leqslant \frac{4\varepsilon^2}{d^2} \xi_{a-1}.$$

Since this upper bound is $o(1)$, we also have

$$\left| \frac{L(\boldsymbol{z}_{S\setminus\{i\}})}{L(\boldsymbol{z}_S)} - 1 \right| \leqslant \frac{5\varepsilon^2}{d^2} \xi_{a-1}.$$

Write $\frac{L(\boldsymbol{z}_{S\setminus\{i\}})}{L(\boldsymbol{z}_S)} - 1 = \frac{5\varepsilon^2}{d^2}\xi_{a-1} b_i$ for $b_i \in [-1, 1]$. By Lemma 5.3.12,

$$\begin{aligned} \| \sum_{i \in S} \frac{d z_i z_i^\dagger - I_d}{z_i^\dagger A z_i} \cdot \left(\frac{L(\boldsymbol{z}_{S\setminus\{i\}})}{L(\boldsymbol{z}_S)} - 1 \right) \|_F &= \frac{5\varepsilon^2}{d^2} \xi_{a-1} \| \sum_{i \in S} \frac{d z_i z_i^\dagger - I_d}{z_i^\dagger A z_i} \cdot b_i \|_F \\ &\leqslant \left(\frac{5\varepsilon^2 n^{1/2}}{d} \gamma + \frac{10\varepsilon^2 n}{d^{3/2}} \right) \xi_{a-1} \leqslant e^{-1} \xi_{a-1}, \end{aligned}$$

using the hypotheses $\gamma \ll d/(\varepsilon^2 n^{1/2})$ and $n \ll d^{3/2}/\varepsilon^2$. By the triangle inequality, equation (5.11), and our choice of D,

$$\|K_S\|_F \leqslant \|K(\boldsymbol{z})\|_F + \sum_{i \in [t]\setminus S} \frac{\|d z_i z_i^\dagger - I_d\|_F}{z_i^\dagger A z_i} \leqslant n^{1/2} d\gamma + 2dD \leqslant \frac{101}{100} n^{1/2} d\gamma.$$

Hence

$$\begin{aligned} \|H_S\|_F &\leqslant \|K_S\|_F + \| \sum_{i \in S} \frac{d z_i z_i^\dagger - I_d}{z_i^\dagger A z_i} \cdot \left(\frac{L(\boldsymbol{z}_{S\setminus\{i\}})}{L(\boldsymbol{z}_S)} - 1 \right) \|_F \\ &\leqslant \frac{101}{100} n^{1/2} d\gamma + e^{-1} \xi_{a-1} \leqslant \xi_a, \end{aligned}$$

as $\frac{101}{100} + 2e^{-1} \leqslant 2$. This completes the induction. Finally,

$$\|H(\boldsymbol{z})\|_F = \|H_{[t]}\|_F \leqslant 2n^{1/2} d\gamma + 4e^{-D} n d^{1/2} = 2n^{1/2} d\gamma + 4n^{1/2} d \leqslant 3n^{1/2} d\gamma. \qquad \square$$

5.3.4 Uniform Frobenius Bound on the $K(\boldsymbol{x}_{\leqslant t})$ Matrix Martingale

In this subsection, we will prove Lemma 5.3.8. The proof mimics the proof of Doob's L^2 maximal inequality. Let $\boldsymbol{x} \sim p_0$, recall that $K_t = K(\boldsymbol{x}_{\leqslant t})$, and define $X = \sup_{1 \leqslant t \leqslant n} \|K_t\|_F$.

Lemma 5.3.14. *We have that* $\mathbb{E}\, X^2 \leqslant 4\, \mathbb{E}\, \|K_n\|_F^2$

Proof. We will first upper bound $\mathbf{Pr} * X \geqslant x$ for all $x > 0$. Consider the stopping time $\tau = \inf\{t : \|K_t\|_F \geqslant x\} \cup \{n\}$. Then,

$$\begin{aligned}
\mathbf{Pr} * X \geqslant x &= \mathbf{Pr} * \|K_\tau\|_F \geqslant x \\
&\leqslant x^{-1}\, \mathbb{E} * \|K_\tau\|_F I\{\|K_\tau\|_F \geqslant x\} \\
&\leqslant x^{-1}\, \mathbb{E} * \mathbb{E} * \|K_n\|_F | \mathcal{F}_\tau I\{\|K_\tau\|_F \geqslant x\} \\
&= x^{-1}\, \mathbb{E} * \|K_n\|_F I\{X \geqslant x\}.
\end{aligned}$$

The first estimate is by Markov's inequality, and the second is by convexity of the norm $\|\cdot\|_F$. Thus,

$$\begin{aligned}
\mathbb{E}\, X^2 &= \int_0^\infty \mathbf{Pr}\, X^2 \geqslant x \, \mathrm{d}x = \int_0^\infty \mathbf{Pr}\, X \geqslant x 2x \, \mathrm{d}x \leqslant \int_0^\infty 2\, \mathbb{E} * \|K_n\|_F I\{X \geqslant x\} \, \mathrm{d}x \\
&= 2\, \mathbb{E} * \|K_n\|_F X \leqslant 2\sqrt{\mathbb{E} * \|K_n\|_F^2\, \mathbb{E} * X^2}.
\end{aligned}$$

Rearranging yields the result. □

Lemma 5.3.15. *We have that* $\mathbb{E} * \|K_n\|_F^2 \lesssim nd^2$.

Proof. We can expand

$$\mathbb{E}\, \|K_n\|_F^2 = \sum_{i=1}^n \mathbb{E} * \| \frac{dx_i x_i^\dagger - I_d}{x_i^\dagger A x_i} \|_F^2 + 2 \sum_{1 \leqslant i < j \leqslant n} \mathbb{E} * \left\langle \frac{dx_i x_i^\dagger - I_d}{x_i^\dagger A x_i}, \frac{dx_j x_j^\dagger - I_d}{x_j^\dagger A x_j} \right\rangle. \quad (5.21)$$

Since

$$\mathbb{E}*\frac{dx_jx_j^\dagger - I_d}{x_j^\dagger Ax_j}|\mathcal{F}_{j-1} = \sum_{x_j}\omega_{x_j}(dx_jx_j^\dagger - I_d) = 0,$$

for any $i < j$ we have

$$\mathbb{E}*\left\langle \frac{dx_ix_i^\dagger - I_d}{x_i^\dagger Ax_i}, \frac{dx_jx_j^\dagger - I_d}{x_j^\dagger Ax_j}\right\rangle = \mathbb{E}*\left\langle \frac{dx_ix_i^\dagger - I_d}{x_i^\dagger Ax_i}, \mathbb{E}*\frac{dx_jx_j^\dagger - I_d}{x_j^\dagger Ax_j}|\mathcal{F}_{j-1}\right\rangle = 0.$$

The other expectation in (5.21) can be bounded by (recalling (5.11))

$$\mathbb{E}*\|\frac{dx_ix_i^\dagger - I_d}{x_i^\dagger Ax_i}\|_F^2 \leqslant 2\,\mathbb{E}*\frac{\langle dx_ix_i^\dagger - I_d, dx_ix_i^\dagger - I_d\rangle}{x_i^\dagger Ax_i} = 2d(d-1)\,\mathbb{E}*\frac{1}{x_i^\dagger Ax_i} = 2d(d-1).$$

Therefore $\mathbb{E}*\|K_n\|_F^2 \leqslant 2nd(d-1) \lesssim nd^2$. □

Proof of Lemma 5.3.8. Follows immediately from Lemmas 5.3.14 and 5.3.15. □

5.4 Lower Bound for Off-Diagonal Perturbations

In this section we consider the family of perturbations which correspond to the "off-diagonal" case. More formally, let $d_1 \geqslant d_2$ and $A \in \mathbb{R}^{d_1\times d_1}$ and $B \in \mathbb{R}^{d_2\times d_2}$ be diagonal matrices with diagonal entries $a_1 \geqslant \cdots \geqslant a_{d_1} > 0$ and $b_1 \geqslant \cdots \geqslant b_{d_2} > 0$ satisfying $2a_{d_1} \geqslant a_1$, $2b_{d_2} \geqslant b_1$, and $\operatorname{tr}(A) + \operatorname{tr}(B) = 1$. We abbreviate $a_{d_1} = a$, $b_{d_2} = b$. With these settings, we consider the task of distinguishing between the following two alternatives:

$$H_0 : \rho = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} \qquad \text{and} \qquad H_1 : \rho = \begin{pmatrix} A & \frac{\varepsilon}{d_2}\overline{G} \\ \frac{\varepsilon}{d_2}\overline{G}^\dagger & B \end{pmatrix}. \tag{5.22}$$

Here, $\overline{G} \sim \mathrm{Gin}_U(d_1, d_2)$ for the U given by Lemma 5.4.1 below.

Lemma 5.4.1. *For $d_1 \geqslant d_2$, there exists $U \subseteq \mathbb{R}^{d_1\times d_2}$ such that if $G \sim \mathrm{Gin}(d_1, d_2)$, then $\mathbf{Pr}\, G \notin U \leqslant \exp(-0.1d_1)$ and on the event $G \in U$, we have $\|G\|_{\mathrm{op}} \leqslant 3$ and*

$\|M\|_1 \geqslant d_2/3$ *for*

$$M = \begin{pmatrix} 0 & G \\ G^\dagger & 0 \end{pmatrix}.$$

We defer the proof of this lemma to Appendix C.3.

Parameter Settings. We will assume the parameters $a, b, d_1, d_2, \varepsilon$ satisfy the following relations:

$$d_1 \gg 1 \qquad \varepsilon \leqslant \frac{1}{10^{20}} \frac{d_2\sqrt{ab}}{\log \frac{1}{a}} \qquad d_1\sqrt{a} \leqslant d_2\sqrt{b} \qquad d_1 \geqslant d_2 \tag{5.23}$$

Remark 5.4.2. For most places, it suffices to use $\varepsilon \leqslant \frac{1}{10^{20}} d_2\sqrt{ab}$ so we will often drop the $\log(1/a)$ except for the few places where it is actually necessary.

Our main result for the distinguishing task (5.22) is the following.

Theorem 5.4.3. *Under the assumed parameter settings, the copy complexity of distinguishing between H_0 and H_1 with incoherent measurements is $\Omega(d_1^{1/2} d_2/\varepsilon^2)$.*

We first record several elementary consequences of the parameters settings.

Fact 5.4.4. *Under the parameter settings, $\rho \sim H_1$ is psd with trace distance at least $\varepsilon/3$ to H_0.*

Proof. The trace distance bound is immediate from the properties of U given by Lemma 5.4.1. To show ρ is psd, note that for any nonzero $x \in \mathbb{C}^{d_1}$, $y \in \mathbb{C}^{d_2}$,

$$\begin{aligned}(x,y)^\dagger \rho (x,y) &= x^\dagger A x + y^\dagger B y + \frac{2\varepsilon}{d_2} x^\dagger \overline{G} y \geqslant a\|x\|^2 + b\|y\|^2 - \frac{6\varepsilon}{d_2}\|x\|\|y\| \\ &\geqslant a\|x\|^2 + b\|y\|^2 - \frac{6}{10^{20}}\sqrt{ab}\|x\|\|y\| > 0.\end{aligned}$$

□

Fact 5.4.5. *Under the parameter settings, $bd_2 \in [1/4, 1]$.*

Proof. Since $d_1 \geqslant d_2$ and $d_1\sqrt{a} \leqslant d_2\sqrt{b}$, we have $d_1 a \leqslant \frac{d_2}{d_1} d_2 b \leqslant d_2 b$. Thus $1 = \mathrm{tr}(A) + \mathrm{tr}(B) \leqslant 2d_1 a + 2d_2 b \leqslant 4d_2 b$ and $1 \geqslant \mathrm{tr}(B) \geqslant d_2 b$. □

Take any learning tree $\mathcal{T}$ corresponding to an algorithm for this task that uses $n \ll d_2 d_1^{1/2}/\varepsilon^2$ incoherent measurements. The parameter settings imply that we may further assume, by taking additional superfluous measurements, that $(\log \frac{n}{\sqrt{ab}})^2/(d_1 a) \leqslant n$. Similarly to the previous section, we let p_0 and p_1 denote the distributions over leaves of $\mathcal{T}$ induced by ρ under H_0 and H_1 respectively, and we will show $d_{\mathrm{TV}}(p_0, p_1) \to 0$.

Because of the block structure in (5.22), we denote leaves of $\mathcal{T}$ by $(\boldsymbol{x}, \boldsymbol{y}) = ((x_1, y_1), \ldots, (x_n, y_n))$. Here, each (x_i, y_i) satisfies $x_i \in \mathbb{C}^{d_1}$, $y_i \in \mathbb{C}^{d_2}$, and $(x_i, y_i) \in \mathbb{C}^{d_1+d_2}$, and corresponds to an outcome from some (adaptively chosen) rank-1 POVM which we write as

$$\{(x, y)(x, y)^\dagger\}_{(x,y)\in\mathcal{P}}.$$

Note that (x, y) are not necessarily unit vectors. We only require that

$$\sum_{(x,y)\in\mathcal{P}} (x, y)(x, y)^\dagger = I_{d_1+d_2}\,.$$

We let $L^*(\cdot)$ denote the likelihood ratio between p_1 and p_0, i.e. $L^*((\boldsymbol{x}, \boldsymbol{y})) \triangleq p_1((\boldsymbol{x}, \boldsymbol{y}))/p_0((\boldsymbol{x}, \boldsymbol{y}))$. Note that

$$L^*((\boldsymbol{x}, \boldsymbol{y})) = \mathbb{E}[\overline{G} \sim \mathrm{Gin}_U(d_1, d_2)] * \prod_{i=1}^{n} \left(1 + \frac{2\varepsilon}{d_2} \cdot \frac{x_i^\dagger \overline{G} y_i}{x_i^\dagger A x_i + y_i^\dagger B y_i}\right)$$

We similarly define the non-truncated estimate

$$L((\boldsymbol{x}, \boldsymbol{y})) = \mathbb{E}[G \sim \mathrm{Gin}(d_1, d_2)] * \prod_{i=1}^{n} \left(1 + \frac{2\varepsilon}{d_2} \cdot \frac{x_i^\dagger \overline{G} y_i}{x_i^\dagger A x_i + y_i^\dagger B y_i}\right). \tag{5.24}$$

We will abuse notation and write $L((\boldsymbol{z}, \boldsymbol{w}))$ for any sequence of unit vectors $(\boldsymbol{z}, \boldsymbol{w}) = ((z_1, w_1), \ldots, (z_t, w_t))$ of length not necessarily n. This is defined identically to (5.24). We let $L((\boldsymbol{x}, \boldsymbol{y}), (\boldsymbol{x}, \boldsymbol{y}))$ denote the value of L on input $((x_1, y_1), (x_1, y_1), \ldots, (x_n, y_n), (x_n, y_n))$.

The following proposition is analogous to Proposition 5.3.4 and will be the main ingredient in our proof.

Proposition 5.4.6. *There exists a subset S of the leaves of $\mathcal{T}$ such that* $\mathbf{Pr}[p_0]S = 1 - o(1)$ *and for all* $(\boldsymbol{x}, \boldsymbol{y}) \in S$, $|L((\boldsymbol{x}, \boldsymbol{y})) - 1| = o(1)$ *and* $L((\boldsymbol{x}, \boldsymbol{y}), (\boldsymbol{x}, \boldsymbol{y})) \leqslant e^{0.02\sqrt{d_1 d_2}}$.

We now prove Theorem 5.4.3 assuming Proposition 5.4.6.

Proof of Theorem 5.4.3. Analogous to the proof of Theorem 5.3.3 assuming Proposition 5.3.4. □

5.4.1 Recursive Evaluation of Likelihood Ratio

Similarly to the previous section, we obtain a recursive expression for L. Let the sequence of unit vectors $(\boldsymbol{z}, \boldsymbol{w}) = ((z_1, w_1), \ldots, (z_t, w_t))$ be as above. For $1 \leqslant i \leqslant t$, let $(\boldsymbol{z}, \boldsymbol{w})_{\sim i}$ be this sequence with (z_i, w_i) omitted. Similarly, for $1 \leqslant i < j \leqslant t$, let $(\boldsymbol{z}, \boldsymbol{w})_{\sim i,j}$ be this sequence with (z_i, w_i) and (z_j, w_j) omitted.

Lemma 5.4.7. *The function L satisfies*

$$L((\boldsymbol{z}, \boldsymbol{w})) = L((\boldsymbol{z}, \boldsymbol{w})_{\sim t}) + \frac{4\varepsilon^2}{d_1 d_2^2} \sum_{i=1}^{t-1} \left[\frac{\langle z_i, z_t \rangle \langle w_i, w_t \rangle}{(z_i^\dagger A z_i + w_i^\dagger A w_i)(z_t^\dagger A z_t + w_t^\dagger A w_t)} \cdot L((\boldsymbol{z}, \boldsymbol{w})_{\sim i,t}) \right].$$

Proof. Analogous to Lemma 5.3.5. The pairwise moments are evaluated by

$$\mathbb{E}[G \sim \mathrm{Gin}(d_1, d_2)](x^\dagger G y)(z^\dagger G w) = \frac{1}{d_1} \langle x, z \rangle \langle y, w \rangle.$$ □

5.4.2 High Probability Bound on Likelihood Ratio at Leaves

This subsection gives the main part of the proof of Proposition 5.4.6. For the sequence of unit vectors $(\boldsymbol{z}, \boldsymbol{w}) = ((z_1, w_1), \ldots, (z_t, w_t))$ as above, define

$$H((\boldsymbol{z}, \boldsymbol{w})) = \sum_{i=1}^{t} \frac{z_i w_i^\dagger}{z_i^\dagger A z_i + w_i^\dagger B w_i} \cdot \frac{L((\boldsymbol{z}, \boldsymbol{w})_{\sim i})}{L((\boldsymbol{z}, \boldsymbol{w}))}.$$

Lemma 5.4.7 can be rewritten as

$$\frac{L((\boldsymbol{z},\boldsymbol{w}))}{L((\boldsymbol{z},\boldsymbol{w})_{\sim t})} = 1 + \frac{4\varepsilon^2}{d_1 d_2^2} \cdot \frac{z_t^\dagger H((\boldsymbol{z},\boldsymbol{w})_{\sim t}) w_t}{z_t^\dagger A z_t + w_t^\dagger B w_t}. \tag{5.25}$$

Further define

$$K((\boldsymbol{z},\boldsymbol{w})) = \sum_{i=1}^{t} \frac{z_i w_i^\dagger}{z_i^\dagger A z_i + w_i^\dagger B w_i} \text{ and } \kappa((\boldsymbol{z},\boldsymbol{w})) = \sup_{b_1,\dots,b_t \in [-1,1]} \|\sum_{i=1}^{t} b_i \frac{z_i w_i^\dagger \|z_i\| \|w_i\|}{(z_i^\dagger A z_i + w_i^\dagger B w_i)^2}\|_F.$$

As in the previous section, K will be our proxy for H. The condition that the error terms in the bootstrapping argument contract correspond to an upper bound on $\kappa((\boldsymbol{z},\boldsymbol{w}))$. In contrast to Lemma 5.3.12, it is no longer true in this setting that boundedness of $K((\boldsymbol{z},\boldsymbol{w}))$ implies the required bound on $\kappa((\boldsymbol{z},\boldsymbol{w}))$ (see Appendix C.4); this will instead be separately proved in Lemma 5.4.10 below.

The following three lemmas are the analogs of Lemmas 5.3.7 and 5.3.8. Lemma 5.4.8 deterministically controls H given bounds on K and κ, and Lemmas 5.4.9 and 5.4.10 give the required high probability bounds on $K((\boldsymbol{x},\boldsymbol{y}))$ and $\kappa((\boldsymbol{x},\boldsymbol{y}))$.

Lemma 5.4.8. *Suppose $\gamma \gg 1$. If $(\boldsymbol{z},\boldsymbol{w}) = ((z_1,w_1),\dots,(z_t,w_t))$ satisfies $t \leqslant n$, $\|K((\boldsymbol{z},\boldsymbol{w}))\|_F \leqslant \sqrt{nd_1d_2}\gamma$ and $\kappa((\boldsymbol{z},\boldsymbol{w})) \leqslant \frac{1}{10^3} d_1 d_2^2/\varepsilon^2$, then $\|H((\boldsymbol{z},\boldsymbol{w}))\|_F \leqslant 3\sqrt{nd_1d_2}\gamma$.*

Lemma 5.4.9. *For $(\boldsymbol{x},\boldsymbol{y}) \sim p_0$, let $(\boldsymbol{x},\boldsymbol{y})_{\leqslant t} = ((x_1,y_1),\dots,(x_t,y_t))$ be the length-t prefix of $(\boldsymbol{x},\boldsymbol{y})$. Then $\mathbb{E} * \sup_{1\leqslant t\leqslant n} \|K((\boldsymbol{x},\boldsymbol{y})_{\leqslant t})\|_F^2 \lesssim nd_1d_2$.*

Lemma 5.4.10. *If $(\boldsymbol{x},\boldsymbol{y}) \sim p_0$, then $\mathbf{Pr}\, \kappa((\boldsymbol{x},\boldsymbol{y})) > \frac{1}{10^3} d_1 d_2^2/\varepsilon^2 = o(1)$.*

We now prove Proposition 5.4.6 assuming Lemmas 5.4.8, 5.4.9, and 5.4.10. These lemmas will be proved in Subsections 5.4.3, 5.4.4, and 5.4.5.

Let α, β be slowly-growing functions with $1 \ll \alpha \ll \beta \ll d_1^{1/2} d_2/(\varepsilon^2 n)$ and furthermore $\alpha^2 \ll d_1^{1/2} d_2/(\varepsilon^2 n)$. This is possible because $n \ll d_1^{1/2} d_2/\varepsilon^2$. Let $(\boldsymbol{x},\boldsymbol{y}) \sim p_0$. For

$1 \leqslant t \leqslant n$, define the filtration $\mathcal{F}_t = \sigma((\boldsymbol{x}, \boldsymbol{y})_{\leqslant t})$ and the sequences

$$H_t = H((\boldsymbol{x}, \boldsymbol{y})_{\leqslant t}), \qquad K_t = K((\boldsymbol{x}, \boldsymbol{y})_{\leqslant t}), \qquad \kappa_t = \kappa((\boldsymbol{x}, \boldsymbol{y})_{\leqslant t}), \qquad \Phi_t = L((\boldsymbol{x}, \boldsymbol{y})_{\leqslant t}).$$

Consider the stopping time (with respect to $\mathcal{F}_t$)

$$\tau = \inf \left\{ t : \|K_t\|_F > \sqrt{nd_1d_2}\alpha \text{ or } \kappa_t > \frac{1}{10^3} d_1 d_2^2 \text{ or } |\Phi_t - 1| > \frac{\varepsilon^2 n}{d_1^{1/2} d_2} \beta \right\} \cup \{\infty\}$$

and stopped sequence $\Psi_t = \Phi_{t \wedge \tau}$.

Claim 5.4.11. With probability $1 - o(1)$, $\|K_t\|_F \leqslant \sqrt{nd_1d_2}\alpha$ for all $1 \leqslant t \leqslant n$.

Proof. Follows from Lemma 5.4.9 and Markov's inequality. □

Claim 5.4.12. With probability $1 - o(1)$, $|\Psi_n - 1| \leqslant \frac{\varepsilon^2 n}{d_1^{1/2} d_2} \beta$.

Proof. This is analogous to Claim 5.3.10, and we only sketch the differences. Note that Ψ_t is a multiplicative martingale. We will bound the quadratic increment $\mathbb{E}\,(\frac{\Psi_t}{\Psi_{t-1}})^2 | \mathcal{F}_{t-1}$. This is 1 if $\tau \leqslant t-1$, and otherwise by (5.25), (because the linear term expects to 0)

$$\mathbb{E} * \left(\frac{\Psi_t}{\Psi_{t-1}} \right)^2 | \mathcal{F}_{t-1} = 1 + \frac{16\varepsilon^2}{d_1^2 d_2^4} \, \mathbb{E} * \frac{(x_t^\dagger H_{t-1} y_t)^2}{(x_t^\dagger A x_t + y_t^\dagger B y_t)^2} | \mathcal{F}_{t-1}.$$

This last expectation is bounded by

$$\begin{aligned} \mathbb{E} * \frac{(x_t^\dagger H_{t-1} y_t)^2}{(x_t^\dagger A x_t + y_t^\dagger B y_t)^2} | \mathcal{F}_{t-1} &= \mathbb{E} * \frac{x_t^\dagger H_{t-1} y_t y_t^\dagger H_{t-1} x_t}{(x_t^\dagger A x_t + y_t^\dagger B y_t)^2} | \mathcal{F}_{t-1} \leqslant \mathbb{E} * \frac{\|y_t\|^2 x_t^\dagger H_{t-1}^2 x_t}{(x_t^\dagger A x_t + y_t^\dagger B y_t)^2} | \mathcal{F}_{t-1} \\ &\leqslant \frac{1}{b} \, \mathbb{E} * \frac{x_t^\dagger H_{t-1}^2 x_t}{x_t^\dagger A x_t + y_t^\dagger B y_t} | \mathcal{F}_{t-1} = \frac{\|H_{t-1}\|_F^2}{b} \leqslant 4d_2 \|H_{t-1}\|_F^2 \end{aligned}$$

using Fact 5.4.5. Since $\tau > t-1$, we have $\|K_{t-1}\|_F \leqslant \sqrt{nd_1d_2}\alpha$ and $\kappa_t \leqslant \frac{1}{10} d_1 d_2^2$.

Thus Lemma 5.4.8 implies $\|H_{t-1}\|_F \leqslant 3\sqrt{nd_1d_2}\alpha$, and

$$\mathbb{E} * \left(\frac{\Psi_t}{\Psi_{t-1}}\right)^2 |\mathcal{F}_{t-1} \leqslant 1 + \frac{64\varepsilon^4}{d_1^2 d_2^3}\|H_{t-1}\|_F^2 \leqslant 1 + \frac{576\varepsilon^4 n}{d_1 d_2^2}\alpha^2.$$

Analogously to the proof of Claim 5.3.10, this implies

$$\mathbb{E}\,(\Psi_n - 1)^2 \leqslant \frac{2 \cdot 576\varepsilon^4 n^2}{d_1 d_2^2}\alpha^2.$$

The result now follows from Markov's inequality. □

Claim 5.4.13. If $\|K_n\|_F \leqslant \sqrt{nd_1d_2}\alpha$ and $\kappa_n \leqslant \frac{1}{10^3}\frac{d_1d_2^2}{\varepsilon^2}$, then $L((\boldsymbol{x},\boldsymbol{y}),(\boldsymbol{x},\boldsymbol{y})) \leqslant e^{0.02\sqrt{d_1d_2}}$.

Proof. Using the elementary inequality $e^z \geqslant 1 + z$ and then Cauchy-Schwarz, we can upper bound $L((\boldsymbol{x},\boldsymbol{y}),(\boldsymbol{x},\boldsymbol{y}))$ by

$$\begin{aligned} L((\boldsymbol{x},\boldsymbol{y}),(\boldsymbol{x},\boldsymbol{y})) &\leqslant \mathbb{E}[G \sim \mathrm{Gin}(d_1,d_2)] * \exp\left(\sum_{i=1}^n \frac{4\varepsilon}{d_2}\frac{x_i^\dagger G y_i}{x_i^\dagger A x_i + y_i^\dagger B y_i} + \left(\frac{2\varepsilon}{d_2}\frac{x_i^\dagger G y_i}{x_i^\dagger A x_i + y_i^\dagger B y_i}\right)^2\right) \\ &\leqslant \sqrt{\mathbb{E}[G \sim \mathrm{Gin}(d_1,d_2)] * \exp\left(\sum_{i=1}^n \frac{8\varepsilon}{d_2}\frac{x_i^\dagger G y_i}{x_i^\dagger A x_i + y_i^\dagger B y_i}\right)} \\ &\quad \times \sqrt{\mathbb{E}[G \sim \mathrm{Gin}(d_1,d_2)] * \exp\left(\sum_{i=1}^n \frac{8\varepsilon^2}{d_2^2}\left(\frac{x_i^\dagger G y_i}{x_i^\dagger A x_i + y_i^\dagger B y_i}\right)^2\right)}. \end{aligned} \tag{5.26}$$

Now, we bound the two terms in the last product in (5.26) separately. First, we have

$$\mathbb{E}[G \sim \mathrm{Gin}(d_1,d_2)] * \exp\left(\sum_{i=1}^n \frac{8\varepsilon}{d_2}\frac{x_i^\dagger G y_i}{x_i^\dagger A x_i + y_i^\dagger B y_i}\right) = \mathbb{E}[G \sim \mathrm{Gin}(d_1,d_2)] * \exp\left(\frac{8\varepsilon}{d_2}\langle G, K_n\rangle\right). \tag{5.27}$$

Note that $\frac{8\varepsilon}{d_2}\langle G, K_n\rangle$ is distributed as a Gaussian with variance $\frac{64\varepsilon^2}{d_1d_2^2}\|K_n\|_F^2 \leqslant \frac{64\varepsilon^2 n}{d_2}\alpha^2$.

So we can bound (5.27) by

$$\mathbb{E}[g \sim \mathcal{N}(0, 64\varepsilon^2 n\alpha^2/d_2))] * \exp(g) = e^{64\varepsilon^2 n\alpha^2/d_2} \ll e^{\sqrt{d_1}}, \tag{5.28}$$

as $\alpha^2 \ll d_1^{1/2} d_2/(\varepsilon^2 n)$ by assumption. Next, we bound the second term in the product in (5.26). Use $\mathsf{vec}(G)$ to denote rearranging G as a vector in $\mathbb{R}^{d_1 d_2}$ (done in a consistent way) and use $\otimes$ to denote the Kronecker product of two matrices. Define $Q \in \mathbb{R}^{d_1 d_2 \times d_1 d_2}$ as

$$Q = \sum_{i=1}^{n} \frac{x_i x_i^\dagger \otimes y_i y_i^\dagger}{(x_i^\dagger A x_i + y_i^\dagger B y_i)^2}\,.$$

Let $X \in \mathbb{C}^{d_1 d_2 \times n}$ be the matrix with columns given by $\frac{\mathsf{vec}(x_i y_i^\dagger)}{\|x_i\|\|y_i\|}$ for $i = 1, 2, \ldots, n$. Let

$$\theta_i = \frac{\|x_i\|^2 \|y_i\|^2}{(x_i^\dagger A x_i + y_i^\dagger B y_i)^2}$$

and let $D \in \mathbb{R}^{n \times n}$ be the diagonal matrix whose diagonal entries are θ_i for $i = 1, 2, \ldots, n$. We can write

$$\|Q\|_F^2 = \sum_{i,j} \theta_i \theta_j \left\langle \frac{x_i y_i^\dagger}{\|x_i\|\|y_i\|}, \frac{x_j y_j^\dagger}{\|x_j\|\|y_j\|} \right\rangle^2 = \langle DX^\dagger X D, X^\dagger X \rangle\,.$$

By Grothendieck's inequality, we can replace the second term $X^\dagger X$ with $\sigma^\dagger \sigma$ for some $\sigma \in [-1, 1]^n$ while incurring at most a factor of 2 in the inequality. Thus, we have

$$\|Q\|_F \leqslant \sqrt{2} \max_{\sigma \in [-1,1]^n} \|\sum_{i=1}^{n} \frac{\sigma_i x_i y_i^\dagger \|x_i\|\|y_i\|}{(x_i^\dagger A x_i + y_i^\dagger B y_i)^2}\|_F \leqslant \sqrt{2}\kappa_n \leqslant \frac{d_1 d_2^2}{700\varepsilon^2}\,.$$

In particular, we have $\|Q\|_{\mathrm{op}} \leqslant (d_1 d_2^2)/(700\varepsilon^2)$ and $\|Q\|_1 \leqslant d_1^{3/2} d_2^{5/2}/(700\varepsilon^2)$. Let $\lambda_1, \ldots, \lambda_{d_1 d_2}$ be the eigenvalues of Q. Returning to the last term in (5.26), we can write

$$\mathbb{E}[G \sim \mathrm{Gin}(d_1, d_2)] * \exp\left(\sum_{i=1}^{n} \frac{8\varepsilon^2}{d_2^2} \left(\frac{x_i^\dagger G y_i}{x_i^\dagger A x_i + y_i^\dagger B y_i}\right)^2\right) \tag{5.29}$$

$$= \mathbb{E}[G \sim \mathrm{Gin}(d_1, d_2)] * \exp\left(\frac{8\varepsilon^2}{d_2^2}\mathsf{vec}(G)^\dagger Q \mathsf{vec}(G)\right) \tag{5.30}$$

$$= \mathbb{E}[v \sim N(0, I_{d_1 d_2})] * \exp\left(\frac{8\varepsilon^2}{d_1 d_2^2} v^\dagger Q v\right) \tag{5.31}$$

$$= \prod_{j=1}^{d_1 d_2} \mathbb{E}[g \sim N(0,1)] * \exp\left(\frac{8\varepsilon^2 g^2}{d_1 d_2^2}\lambda_j\right) \tag{5.32}$$

$$\leqslant e^{10\varepsilon^2(\lambda_1 + \cdots + \lambda_{d_1 d_2})/(d_1 d_2^2)} \tag{5.33}$$

$$\leqslant e^{\sqrt{d_1 d_2}/70}, \tag{5.34}$$

where in the fourth step we used our bound on $\|Q\|_{\mathrm{op}}$ together with the fact that $\mathbb{E}[g]e^{cx^2} = (1-2c)^{-1} \leqslant e^{5c/4}$ for sufficiently small c, and in the last step we used our bound on $\|Q\|_1$.

Putting (5.26), (5.27), 5.28, and (5.34) together, we conclude $L((\boldsymbol{x}, \boldsymbol{y}), (\boldsymbol{x}, \boldsymbol{y})) \ll e^{\sqrt{d_1 d_2}}$ as desired. □

Proof of Proposition 5.4.6. Define the event

$$S = \left\{ \sup_{1 \leqslant t \leqslant n} \|K_t\|_F \leqslant \sqrt{n d_1 d_2}\alpha \text{ and } \kappa_n \leqslant \frac{1}{10^3}\frac{d_1 d_2^2}{\varepsilon^2} \text{ and } \Psi_n \leqslant \frac{\varepsilon^2 n}{d_1^{1/2} d_2}\beta \right\}.$$

By Lemma 5.4.10 and Claims 5.4.11 and 5.4.12, $\mathbf{Pr}[p_0]S = 1 - o(1)$. If S holds, then $\tau = \infty$: indeed, if $\tau = t < \infty$, then one of $\|K_t\|_F > \sqrt{n d_1 d_2}\alpha$, $\kappa_t > \frac{1}{10^3}\frac{d_1 d_2^2}{\varepsilon^2}$, and $|\Phi_t - 1| > \frac{\varepsilon^2 n}{d_1^{1/2} d_2}\beta$ holds. Since $\Psi_n = \Phi_t$ and $\kappa_n \geqslant \kappa_t$ (because in the definition of $\kappa((\boldsymbol{x}, \boldsymbol{y}))$ we can take $b_{t+1} = \cdots = b_n = 0$) this contradicts S.

So, $\tau = \infty$ on S. This implies $|L((\boldsymbol{x}, \boldsymbol{y})) - 1| = |\Phi_n - 1| \leqslant \frac{\varepsilon^2 n}{d_1^{1/2} d_2}\beta = o(1)$. Moreover, $\|K_n\|_F \leqslant \sqrt{n d_1 d_2}\alpha$, so by Claim 5.4.13 we have $L((\boldsymbol{x}, \boldsymbol{y}), (\boldsymbol{x}, \boldsymbol{y})) \leqslant e^{0.02\sqrt{d_1 d_2}}$. □

5.4.3 Bounding H in Frobenius Norm by Bootstrapping

In this section, we will prove Lemma 5.4.8. Let $(\boldsymbol{z}, \boldsymbol{w}) = ((z_1, w_1), \ldots, (z_t, w_t))$ be a sequence of unit vectors satisfying $t \leqslant n$, $\|K((\boldsymbol{z}, \boldsymbol{w}))\|_F \leqslant \sqrt{n d_1 d_2}\alpha$, and

$\kappa((\boldsymbol{z},\boldsymbol{w})) \leqslant \frac{1}{10} d_1 d_2^2$, where $\gamma \gg 1$. For $S \subseteq [t]$, let $(\boldsymbol{z},\boldsymbol{w})_S = ((z_i, w_i))_{i\in S}$. Let

$$H_S = \sum_{i\in S} \frac{z_i w_i^\dagger}{z_i^\dagger A z_i + w_i^\dagger B w_i} \cdot \frac{L((\boldsymbol{z},\boldsymbol{w})_{S\setminus i})}{L((\boldsymbol{z},\boldsymbol{w})_S)} \qquad \text{and} \qquad K_S = \sum_{i\in S} \frac{z_i w_i^\dagger}{z_i^\dagger A z_i + w_i^\dagger B w_i}.$$

Lemma 5.4.14. *For all $S \subseteq [t]$, $\|H_S\|_F \leqslant n/\sqrt{ab}$.*

Proof. For any fixed $\overline{G} \in U$, for the U given by Lemma 5.4.1, and any unit vector (z, w),

$$\frac{2\varepsilon}{d_2} \left| \frac{z^\dagger \overline{G} w}{z^\dagger A z + w^\dagger B w} \right| \leqslant \frac{2\varepsilon}{d_2} \cdot \frac{3\|z\|\|w\|}{a\|z\|^2 + b\|w\|^2} \leqslant \frac{3\varepsilon}{d_2\sqrt{ab}} \leqslant \frac{3}{10^6}.$$

Thus, for all i, $L((\boldsymbol{z},\boldsymbol{w})_S)/L((\boldsymbol{z},\boldsymbol{w})_{S\setminus i}) \in [1/(1+10^{-5}), 1+10^{-5}]$ which implies

$$\frac{L((\boldsymbol{z},\boldsymbol{w})_{S\setminus i})}{L((\boldsymbol{z},\boldsymbol{w})_S)} \in \left[\frac{1}{1+10^{-5}}, 1+10^{-5}\right]. \tag{5.35}$$

Thus

$$\|H_S\|_F \leqslant \sum_{i\in S} \|\frac{z_i w_i^\dagger}{z_i^\dagger A z_i + w_i^\dagger B w_i}\|_F \cdot (1+10^{-5}) \leqslant \frac{n}{2\sqrt{ab}} \cdot (1+10^{-5}) \leqslant \frac{n}{\sqrt{ab}}. \qquad \square$$

Proof of Lemma 5.4.8. Let $D = \log(n/\sqrt{ab})$. If $t < D$, then by equation (5.35) and the assumption $(\log \frac{n}{\sqrt{ab}})^2/(d_1 a) \leqslant n$,

$$\|H((\boldsymbol{z},\boldsymbol{w}))\|_F \leqslant \sum_{i=1}^{t} \|\frac{z_i w_i^\dagger}{z_i^\dagger A z_i + w_i^\dagger B w_i}\|_F \cdot (1+10^{-5}) \leqslant \frac{D}{\sqrt{ab}} \leqslant 3\sqrt{n d_1 d_2}\gamma.$$

Otherwise $t \geqslant D$. We will prove by induction on $a \geqslant 0$ that if $S \subseteq [t]$ satisfies $|S| = t - D + a$, then

$$\|H_S\|_F \leqslant \xi_a \triangleq 2\sqrt{n d_1 d_2}\gamma + e^{-a}\frac{n}{\sqrt{ab}}.$$

The base case $a = 0$ holds by Lemma 5.4.14. For the inductive step, assume $a \geqslant 1$.

By the inductive hypothesis and equation (5.25),

$$\left|\frac{L((\boldsymbol{z},\boldsymbol{w})_S)}{L((\boldsymbol{z},\boldsymbol{w})_{S\setminus i})}-1\right| \leqslant \frac{4\varepsilon^2}{d_1d_2^2}\cdot\left|\frac{z_i^\dagger H_{S\setminus i}w_i}{z_i^\dagger Az_i+w_i^\dagger Bw_i}\right| \leqslant \frac{4\varepsilon^2\xi_{a-1}}{d_1d_2^2}\cdot\frac{\|z_i\|\|w_i\|}{z_i^\dagger Az_i+w_i^\dagger Bw_i}.$$

Thus,

$$\|H_S\|_F \leqslant \|K_S\|_F + \|\sum_{i\in S}\frac{z_iw_i^\dagger}{z_i^\dagger Az_i+w_i^\dagger Bw_i}\cdot\left(\frac{L((\boldsymbol{z},\boldsymbol{w})_{S\setminus i})}{L((\boldsymbol{z},\boldsymbol{w})_S)}-1\right)\|_F$$

These terms are bounded by (in light of (5.35))

$$\|K_S\|_F \leqslant \|K((\boldsymbol{z},\boldsymbol{w}))\|_F + (1+10^{-5})\sum_{i\in[t]\setminus S}\|\frac{z_iw_i^\dagger}{z_i^\dagger Az_i+w_i^\dagger Bw_i}\|_F \tag{5.36}$$

$$\leqslant \sqrt{nd_1d_2}\gamma + \frac{D}{\sqrt{ab}} \leqslant 1.01\sqrt{nd_1d_2}\gamma \tag{5.37}$$

and, for some $b_1,\ldots,b_t\in[-1,1]$,

$$\begin{aligned}\|\sum_{i\in S}\frac{z_iw_i^\dagger}{z_i^\dagger Az_i+w_i^\dagger Bw_i}\cdot\left(\frac{L((\boldsymbol{z},\boldsymbol{w})_{S\setminus i})}{L((\boldsymbol{z},\boldsymbol{w})_S)}-1\right)\|_F &= \frac{4\varepsilon^2\xi_{a-1}}{d_1d_2^2}\|\sum_{i=1}^t b_i\frac{z_iw_i^\dagger\|z_i\|\|w_i\|}{(z_i^\dagger Az_i+w_i^\dagger Bw_i)^2}\|_F\\ &\leqslant \frac{4\varepsilon^2\kappa((\boldsymbol{z},\boldsymbol{w}))}{d_1d_2^2}\xi_{a-1} \leqslant e^{-1}\xi_{a-1}.\end{aligned}$$

Since $1.01+\frac{2}{e}\leqslant 2$, this implies $\|H_S\|\leqslant\xi_a$. Therefore

$$\|H((\boldsymbol{z},\boldsymbol{w}))\|_F = \|H_{[t]}\|_F \leqslant 2\sqrt{nd_1d_2}\gamma + e^{-D}\frac{n}{\sqrt{ab}} = 2\sqrt{nd_1d_2}\gamma+1 \leqslant 3\sqrt{nd_1d_2}\gamma.\square$$

5.4.4 Uniform Frobenius Bound on $K((\boldsymbol{x},\boldsymbol{y})_{\leqslant t})$

In this subsection, we will prove Lemma 5.4.9. Let $(\boldsymbol{x},\boldsymbol{y})\sim p_0$, $K_t = K((\boldsymbol{x},\boldsymbol{y})_{\leqslant t})$ and $X = \sup_{1\leqslant t\leqslant n}\|K_t\|_F$.

Lemma 5.4.15. *We have that* $\mathbb{E}\,X^2 \leqslant 4\,\mathbb{E}\,\|K_n\|_F^2$.

Proof. Analogous to Lemma 5.3.14. $\square$

Lemma 5.4.16. *We have that* $\mathbb{E}\,\|K_n\|_F^2 \lesssim nd_1d_2$.

Proof. We expand

$$\mathbb{E}\,\|K_n\|_F^2 = \sum_{i=1}^{n} \mathbb{E}\,*\|\frac{x_i y_i^\dagger}{x_i^\dagger A x_i + y_i^\dagger B y_i}\|_F^2 + 2\sum_{1\leqslant i<j\leqslant n} \mathbb{E}\,*\left\langle \frac{x_i y_i^\dagger}{x_i^\dagger A x_i + y_i^\dagger B y_i}, \frac{x_j y_j^\dagger}{x_j^\dagger A x_j + y_j^\dagger B y_j} \right\rangle$$

The cross terms have expectation 0 like in the proof of Lemma 5.3.15. Now,

$$\mathbb{E}\,*\|\frac{x_i y_i^\dagger}{x_i^\dagger A x_i + y_i^\dagger B y_i}\|_F^2 = \mathbb{E}\,*\frac{\|x_i\|^2\|y_i\|^2}{(x_i^\dagger A x_i + y_i^\dagger B y_i)^2} \leqslant \frac{1}{b}\,\mathbb{E}\,*\frac{\|x_i\|^2}{x_i^\dagger A x_i + y_i^\dagger B y_i} = \frac{d_1}{b} \lesssim d_1 d_2$$

in light of Fact 5.4.5. □

Proof of Lemma 5.4.9. Follows from Lemmas 5.4.15 and 5.4.16. □

5.4.5 Balancedness of Realizations

Finally, it remains to prove Lemma 5.4.10. Recall that

$$\kappa((\boldsymbol{x},\boldsymbol{y})) = \sup_{b_1,\ldots,b_n\in[-1,1]} \|\sum_{i=1}^{n} b_i \frac{x_i y_i^\dagger \|x_i\|\|y_i\|}{(x_i^\dagger A x_i + y_i^\dagger B y_i)^2}\|_F\,. \tag{5.38}$$

We will first prove a few preliminary inequalities about the individual terms in the summation above. In particular, since x_i, y_i are the measurements obtained from a POVM, we argue that over the randomness in the ith measurement, the term

$$\frac{x_i y_i^\dagger \|x_i\|\|y_i\|}{(x_i^\dagger A x_i + y_i^\dagger B y_i)^2}$$

is not too aligned with any given direction. Thus, intuitively, over $i = 1,\ldots,n$, the individual terms will be very weakly correlated and this will allow us to bound the signed sum. In the next two claims below, we think of $\mathcal{P}$ as a POVM that we will use to measure our state.

Claim 5.4.17. Let $\mathcal{P}$ be a set of vectors (not necessarily unit vectors) in $\mathbb{C}^d$ such that

$$\sum_{x\in\mathcal{P}} xx^\dagger = I\,.$$

Then for any coefficients c_x for all $x \in \mathcal{P}$, we have

$$\|\sum_{x\in\mathcal{P}} c_x x\| \leqslant \sqrt{\sum_{x\in\mathcal{P}} c_x^2}\,.$$

Proof. For any unit vector v,

$$\left\langle v, \sum_{x\in\mathcal{P}} c_x x \right\rangle \leqslant \sum_{x\in\mathcal{P}} c_x |\langle v, x\rangle| \leqslant \left(\sqrt{\sum_{x\in\mathcal{P}} c_x^2}\right)\left(\sqrt{\sum_{x\in\mathcal{P}} |\langle v, x\rangle|^2}\right) = \sqrt{\sum_{x\in\mathcal{P}} c_x^2}$$

where we used Cauchy-Schwarz and the hypothesis. Because $\|u\| = \max_{\|v\|=1}\langle v, u\rangle$ for all vectors u, we are done. □

Claim 5.4.18. Let $\mathcal{P} = \{(x, y)\}$ be a set of vectors where $x \in \mathbb{C}^{d_1}, y \in \mathbb{C}^{d_2}$ such that

$$\sum_{z=(x,y),z\in\mathcal{P}} zz^\dagger = I\,.$$

Then for any choice of $c_{x,y} \in [-1, 1]$ for each $(x, y) \in \mathcal{P}$, we have

$$\|\sum_{(x,y)\in\mathcal{P}} \frac{c_{x,y}x\|x\|\|y\|^2}{x^\dagger Ax + y^\dagger By}\| \leqslant \frac{\sqrt{d_1}}{b}\,.$$

Proof. For all choices of $c_{x,y} \in [-1, 1]$

$$\|\sum_{(x,y)\in\mathcal{P}} \frac{c_{x,y}x\|x\|\|y\|^2}{x^\dagger Ax + y^\dagger By}\| \leqslant \frac{1}{b}\|\sum_{(x,y)\in\mathcal{P}} c_{x,y}\|x\|x\|_F\,.$$

By Claim 5.4.17,

$$\|\sum_{(x,y)\in\mathcal{P}} c_{x,y}\|x\|x\|_F \leqslant \sqrt{\sum_{(x,y)\in\mathcal{P}} c_{x,y}^2\|x\|^2} \leqslant \sqrt{d_1}.$$ □

We can think of the bound in Claim 5.4.18 as a bound on a single term in (5.38) over the randomness of the measurement. Now, we will use Claim 5.4.18 on all of the terms in (5.38) to bound the signed sum. Of course, the sum in the expression in Claim 5.4.18 is inside the norm and it is not immediately clear how to use this to reason about the sum in (5.38). Actually relating the two expressions requires several additional arguments.

Claim 5.4.19. Consider measuring the state H_0 with respect to POVMs $\mathcal{P}_1, \dots, \mathcal{P}_n$ (which may be chosen adaptively). WLOG, the POVMs are rank-1 and can be viewed as sets of vectors such that for all $i \in [n]$

$$\sum_{z \in \mathcal{P}_i} zz^\dagger = I \,.$$

Let the results of the measurements be $(x_1, y_1), \dots, (x_n, y_n)$. If $n \leqslant d_2 d_1^{1/2}/(10^{20}\varepsilon^2)$. Then with probability $1 - e^{-5d_1}$ over the randomness in the measurements, we have the following inequality for any choice of $c_1, \dots, c_n \in [-1, 1]$:

$$\|\sum_{i=1}^{n} \frac{c_i x_i y_i^\dagger \|x_i\|\|y_i\|}{(x_i^\dagger A x_i + y_i^\dagger B y_i)^2}\|_F \leqslant \frac{d_1 d_2}{4 \cdot 10^3 b \varepsilon^2} \,.$$

Proof. For each $i \in [n]$, define

$$\theta_i = \frac{\|x_i\|\|y_i\|}{(x_i^\dagger A x_i + y_i^\dagger B y_i)^2} \,.$$

Let Q be the expression inside the Frobenius norm on the LHS of the desired inequality. Define the matrix $X \in \mathbb{C}^{d_1 \times n}$ to have columns given by $x_1, \dots, x_n$ and the matrix $Y \in \mathbb{C}^{d_2 \times n}$ to have columns given by $y_1, \dots, y_n$. Let $N, D \in \mathbb{R}^{n \times n}$ be the diagonal matrices whose entries are $\|y_1\|, \dots, \|y_n\|$ and $c_1\theta_1, \dots, c_t\theta_n$ respectively. Now we can rewrite

$$\|Q\|_F^2 = \sum_{i,j} c_i c_j \theta_i \theta_j \langle x_i, x_j \rangle \langle y_i, y_j \rangle = \left\langle (XD)^\dagger XD, Y^\dagger Y \right\rangle = \left\langle (XDN)^\dagger (XDN), (YN^{-1})^\dagger (YN^{-1}) \right\rangle \,.$$

Now by Grothendieck's inequality, we can replace the expression $(N^{-1}Y)^\dagger(N^{-1}Y)$ with $\sigma^\dagger\sigma$ for some sign vector $\sigma \in \{-1,1\}^n$ while incurring at most a factor of 2 loss. Thus,

$$\max_{c_i}\|Q\|_F^2 \leqslant 2\max_{c_i}\|\sum_{i=1}^n \frac{c_i x_i\|x_i\|\|y_i\|^2}{(x_i^\dagger A x_i + y_i^\dagger B y_i)^2}\|^2 \,. \tag{5.39}$$

Let R denote the quantity inside the norm on the RHS above. Now it suffices to bound $\|R\|$. First, consider any fixed unit vector v. Note that

$$\langle R, v\rangle \leqslant \sum_{i=1}^n \frac{\|x_i\|\|y_i\|^2}{(x_i^\dagger A x_i + y_i^\dagger B y_i)^2}|\langle x_i, v\rangle| \,.$$

Next, by Claim 5.4.18, we have that

$$\boldsymbol{E}\left[\frac{\|x_i\|\|y_i\|^2}{(x_i^\dagger A x_i + y_i^\dagger B y_i)^2}|\langle x_i, v\rangle|\right] \leqslant \frac{\sqrt{d_1}}{b}$$

where the randomness is over the ith measurement. Also note that the individual terms

$$\frac{\|x_i\|\|y_i\|^2}{(x_i^\dagger A x_i + y_i^\dagger B y_i)^2}|\langle x_i, v\rangle|$$

are always bounded in magnitude by $1/(ab)$. Note that the above two observations also imply that

$$\boldsymbol{E}\left[\left(\frac{\|x_i\|\|y_i\|^2}{(x_i^\dagger A x_i + y_i^\dagger B y_i)^2}|\langle x_i, v\rangle|\right)^2\right] \leqslant \frac{\sqrt{d_1}}{ab^2} \,.$$

Thus, by Freedman's inequality, we have

$$\langle R, v\rangle \leqslant \frac{d_1 d_2}{10^4 b\varepsilon^2}$$

with failure probability at most

$$\exp\left(-\frac{1}{2}\frac{\frac{d_1^2 d_2^2}{10^8 b^2\varepsilon^4}}{\frac{d_1 d_2}{10^{20}\varepsilon^2 ab^2} + \frac{d_1 d_2}{10^4 ab^2\varepsilon^2}}\right) \leqslant \exp\left(-\frac{d_1 d_2 a}{10^5\varepsilon^2}\right) \leqslant \exp(-10d_1) \,,$$

where in the second step we used that $\frac{d_2 a}{\varepsilon^2} \geqslant \frac{10^{40}}{d_2 b} \geqslant 10^{40}$. Finally, we can take some

0.1-net of possible choices of v, which we call Γ, and union bound over all elements of Γ. If the desired inequality were false i.e.

$$\|Q\|_F \geqslant \frac{d_1 d_2}{4 \cdot 10^3 b \varepsilon^2}$$

then by (5.39), we must have

$$\|R\| \geqslant \frac{d_1 d_2}{8 \cdot 10^3 b \varepsilon^2} .$$

By the construction of Γ, there must be some $v \in \Gamma$ such that

$$\langle R, v \rangle \geqslant 0.9 \|R\| > \frac{d_1 d_2}{10^4 b \varepsilon^2}$$

and thus we are done. □

Proof of Lemma 5.4.10. Plugging in the parameter settings at the beginning of Section 5.4, and using Claim 5.4.19, we get the desired property. □

5.5 Instance Near-Optimal Lower Bounds

Our main result on general state certification is the following. Recall that for two quantum states σ, ρ, the fidelity F between them is defined in Definition 1.4.7.

Theorem 5.5.1. *Let $0 < \varepsilon < \widetilde{O}(1/\log\log(d))$. Let $\sigma \in \mathbb{C}^{d \times d}$ be a density matrix. Then any algorithm that uses incoherent measurements which, given n copies of $\rho \in \mathbb{C}^{d \times d}$, can distinguish between the case where $\rho = \sigma$ and $\|\rho - \sigma\|_1 > \varepsilon$ with probability at least $2/3$ must satisfy*

$$n \geqslant \Omega\left(\frac{d\sqrt{d_{\mathsf{eff}}}}{\varepsilon^2 \operatorname{polylog}(d/\varepsilon)} \cdot F\left(\sigma^*, I_d/d\right)\right). \tag{5.40}$$

Here, σ^* is an explicit density matrix given by zeroing out $O(\varepsilon)$ mass from σ and normalizing, and d_{eff} is the rank of σ^* .

As before, the choice of failure probability can be taken to be any constant greater than 1/2.

In this section we use Theorems 5.3.3 and 5.4.3 to give a simple proof of a slightly weaker version of Theorem 5.5.1 where the construction of σ^* requires removing up to $O(\varepsilon \log(d/\varepsilon))$ mass. The analysis is a simplified version of the analysis from Sections 5.5 and A.3 of [65]. Later, in Appendix C.1 and C.2 we give a full proof of Theorem 5.5.1, which involves slightly generalizing Theorem 5.3.3 and carrying out a more delicate version of the analysis below.

As a first step, notice that since we are given an explicit description of σ, by applying an appropriate rotation, we may assume without loss of generality that σ is diagonal. For the remainder of this section, we will let $\sigma_1 \geqslant \ldots \geqslant \sigma_d$ be its eigenvalues (equivalently, its diagonal entries in sorted order).

5.5.1 Bucketing and Mass Removal

For $j \in \mathbb{Z}_{\geqslant 0}$, let S_j denote the set of indices $i \in [d]$ for which $2^{-j-1} < \sigma_i \leqslant 2^{-j}$, and define $d_j \triangleq |S_j|$. Let $\mathcal{J}$ denote the set of j for which $S_j \neq \emptyset$. We will refer to $j \in \mathcal{J}$ as *buckets*. Given $i \in [d]$, let $j(i)$ denote the index of the bucket for which $i \in S_j$.

Let $\mathcal{J}^* \subseteq \mathcal{J}$ denote the buckets j for which $\sum_{i \in S_j} \sigma_i \geqslant \varepsilon$, and let $S_{\mathsf{light}} \subseteq [d]$ denote all $i \in [d]$ for which $j(i) \in \mathcal{J}^*$. Let σ' denote the unnormalized density matrix given by zeroing out the i-th entry of σ for every $i \in S_{\mathsf{light}}$, and let σ^* denote the density matrix $\sigma' / \operatorname{tr}(\sigma')$.

Fact 5.5.2. *$|\mathcal{J}^*| \leqslant O(\log(d/\varepsilon))$. In particular, σ' is given by removing $O(\varepsilon \log(d/\varepsilon))$ mass from σ.*

Proof. Note that for any $j \in \mathcal{J}^*$, $2^{-j} > \sigma_i \geqslant \varepsilon/d$ for all $i \in S_j$, so $j < \log_2(\varepsilon/d)$. $\square$

5.5.2 Helper Lemmas

Here we collect some elementary observations that will be useful in our proof of the weaker version of Theorem 5.5.1. We begin by noting an alternative way of representing fidelity with respect to the maximally mixed state.

Fact 5.5.3. *Given psd matrix $\sigma \in \mathbb{C}^{d\times d}$, let $\widehat{\sigma} \triangleq \sigma/\operatorname{tr}(\sigma)$. Then $F(\widehat{\sigma}, \frac{1}{d}I_d) = \frac{1}{d}\|\sigma\|_{1/2} \cdot \operatorname{tr}(\sigma)^{-2}$.*

Fact 5.5.4. *Let S be any set of distinct positive integers. Given a collection of numbers $\{d_j\}_{j\in S}$ satisfying $\sum_j d_j 2^{-j} \leqslant 2$, let p be the vector with d_j entries equal to 2^{-j} for every j. Then $\|p\|_{1/2} \leqslant |S|^2 \cdot \max_j d_j^2 2^{-j}$.*

Proof. Let $j^* \triangleq \arg\max d_j^2 2^{-j}$. Then $\|p\|_{1/2}^{1/2} = \sum_j d_j 2^{-j/2} \leqslant |S| \cdot d_{j^*} 2^{-j^*/2}$. □

Our lower bound instances in the proof of the weaker version of Theorem 5.5.1 will be based on perturbing certain submatrices of σ. We will use the following basic fact to analyze these instances.

Lemma 5.5.5. *Consider the task of distinguishing between the following alternatives:*

$$H_0 : \rho = \begin{pmatrix} S & 0 \\ 0 & P \end{pmatrix} \qquad \text{and} \qquad H_1 : \rho = \begin{pmatrix} \widetilde{S} & 0 \\ 0 & P \end{pmatrix} \tag{5.41}$$

where $S \in \mathbb{R}^{d'\times d'}$ and $P \in \mathbb{R}^{(d-d')\times(d-d')}$ are deterministic psd matrices for which $\operatorname{tr}(S) + \operatorname{tr}(P) = 1$, and $\widetilde{S} \in \mathbb{R}^{d'\times d'}$ is drawn from some distribution over psd matrices with trace $\operatorname{tr}(S)$.

Then the copy complexity of this task using incoherent measurements is $\Omega(n/\operatorname{tr}(S)) \geqslant \Omega(n)$, where n is the copy complexity of the following distinguishing task:

$$H_0 : \rho = S/\operatorname{tr}(S) \qquad \text{and} \qquad H_1 : \rho = \widetilde{S}/\operatorname{tr}(S) \tag{5.42}$$

using incoherent measurements.

Theorem 5.3.3 gives a $\Omega(d^{3/2}/\varepsilon^2)$ lower bound for mixedness testing for d larger than some absolute constant. In the following lemma, we complement this with a weaker lower bound that holds for all d, based on the classical lower bound for uniformity testing. For this, consider the task of distinguishing between the two alternatives:

$$H_0 : \rho = \frac{1}{d}A \qquad \text{and} \qquad H_1 : \rho = \frac{1}{d}\big(A + \varepsilon P Z P^\top\big), \tag{5.43}$$

where $A \in \mathbb{R}^{d\times d}$ is a diagonal matrix with diagonal entries $a_1 \geqslant \cdots \geqslant a_d > 0$ and $\operatorname{tr}(A) = d$, where $Z = \operatorname{diag}(1, \cdots, -1, \cdots)$ if d is even and $Z = \operatorname{diag}(1, \cdots, -1, \cdots, -1, 0)$ otherwise, and where $P \in \mathbb{R}^{d\times d}$ is a random permutation matrix on the first $2\lfloor d/2 \rfloor$ coordinates.

Lemma 5.5.6. *For all $d > 1$ and $\varepsilon < 1$, the copy of complexity of distinguishing between H_0 and H_1 with incoherent measurements is $\Omega(\sqrt{d}/\varepsilon^2)$.*

Proof. By Lemma 5.5.5, it suffices to prove the lemma when d is even. As the states under H_0 and H_1 are both diagonal, we can assume without loss of generality that the measurements are all in the standard basis. Let p_0 denote the uniform distribution over $[d]$. Given $S \subset [d]$ of size $d/2$, let p_S denote the discrete distribution over $[d]$ which places mass $\frac{1+\varepsilon}{d}$ on elements in S and mass $\frac{1-\varepsilon}{d}$ on elements in $[d]\backslash S$. Under H_0, if one measures n copies of ρ, the n measurement outcomes are a sample from $p_0^{\otimes n}$. Under H_1, if one measures n copies of ρ, the measurement outcomes are a sample from $\mathbb{E}[S]p_S^{\otimes n}$ where S is a random subset of $[d]$ of size $d/2$. It is a standard result in distribution testing that distinguishing between $d_{\mathrm{TV}}(p_0^{\otimes n}, \mathbb{E}[S]p_S^{\otimes n}) = o(1)$ if $n = o(\sqrt{d}/\varepsilon^2)$ (see e.g. the proof of [216, Theorem 24.1] which is based on [174]). □

Finally, we will use the following lower bound to handle a corner case where σ has one especially large eigenvalue.

Lemma 5.5.7 (Lemma 5.24 from [65])**.** *Let $\varepsilon \leqslant 1/2$. If $\sigma_1 \geqslant 3/4$, then state certification to error ε with respect to σ using incoherent measurements is $\Omega(1/\varepsilon^2)$.*

5.5.3 Proof of Weaker Variant of Theorem 5.5.1

We now give a simple proof of a slight weakening of Theorem 5.5.1 where one removes $O(\varepsilon \log(d/\varepsilon))$ mass from σ, instead of $O(\varepsilon)$. We strengthen this analysis in Appendix C.2.

Proof. Note that $\mathrm{tr}(\sigma') \geqslant 1 - O(\varepsilon \log(d/\varepsilon)) \geqslant \Omega(1)$, so by Fact 5.5.3 it suffices to lower bound the copy complexity by

$$\Omega\left(d_{\mathsf{eff}} \|\sigma'\|_{1/2} / (\varepsilon^2 \log(d/\varepsilon))\right). \tag{5.44}$$

We proceed by casework depending on whether or not $d_j = 1$ for all $j \in \mathcal{J}^*$.

Case 1. $d_j = 1$ for all $j \in \mathcal{J}^*$. Note that in this case,

$$\|\sigma'\|_{1/2}^{1/2} = \sum_{j \in \mathcal{J}^*} 2^{-j/2} = O(1). \tag{5.45}$$

and $\|\sigma^*\|_{1/2} = \Theta(\|\sigma'\|_{1/2})$. As $d_{\mathsf{eff}} = 1$, it thus suffices to show a copy complexity lower bound of $\Omega(1/\varepsilon^2)$ in this case.

If additionally we have $|\mathcal{J}^*| = 1$, then for $\varepsilon \leqslant \widetilde{O}(1/\log d)$ sufficiently small, the maximum entry of σ is at least $3/4$, so we can apply Lemma 5.5.7 to obtain a lower bound of $\Omega(1/\varepsilon^2)$ as desired.

Otherwise, let j, j' be the two smallest bucket indices in $\mathcal{J}^*$, and let i, i' be the elements of the singleton sets $S_j, S_{j'}$. If $\varepsilon \leqslant c 2^{-j/2 - j'/2 - 1}$ for sufficiently small constant $c > 0$, we can invoke [65, Lemma A.4] to conclude a copy complexity lower bound of $\Omega(1/\varepsilon^2)$.*

Otherwise, suppose $\varepsilon > c 2^{-j/2 - j'/2 - 1}$. Because $2^{-j} > 2^{-j'}$, we know that $2^{-j'} \leqslant O(\varepsilon)$. In particular, consider the state σ^{**} given by zeroing out $\sigma_{i'}$ from σ' and normalizing.

*Note that [65, Lemma A.4] gives a (suboptimal) lower bound for the distinguishing task in Section 5.4. The reason we invoke it instead of Theorem 5.4.3 is that unlike the latter, it holds for the setting $d_j = d_{j'} = 1$ that we consider here.

For this matrix, $d_{\mathsf{eff}} = 1$ and $\|\sigma^{**}\|_{1/2} = O(1)$. Furthermore, because $\varepsilon \leqslant \widetilde{O}(1/\log(d))$, we have $\sigma_i \geqslant 3/4$, so we can apply Lemma 5.5.7 to conclude a lower bound of $\Omega(1/\varepsilon^2)$.

Case 2. $d_j > 1$ for some $j \in \mathcal{J}^*$. In this case, let $j_1 \triangleq \arg\max_{j\in\mathcal{J}^*} d_j$ and $j_2 \triangleq \arg\max_{j\in\mathcal{J}^*} d_j^2 2^{-j}$.

If $\varepsilon \leqslant c d_{j_2} 2^{-j_1/2-j_2/2-1}/j_1$ for sufficiently small constant $c > 0$, we can apply the lower bound instance in Section 5.4 to these two buckets. Let P denote the submatrix of σ containing the diagonal entries of σ outside of $S_{j_1} \cup S_{j_2}$. Let Σ_{j_1} and Σ_{j_2} denote the submatrices of σ containing the diagonal entries indexed by S_{j_1} and S_{j_2}.

If $j_1 \neq j_2$, then consider the distinguishing task

$$H_0 : \rho = \begin{pmatrix} \Sigma_{j_1} & 0 & 0 \\ 0 & \Sigma_{j_2} & 0 \\ 0 & 0 & P \end{pmatrix} (= \sigma) \qquad \text{and} \qquad H_1 : \rho = \begin{pmatrix} \Sigma_{j_1} & \frac{\varepsilon}{d_{j_1}^{1/2} d_{j_2}} \overline{G} & 0 \\ \frac{\varepsilon}{d_{j_1}^{1/2} d_{j_2}} \overline{G}^\top & \Sigma_{j_2} & 0 \\ 0 & 0 & P \end{pmatrix}, \tag{5.46}$$

where $\overline{G}$ is a C-truncated $d_{j_1} \times d_{j_2}$ Ginibre matrix. If d_{j_1} is sufficiently large that Theorem 5.4.3 applies, then by Lemma 5.5.5 and Theorem 5.4.3, this has copy complexity at least

$$\Omega\left(\frac{\sqrt{d_{j_1}} \cdot d_{j_2}}{(\varepsilon/(d_{j_1}2^{-j_1} + d_{j_2}2^{-j_2}))^2}\right) \geqslant \Omega(\sqrt{d_{j_1}} \cdot d_{j_2}^2 2^{-j_2}/\varepsilon^2) \geqslant \Omega(d_{\mathsf{eff}}\|\sigma'\|_{1/2}/(\varepsilon^2 \operatorname{polylog}(d/\varepsilon))), \tag{5.47}$$

where in the first step we used that $d_{j_2} 2^{-j_2} \leqslant 2$, and in the last step we used Fact 5.5.2 and Fact 5.5.4. Note that $\|\sigma^*\|_{1/2} = \Theta(\|\sigma'\|_{1/2})$. Otherwise, if $d_{j_1} = O(1)$ and Theorem 5.4.3 doesn't apply, we can still apply [65, Lemma A.6] which only differs in its suboptimal dependence of $d_{j_1}^{1/3}$ on the parameter d_{j_1}, which does not affect our overall bound as $d_{j_1} = O(1)$ in this case.

If $j_1 = j_2$, then because we are in Case 2 we know $d_{j_1} > 1$, so let $\Sigma_{j_1}^{(1)}$ and $\Sigma_{j_1}^{(2)}$ denote an arbitrary partition of Σ_{j_1} into two $\frac{d_1}{2} \times \frac{d_1}{2}$ diagonal submatrices. Consider the

distinguishing task

$$H_0 : \rho = \begin{pmatrix} \Sigma_{j_1}^{(1)} & 0 & 0 \\ 0 & \Sigma_{j_1}^{(2)} & 0 \\ 0 & 0 & P \end{pmatrix} (= \sigma) \qquad \text{and} \qquad H_1 : \rho = \begin{pmatrix} \Sigma_{j_1}^{(1)} & \frac{\varepsilon}{d_{j_1}^{1/2} d_{j_2}} \overline{G} & 0 \\ \frac{\varepsilon}{d_{j_1}^{1/2} d_{j_2}} \overline{G}^\top & \Sigma_{j_1}^{(2)} & 0 \\ 0 & 0 & P \end{pmatrix}, \tag{5.48}$$

where $\overline{G}$ is a C-truncated $\frac{d_{j_1}}{2} \times \frac{d_{j_1}}{2}$ Ginibre matrix. If d_{j_1} is sufficiently large that Theorem 5.4.3 applies, then by Lemma 5.5.5 and Theorem 5.4.3, this has copy complexity at least

$$\Omega\left(\frac{\sqrt{d_{j_1}} \cdot d_{j_1}}{(\varepsilon/d_{j_1} 2^{-j_1} + d_{j_2} 2^{-j_2})^2}\right) \geqslant \Omega(\sqrt{d_{j_1}} \cdot d_{j_1}^2 2^{-j_1}/\varepsilon^2) \geqslant \Omega(d_{\mathsf{eff}} \|\sigma'\|_{1/2}/(\varepsilon^2 \operatorname{polylog}(d/\varepsilon))), \tag{5.49}$$

where in the first step we used that $d_{j_1} 2^{-j_1} \leqslant 2$, and in the last step we used Fact 5.5.2 and Fact 5.5.4. Otherwise, if $d_{j_1} = O(1)$, we can apply [65, Lemma A.4] as above.

It remains to consider the case where $\varepsilon > c 2^{-j_1/2 - j_2/2 - 1}/j_1$. Let $j^* \triangleq \arg\max_{j \in \mathcal{J}^*} d_j 2^{-5j/2}$. We can apply the lower bound instance in Section 5.3 to bucket j^*. Letting Σ_{j^*} denote the submatrix of σ containing the diagonal entries of S_{j^*} and P denote the submatrix containing the remaining diagonal entries, we consider the distinguishing task

$$H_0 : \rho = \begin{pmatrix} \Sigma_{j^*} & 0 \\ 0 & P \end{pmatrix} (= \sigma) \qquad \text{and} \qquad \rho = \begin{pmatrix} \Sigma_{j^*} + \frac{\varepsilon}{d_{j^*}} \cdot \overline{M} & 0 \\ 0 & P \end{pmatrix}, \tag{5.50}$$

where $\overline{M}$ is a C-truncated trace-centered GOE matrix. By Lemma 5.5.5 together with either Theorem 5.3.3 if $d \gg 1$ or Lemma 5.5.6 if $d = O(1)$, this has copy complexity at least

$$\Omega\left(\frac{d_{j^*}^{3/2}}{(\varepsilon/(d_{j^*} 2^{-j^*}))^2} \cdot \frac{1}{\operatorname{tr}(\Sigma_{j^*})}\right) = \Omega(d_{j^*}^{5/2} 2^{-j^*}/\varepsilon^2) \geqslant \Omega(d_{j_1}^{5/2} 2^{-j_1}/\varepsilon^2). \tag{5.51}$$

To complete the proof of the theorem, it suffices to show that

$$d_{j_1}^{5/2} 2^{-j_1} \operatorname{polylog}(d/\varepsilon) \geqslant \Omega\left(\sqrt{d_{j_1}} d_{j_2}^2 2^{-j_2}\right) \tag{5.52}$$

Suppose to the contrary. Then we would get

$$d_{j_1}^2 2^{-j_1} \operatorname{polylog}(d/\varepsilon) = o\left(d_{j_2}^2 2^{-j_2}\right). \tag{5.53}$$

But by assumption on ε,

$$c d_{j_2} 2^{-j_1/2 - j_2/2 - 1} / j_1 \leqslant \varepsilon \leqslant d_{j_1} 2^{-j_1}, \tag{5.54}$$

where in the last step we used the fact that $d_j 2^{-j} \geqslant \Omega(\varepsilon)$ for any $j \in \mathcal{J}^*$. Squaring both sides and rearranging, we find that

$$d_{j_2}^2 2^{-j_2} \leqslant O(d_{j_1}^2 2^{-j_1} j_1^2) \leqslant O(d_{j_1}^2 2^{-j_1} \log^2(d/\varepsilon)), \tag{5.55}$$

where the last step follows by the fact that $j_1 \leqslant \log(d/\varepsilon)$ because $2^{-j_1} d \geqslant 2^{-j_1} d_{j_1} \geqslant \varepsilon$, contradicting (5.53). □

Chapter 6

Tradeoffs for Few-Copy Measurements

In Chapter 4 we studied the basic task of quantum state tomography, and in this chapter we return to this question. Recall that in quantum state tomography, one is given n copies of a d-dimensional mixed state ρ, and the goal is to estimate its density matrix to error ε, e.g. in trace distance, by measuring copies of ρ.

As discussed previously, [102] and [170] settled the optimal copy complexity of this problem, showing that $n = \Theta(d^2/\varepsilon^2)$ copies are necessary and sufficient but these protocols require fully coherent measurements across the joint state $\rho^{\otimes n}$, and thus are impractical on near-term setups, likely requiring fault-tolerant quantum computers to be tractable.

Motivated by these constraints, in Chapter 4, we studied what one can do with *incoherent* i.e. single-copy measurements. These algorithms are much more practical to run and in Chapter 4 we showed that the optimal copy complexity in this setting is $n = \Theta(d^3/\varepsilon^2)$ — and thus incoherent measurements are provably weaker than coherent measurements.

In this chapter, we study what happens between the incoherent and fully coherent regimes. This is a particularly timely consideration given recent advancements in quantum devices and the fact that, while still far from the unconstrained fully co-

herent regime, they may soon have capabilities that go beyond purely incoherent measurements. Specifically, we study what happens if the learner can make entangled measurements over a limited number, say t copies of ρ at a time, which we call t-entangled measurements. This naturally interpolates between the incoherent regime ($t = 1$) and the fully coherent regime ($t = n$). We will show a tight characterization of the copy complexity of state tomography with t-entangled measurements, for a large range of t.

6.1 Results

Our main results in this chapter are the following pair of theorems.

Theorem 6.1.1 (informal upper bound, see Theorem 6.4.8)**.** *Let $t \leqslant \min(d^2, (\sqrt{d}/\varepsilon)^c)$, for some (sufficiently small) universal constant c. There is an algorithm that uses*

$$n = \widetilde{O}\left(\frac{d^3}{\sqrt{t}\varepsilon^2}\right) .$$

total copies of ρ by taking n/t separate t-entangled measurements, and outputs $\hat{\rho}$ so that $\|\rho - \hat{\rho}\|_1 \leqslant \varepsilon$ with high probability.

Theorem 6.1.2 (informal lower bound, see Theorem 6.6.13)**.** *Let $t \leqslant 1/\varepsilon^c$ for some (sufficiently small) universal constant c. Any algorithm that uses n total copies of ρ by taking taking n/t separate (but possible adaptively chosen) t-entangled measurements must require*

$$n = \Omega\left(\frac{d^3}{\sqrt{t}\varepsilon^2}\right)$$

to succeed at state tomography, with probability $\geqslant 0.01$.

Together, these theorems imply that the copy complexity of tomography with t-entangled measurements is, up to logarithmic factors, exactly $\widetilde{\Theta}\left(\frac{d^3}{\sqrt{t}\varepsilon^2}\right)$, for all $t \leqslant 1/\varepsilon^c$. In other words, for this regime of t, the sample complexity smoothly improves with t. To our knowledge, this is the first nontrivial setting in which the copy complexity has

been sharply characterized to smoothly depend on the amount of entanglement. Prior works on the power of t-entangled measurements either did not obtain sharp tradeoffs, or studied settings where simply taking t to be a large enough constant sufficed to obtain optimal copy complexity (e.g. Bell sampling [66, 164]).

We pause here to make a few remarks about our result. First, as mentioned above, the most practically pertinent regime of t is the regime where t is small but > 1. Our result fully characterizes this setting, so long as $\varepsilon = o(1)$, and obtains a nontrivial partial tradeoff even for constant ε. We believe that the restriction that $t \leqslant 1/\varepsilon^c$ is ultimately an artifact of our techniques, and we conjecture that $\widetilde{\Theta}\left(\frac{d^3}{\sqrt{t}\varepsilon^2}\right)$ is the right rate for all $t \leqslant O(d^2)$. We leave resolving this question as an interesting future direction.

Second, we note that our result implies that when ε is sufficiently small, entanglement only helps up until a certain point. When $\varepsilon \ll 1/d$, we can take $t = \Theta(d^2)$, and our algorithm achieves a rate of $n = \widetilde{O}(\frac{d^2}{\varepsilon^2})$, which matches (up to logarithmic factors) the lower bound against general quantum algorithms with any degree of entanglement. In other words, we can match the rate of the fully coherent algorithm, with asymptotically less coherence. To our knowledge, this is the first instance for a natural learning problem where a super-constant, but still partial, amount of entanglement is the minimal amount of entanglement required to obtain the statistically optimal rate.

Our techniques. As previously mentioned, we need a number of novel ideas to overcome the difficulties with prior techniques for the t-entangled setting. In this section, we describe some of these conceptual contributions at a high level.

A key idea, which is crucial to both our upper and lower bounds, and which to our knowledge is novel in the literature, is something we call *linearization*. For any state ρ, we can write it as $\rho = \frac{I_d}{d} + E$, where E is some traceless Hermitian matrix. Then,

for any integer t, we observe that $\rho^{\otimes t}$ can be approximated as follows

$$\rho^{\otimes t} \approx \left(\frac{1}{d} I_d\right)^{\otimes t} + \sum_{\text{sym}} E \otimes \left(\frac{1}{d} I_d\right)^{\otimes t-1} , \tag{6.1}$$

where here and throughout the paper, we let $\sum_{\text{sym}}$ denote the symmetric sum, i.e.

$$\sum_{\text{sym}} A_1 \otimes \ldots \otimes A_t = \sum_{\pi \in S_t} A_{\pi(1)} \otimes \ldots \otimes A_{\pi(t)} .$$

We call the expression on the RHS of (6.1) the *t-linearization* of ρ, as it is the linear term (in E) of the expansion of $\rho^{\otimes t}$. Note that while the linearization of the state is not necessarily itself a mixed state (as it need not be PSD), we can still perform formal calculations with it. For our purposes, the linearization of ρ has a number of crucial properties. While estimators based on weak Schur sampling do not yield an unbiased estimate of ρ when applied to $\rho^{\otimes t}$, they do yield an unbiased estimate of E (up to a known correction term) when "applied" to the t-linearization of ρ. Therefore, as long as the linearization of ρ is a good approximation for $\rho^{\otimes t}$, we may average the result of many independent trials of such estimators to obtain a better estimate for E, and consequently, ρ. For the lower bound, it turns out that the amount of information gained by the algorithm from some measurement outcome (more formally, the likelihood ratio of the posterior distribution for the algorithm starting from two different states, see Lemma 6.5.2) is controlled by the Frobenius norm of some linear transformation of the corresponding POVM element.

Another important and novel algorithmic idea is that of *quantum splitting*. It turns out that our techniques work best for states which are relatively spread-out, i.e. states with spectral norm at most $O(1/d)$. To extend our techniques to general mixed states, we give a novel procedure which reversibly transforms any state into another state of marginally larger dimension that is spectrally bounded. One can view this as a quantum generalization of the splitting procedure introduced in [75] for classical distribution learning and testing. There, as well, the goal was to take an arbitrary distribution over $[d]$ elements and produce a "well-spread" distribution over a slightly

larger domain. However, more care must be taken in the quantum setting, as our splitting procedure must be done obliviously of the (unknown) eigenvectors of the true state ρ, and so any such procedure will non-trivially alter the spectral properties of ρ.

6.1.1 Additional Related work

The results in this chapter follow a flurry of recent works characterizing the effect that near-term constraints like noise and limited quantum memory have on statistical rates for quantum learning tasks. Many of these have focused on proving lower bounds for protocols that can only make incoherent measurements, e.g. [54, 65, 67, 64, 66, 5, 79, 62]. Among these works, the two most relevant ones are [62] and [64]. The work in [62] has been covered in Chapter 4.

In [64], the authors are motivated by the same general question as the present work: are there tasks for which being able to perform t-entangled measurements for large t is strictly more powerful than for small t? Instead of tomography, they studied a certain hypothesis testing task based on distinguishing approximate state k-designs on n qubits from maximally mixed. They show that when $t < k$, any protocol must use exponentially many copies, but with fully-entangled measurements, $\mathrm{poly}(n, k, 1/\varepsilon)$ copies suffice because one can simply run quantum hypothesis selection [26]. Unlike here, they do not manage to give a full tradeoff. Additionally, they prove some partial results for using t-entangled measurements for mixedness testing [169, 54, 67], where the goal is to distinguish the maximally mixed state from states that are ε-far in trace distance from maximally mixed. Specifically, when $t = o(\log(d))/\varepsilon^2$, the polynomial dependence on d in their lower bound matches that of the lower bound of [54] for $t = 1$. Although that lower bound is suboptimal even for $t = 1$, and although [64] did not achieve a full tradeoff for mixedness testing, we remark that it hints at an interesting phenomenon. Unlike in tomography, where copy complexity smoothly decreases as t increases, for mixedness testing, the number of copies t must exceed some *ε-dependent threshold* before one can improve upon the single-copy rate.

Lastly, we remark that our analysis of the upper bound makes crucial use of certain

representation-theoretic estimates proved by [170]. While these estimates were originally used to analyze Keyl's estimator in the fully-entangled setting, in our analysis we leverage them in a novel way to prove guarantees in the partially-entangled setting
.

6.2 Preliminaries

Throughout, let $\rho \in \mathbb{C}^{d \times d}$ be the unknown state that we are trying to learn.

t-Entangled Measurements An algorithm that uses t-entangled measurements and n total copies of ρ operates as follows: it is given $m = n/t$ copies of $\rho^{\otimes t}$. It then iteratively measures the i-th copy of $\rho^{\otimes t}$ (for $i = 1, 2, \ldots, m$) using a POVM in $\mathbb{C}^{d^t \times d^t}$ which could depend on the results of previous measurements, records the outcome, and then repeats this process on the $(i+1)$-th copy of $\rho^{\otimes t}$. After performing all m measurements, it must output an estimate of ρ based on the (classical) sequence of outcomes it has received.

Remark 6.2.1. Note that limiting the batch size to exactly t is not actually a limitation because we can simulate an algorithm that adaptively chooses batch sizes of at most t with an algorithm that uses batch sizes of exactly t up to at most a factor of 2 in the total copy complexity. This is because we can simply combine batches whenever their total size is at most t.

6.2.1 Representation Theory Basics

We will introduce some notation and facts from representation theory for analyzing entangled quantum measurements. Our exposition closely follows [170, 171]. For a more detailed explanation of the elementary representation theory results, see e.g. [215]. We use GL_d to denote the general linear group in $\mathbb{C}^{d \times d}$ and U_d to denote the unitary group in $\mathbb{C}^{d \times d}$. Next, we introduce some notation for partitions.

Definition 6.2.2. Given a positive integer n, a partition of n into d parts is a list

of positive integers $\lambda = (\lambda_1, \ldots, \lambda_d)$ with $\lambda_1 \geqslant \ldots \geqslant \lambda_d$ such that $\lambda_1 + \cdots + \lambda_d = n$. We write $\lambda \vdash n$ to denote that λ is a partition of n. We use $|\lambda|$ to denote the total number of elements in λ and $\ell(\lambda)$ to denote the number of nonzero parts in λ, e.g. in the previous example $|\lambda| = n$ and $\ell(\lambda) = d$.

Definition 6.2.3. Given two partitions $\lambda, \lambda' \vdash n$ where $\lambda = (\lambda_1, \ldots, \lambda_d)$ and $\lambda' = (\lambda'_1, \ldots, \lambda'_{d'})$, we say λ majorizes λ' if

$$\lambda_1 + \cdots + \lambda_i \geqslant \lambda'_1 + \cdots + \lambda'_i$$

for all i where λ_j, λ'_j are defined to be 0 whenever j exceeds the number of parts in the partition. If λ majorizes λ', we write $\lambda \succeq \lambda'$.

Fact 6.2.4 ([184])**.** *For any n, the number of distinct partitions $\lambda \vdash n$ is at most $2^{3\sqrt{n}}$.*

Definition 6.2.5. Given a partition $\lambda \vdash n$, let $f_i(\lambda)$ be the number of parts of λ that are equal to i for each integer i. We define $\mathsf{count}(\lambda) = f_1(\lambda)! \cdots f_n(\lambda)!$.

Definition 6.2.6. Given a partition $\lambda \vdash n$, we let $\lambda^T \vdash n$ denote its transpose i.e. λ^T_1 the number of parts of λ that are at least 1, λ^T_2 is the number of parts of λ that are at least 2 and so on.

Definition 6.2.7. [Young Tableaux] We have the following standard definitions:

- Given a partition $\lambda \vdash n$, a Young diagram of shape λ is a left-justified set of boxes arranged in rows, with λ_i boxes in the ith row from the top.

- A standard Young tableaux (SYT) T of shape λ is a Young diagram of shape λ where each box is filled with some integer in $[n]$ such that the rows are strictly increasing from left to right and the columns are strictly increasing from top to bottom.

- A semistandard Young tableaux (SSYT) T of shape λ is a Young diagram of shape λ where each box is filled with some integer in $[d]$ for some d and the rows

are weakly increasing from left to right and the columns are strictly increasing from top to bottom.

Now we review the correspondence between Young tableaux and representations of the symmetric and general linear groups.

Definition 6.2.8. We say a representation μ of GL_d over a complex vector space $\mathbb{C}^m$ is a polynomial representation if for any $U \in \mathbb{C}^{d\times d}$, $\mu(U) \in \mathbb{C}^{m\times m}$ is a polynomial in the entries of U.

Fact 6.2.9 ([187])**.** *The irreducible representations of the symmetric group S_n are exactly indexed by the partitions $\lambda \vdash n$ and have dimensions* $\dim(\lambda)$ *equal to the number of standard Young tableaux of shape λ. We denote the corresponding vector space* Sp_λ.

Fact 6.2.10 ([93])**.** *For each $\lambda \vdash n$, there is a (unique) irreducible polynomial representation of GL_d corresponding to λ. We denote the corresponding map and vector space $(\pi_\lambda, V_\lambda^d)$. The dimension* $\dim(V_\lambda^d)$ *is equal to the number of semistandard Young tableaux of shape λ with entries in $[d]$. This representation, restricted to U_d is also an irreducible representation.*

Theorem 6.2.11 (Schur-Weyl Duality [93])**.** *Consider the representation of $S_n \times GL_d$ on $(\mathbb{C}^d)^{\otimes n}$ where the action of the permutation $\pi \in S_n$ permutes the different copies of $\mathbb{C}^d$ and the action of $U \in GL_d$ is applied independently to each copy. This representation can be decomposed as a direct sum*

$$(\mathbb{C}^d)^{\otimes n} = \bigoplus_{\substack{\lambda\vdash n \\ \ell(\lambda)\leqslant d}} \mathsf{Sp}_\lambda \otimes V_\lambda^d \,.$$

Definition 6.2.12 (Schur Subspace)**.** We call $\mathsf{Sp}_\lambda \otimes V_\lambda^d$ the λ-Schur subspace. Given integers n, d and $\lambda \vdash n$, we define $\Pi_\lambda^d : (\mathbb{C}^d)^{\otimes n} \to \mathsf{Sp}_\lambda \otimes V_\lambda^d$ to project onto the λ-Schur subspace.

Theorem 6.2.13 (Gelfand-Tsetlin Basis [93])**.** *Let n, d be positive integers. For each partition $\lambda \vdash n$ where λ has at most d parts, there is a basis $v_1, \ldots, v_m$ of V_λ^d with $m = \dim(V_\lambda^d)$ such that for any matrix $D_\alpha = \mathrm{diag}(\alpha_1, \ldots, \alpha_d)$, we have $v_i^\dagger \pi_\lambda(D_\alpha) v_i = \alpha^{f^{(i)}}$ for all i where $f^{(i)}$ are each d-tuples that give the frequencies of $1, 2, \ldots, d$ in each of the different semi-standard tableaux of shape λ.*

We now present a few consequences of Theorem 6.2.11 and Theorem 6.2.13 that will be used in our learning primitives. Roughly, we give a more explicit representation of the Gelfand-Tsetlin basis when embedded in the representation of $(\mathbb{C}^d)^{\otimes n}$.

Definition 6.2.14. For integers n, d and a d-tuple $f_1, \ldots, f_d$ such that $f_1 + \cdots + f_d = n$, we define $G_{f_1,\ldots,f_d}[n \to d]$ to be the set of all functions $g : [n] \to [d]$ where the multiset $\{g(1), \ldots, g(n)\}$ has 1 with frequency f_1, 2 with frequency f_2 and so on.

Lemma 6.2.15. *Let $f = (f_1, \ldots, f_d)$ where $f_1, \ldots, f_d \geqslant 0$ are integers and $f_1 + \cdots + f_d = n$. Let $v_1, \ldots, v_d$ be orthogonal vectors in $\mathbb{C}^d$. If λ is a partition that majorizes $(f_1, \ldots, f_d)$ (after sorting in decreasing order) then there are weights $\{w_g\}_{g \in G_{f_1,\ldots,f_d}[n \to d]}$ such that the vector*

$$v_f = \sum_{g \in G_{f_1,\ldots,f_d}[n \to d]} w_g (v_{g(1)} \otimes \cdots \otimes v_{g(n)})$$

is in the Schur subspace $\mathsf{Sp}_\lambda \otimes V_\lambda^d$. Furthermore, we can choose $\dim(\lambda)$ orthogonal unit vectors $v_{f,1}, \ldots, v_{f,\dim(\lambda)}$ of the above form the weights so that

$$\mathop{\mathbb{E}}_U \left[U^{\otimes n} \left(\sum_{j=1}^{\dim(\lambda)} v_{f,j} v_{f,j}^\dagger \right) (U^\dagger)^{\otimes n} \right] = \frac{1}{\dim(V_\lambda^d)} \Pi_\lambda^d$$

where the expectation is over Haar random unitaries U.

Proof. First, WLOG $v_1, \ldots, v_d$ are the standard basis $e_1, \ldots, e_d$. This is because by Theorem 6.2.11, we can always rotate all of the n copies simultaneously by the same unitary while staying within the same Schur subspace. Now imagine decomposing $\mathbb{C}^{d^n}$

into the subspaces given by Theorem 6.2.11. Take one of the copies of V_λ^d. Note that since λ majorizes $(f_1, \ldots, f_d)$, there is a semi-standard tableaux of shape λ where the numbers $1, 2, \ldots, d$ occur exactly $f_1, \ldots, f_d$ times respectively. Thus, we can apply Theorem 6.2.13 to find the vector v_f in it such that

$$v_f^\dagger D_\alpha^{\otimes n} v_f = \alpha^f$$

where $D_\alpha = \mathrm{diag}(\alpha_1, \ldots, \alpha_d)$. Note that both sides of the above are polynomials in the $\alpha_1, \ldots, \alpha_d$ and it holds for all values, so it must actually hold as a polynomial identity. Thus, v_f must be contained in the eigenspace of $D_\alpha^{\otimes n}$ with eigenvalue α^f which is exactly the subsapce spanned by $e_{g(1)} \otimes \cdots \otimes e_{g(n)}$ as g ranges over all of the functions in $G_{f_1,\ldots,f_d}[n \to d]$. This immediately implies the first statement.

For the second statement, the fact that V_λ^d is an irreducible representation of U_d implies that

$$\mathop{\mathbb{E}}_U \left[U^{\otimes n} v_f v_f^\dagger (U^\dagger)^{\otimes n} \right]$$

is equal to $\frac{1}{\dim(V_\lambda^d)}$ times a projection onto the copy of V_λ^d that v_f is in. Thus, we can simply use the same construction and pick $v_{f,1}, \ldots, v_{f,\dim(\lambda)}$, one from each of the copies of V_λ^d in $\mathbb{C}^{d^n}$ and then we get the desired statement. □

In light of Lemma 6.2.15, we make the following definition.

Definition 6.2.16. Given integers n, d and a partition $\lambda \vdash n$, we define the vectors $v_{\lambda,1}^d, \ldots, v_{\lambda,\dim(\lambda)}^d \in \mathbb{C}^{d^n}$ to be the vectors constructed in Lemma 6.2.15 where we choose $v_1, \ldots, v_d$ to be the standard basis and $(f_1, \ldots, f_d) = (\lambda_1, \ldots, \lambda_d)$ (when the vectors are not unique, we pick arbitrary ones, the choice will not matter when we use this later on). We define

$$M_\lambda^d = \sum_{j=1}^{\dim(\lambda)} v_{\lambda,j}^d (v_{\lambda,j}^d)^\dagger .$$

We may drop the d in the superscript and simply write M_λ when d is clear from context.

We also have the following lemma that provides a type of converse to Lemma 6.2.15.

Lemma 6.2.17. *Let $g : [n] \to [d]$ be a function in $G_{f_1,\ldots,f_d}[n \to d]$. Then for any partition $\lambda \vdash n$ that does not majorize $(f_1, \ldots, f_d)$ (after sorting in decreasing order), we have*

$$\Pi^d_\lambda(v_{g(1)} \otimes \cdots \otimes v_{g(n)}) = 0\,.$$

Proof. As before, due to Theorem 6.2.11, WLOG $v_1, \ldots, v_d$ are the standard basis $e_1, \ldots, e_d$. Take any of the copies of V^d_λ. Note that by assumption, there does not exist any semi-standard tableaux of shape λ where $1, 2, \ldots, d$ occur with frequencies $f_1, \ldots, f_d$ respectively. Now consider a basis $v_{f^{(1)}}, \ldots, v_{f^{(m)}}$ for V^d_λ as given by Theorem 6.2.13. As in Lemma 6.2.15, we have that each vector $v_{f^{(i)}}$ must be contained in the span of vectors $v_{g'(1)} \otimes \cdots \otimes v_{g'(n)}$ as g' ranges over all functions in $G_{f^{(i)}_1,\ldots,f^{(i)}_d}[n \to d]$. However, as no semi-standard tableaux of shape λ has $1, 2, \ldots, d$ occur with frequencies $f_1, \ldots, f_d$, these subsapces are all orthogonal to $v_{g(1)} \otimes \cdots \otimes v_{g(n)}$ and thus $v_{f^{(i)}}$ is orthogonal as well. Since the $v_{f^{(i)}}$ form a basis for (one copy of) V^d_λ and we can repeat the same argument for all of the other copies, $v_{g(1)} \otimes \cdots \otimes v_{g(n)}$ is actually orthogonal to the entire λ-Schur subspace and we are done. □

Equipped with the above constructions, we can define the following POVMs.

Definition 6.2.18 (Weak Schur Sampling)**.** We use the term weak Schur sampling to refer to the POVM on $\mathbb{C}^{d^n \times d^n}$ with elements given by Π^d_λ for λ ranging over all partitions of n into at most d parts.

Our algorithm will make use of a POVM introduced by Keyl in [135]. However, the actual estimator we construct and its analysis will (necessarily) be very different from previous works.

Definition 6.2.19 (Keyl's POVM [135])**.** We define the following POVM on $\mathbb{C}^{d^n \times d^n}$: first perform weak Schur sampling to obtain $\lambda \vdash n$. Then within the subspace $\mathsf{Sp}_\lambda \otimes V^d_\lambda$,

measure according to

$$\dim(V_\lambda^d)\{U^{\otimes n} M_\lambda (U^\dagger)^{\otimes n})\}_U$$

where U ranges over Haar random unitaries. Note that the outcome of the measurement consists of a partition $\lambda \vdash n$ and a unitary $U \in \mathbb{C}^{d\times d}$.

Remark 6.2.20. The fact that this is a valid POVM follows from Lemma 6.2.15.

6.2.2 Schur Sampling and the Schur-Weyl Distribution

Next, we introduce a few standard tools for analyzing entangled measurements involving Schur sampling and related POVMs.

Definition 6.2.21. For a d-tuple of variables $x = (x_1, \ldots, x_d)$ and integers $\lambda = (\lambda_1, \ldots, \lambda_d)$, we use $\sum_{\text{sym}} x^\lambda$ to denote the symmetric polynomial that is a sum over all distinct monomials of the form $x_1^{\lambda_{\pi(1)}} \cdots x_d^{\lambda_{\pi(d)}}$ for permutations π on $[d]$.

Definition 6.2.22 (Schur Polynomial)**.** Given integers n, d and a partition $\lambda \vdash n$, we define the d-variate polynomial $s_\lambda(x_1, \ldots, x_d)$ as

$$s_\lambda(x_1, \ldots, x_d) = \sum_{\substack{\text{SSYT } T: \\ T \text{ has shape } \lambda}} x^T$$

where $x^T = x_1^{f_1} \cdots x_d^{f_d}$ where $f_1, \ldots, f_d$ are the frequencies of $1, 2, \ldots, d$ in the tableaux T.

Fact 6.2.23 ([93])**.** *We have the following properties:*

- *The Schur polynomial $s_\lambda(x_1, \ldots, x_d)$ is symmetric in the variables $x_1, \ldots, x_d$*
- *The Schur polynomial is equal to the character of the representation π_λ i.e. for a matrix $M \in \mathbb{C}^{d\times d}$ with eigenvalues $\alpha_1, \ldots, \alpha_d$, $\operatorname{tr}(\pi_\lambda(M)) = s_\lambda(\alpha_1, \ldots, \alpha_d)$.*

- *We have for any $\alpha_1, \ldots, \alpha_d$, the identity*

$$\sum_{\lambda \vdash n} \dim(\lambda) s_\lambda(\alpha_1, \ldots, \alpha_d) = (\alpha_1 + \cdots + \alpha_d)^n$$

Definition 6.2.24 (Schur Weyl Distribution)**.** Given integers n, d and a tuple $(\alpha_1, \ldots, \alpha_d)$ with $\alpha_i \geqslant 0$ and $\alpha_1 + \cdots + \alpha_d = 1$, the Schur-Weyl distribution $\mathrm{SW}^n(\alpha)$ is a distribution over partitions $\lambda \vdash n$ into at most d parts where the probability of sampling a given λ is $\dim(\lambda) \cdot s_\lambda(\alpha)$. When α is uniform, we may write SW^n_d instead.

Fact 6.2.25. *For a quantum state $\rho \in \mathbb{C}^{d \times d}$ with eigenvalues $\alpha_1, \ldots, \alpha_d$, if we perform weak Schur sampling on $\rho^{\otimes n}$, then the resulting distribution over λ is exactly* $\mathrm{SW}^n(\alpha)$.

One of the key tools for understanding the Schur-Weyl distribution is the following combinatorial characterization.

Fact 6.2.26 (The RSK-Correspondence, Greene's Theorem [184])**.** *Consider $\lambda \sim \mathrm{SW}^n(\alpha)$. Alternatively, sample a sequence of n tokens in $[d]$ say $x = (x_1, \ldots, x_n)$ where each token is drawn independently from the distribution $(\alpha_1, \ldots, \alpha_d)$. For each $k \leqslant n$, let l_k be the maximum length of the union of k disjoint weakly increasing subsequences of x and let l'_k be the maximum length of the union of k disjoint strictly decreasing subsequences of k. Then we have that:*

- *The distribution of λ is the same as the distribution of $(l_1, l_2 - l_1, l_3 - l_2, \ldots, l_n - l_{n-1})$*

- *The distribution of λ^T is the same as the distribution of $(l'_1, l'_2 - l'_1, \ldots, l'_n - l'_{n-1})$*

We now prove a few basic inequalities about the Schur-Weyl distribution. Compared to [170, 171] where inequalities of a similar flavor are used, the inequalities here are stronger in the regime $n \ll d^2$, which will be crucial later on.

Claim 6.2.27. Let $\alpha = (\alpha_1, \ldots, \alpha_d)$ be a vector of nonnegative weights summing to

1. Then for $\lambda \sim \mathrm{SW}^n(\alpha)$, with probability at least $1/2$,

$$\sum_{i=1}^{d} \lambda_i^2 \geqslant \frac{n^{1.5}}{4}.$$

Proof. Consider sampling a sequence $x = (x_1, \ldots, x_n)$ in $[d]^n$ where each token is drawn independently from the distribution $(\alpha_1, \ldots, \alpha_d)$ as in Fact 6.2.26. Now we pair up each sequence x with its reverse, say x'. These sequences occur with equal probability. Next, let $\lambda(x)$ be the partition corresponding to x as defined in Fact 6.2.26. Note that by construction, $\lambda(x') \succ \lambda(x)^T$. On the other hand, for any partition $\lambda \vdash n$, we claim that

$$\max\left(\sum_{i=1}^{n} \lambda_i^2, \sum_{i=1}^{n} (\lambda_i^T)^2\right) \geqslant \frac{n^{1.5}}{4}.$$

Once we prove the above we are done, since x and x' occur with the same probability and then we would get that either $\lambda(x)$ or $\lambda(x')$ satisfy the desired property. To see why the above holds, if $\lambda_1 + \cdots + \lambda_{\sqrt{n}/2} \geqslant \frac{n}{2}$, then $\sum_{i=1}^{\sqrt{n}/2} \lambda_i^2 \geqslant n^{1.5}/2$ by Cauchy Schwarz. On the other hand, if $\lambda_1 + \cdots + \lambda_{\sqrt{n}/2} \leqslant \frac{n}{2}$, then $\sum_{i>\sqrt{n}/2} \lambda_i \geqslant \frac{n}{2}$ which actually implies that

$$\sum_{i, \lambda_i^T \geqslant \frac{\sqrt{n}}{2}} \lambda_i^T \geqslant \frac{n}{2}.$$

The above then implies that $\sum_{i=1}^{n} (\lambda_i^T)^2 \geqslant \frac{n}{2} \cdot \frac{\sqrt{n}}{2} = \frac{n^{1.5}}{4}$. This completes the proof. □

Claim 6.2.28. Let $\alpha = (\alpha_1, \ldots, \alpha_d)$ be a vector of nonnegative weights summing to 1. Then for $\lambda \sim \mathrm{SW}^n(\alpha)$,

$$\mathbb{E}\left[\sum_{i=1}^{d} \lambda_i^2\right] \leqslant 2((\alpha_1^2 + \cdots + \alpha_d^2)n^2 + n^{1.5})$$

Proof. Define the polynomial

$$p_2^*(\lambda) = \sum_{i=1}^{\ell(\lambda)} \left(\left(\lambda_i - i + \frac{1}{2}\right)^2 - \left(i - \frac{1}{2}\right)^2\right).$$

It is a standard fact (see [169]) that

$$\mathop{\mathbb{E}}_{\lambda \sim SW^n(\alpha)} [p_2^*(\lambda)] = n(n-1) \sum_{i=1}^{d} \alpha_i^2 \,. \tag{6.2}$$

Next, note that

$$\sum_{i=1}^{\ell(\lambda)} (2i-1)\lambda_i = \sum_{j=1}^{\ell(\lambda_j^T)} (\lambda_j^T)^2 \,.$$

Finally, recall from Fact 6.2.26 that λ_1^T has the same distribution as the longest strictly decreasing subsequence of a random sequence of n tokens in $[d]$ drawn independently from the distribution $(\alpha_1, \ldots, \alpha_d)$. This is dominated by the longest decreasing subsequence of a random permutation and it is known (see e.g. [215]) that this has expected value at most $2\sqrt{n}$. Thus,

$$\mathbb{E}[\lambda_1^T] \leqslant 2\sqrt{n}$$

so

$$\mathbb{E}\left[\sum_{i=1}^{\ell(\lambda)} (2i-1)\lambda_i\right] = \mathbb{E}\left[\sum_{j=1}^{\ell(\lambda_j^T)} (\lambda_j^T)^2\right] \leqslant 2n^{1.5} \,.$$

Combining the above with (6.2) implies

$$\mathbb{E}\left[\sum_{i=1}^{d} \lambda_i^2\right] \leqslant 2((\alpha_1^2 + \cdots + \alpha_d^2)n^2 + n^{1.5})$$

as desired. □

6.2.3 Tensor Manipulations

We will also need some general notation for working with vectors, matrices and tensors.

Definition 6.2.29. For a vector $v \in \mathbb{C}^{d^t}$, we define $T(v)$ to be the $d \otimes \cdots \otimes d$ tensor obtained by reshaping v (we assume that this is done in a canonical and consistent way throughout this paper).

Definition 6.2.30. For matrices (or tensors) $A_1, \ldots, A_t$ with all dimensions equal, we denote the symmetric sum

$$\sum_{\text{sym}} A_1 \otimes \cdots \otimes A_t = \sum_{\pi \in S_t} A_{\pi(1)} \otimes \cdots \otimes A_{\pi(t)} \,.$$

If we additionally have positive integers $k_1, \ldots, k_t$, then

$$\sum_{\text{sym}} A_1^{\otimes k_1} \otimes \cdots \otimes A_t^{\otimes k_t} = \sum_{f \in \mathcal{S}_{k_1,\ldots,k_t}} A_{f(1)} \otimes \cdots \otimes A_{f(k_1+\cdots+k_t)} \,.$$

where $\mathcal{S}_{k_1,\ldots,k_t}$ consists of all distinct permutations of the multiset with k_1 elements equal to 1 , k_2 elements equal to 2 and so on.

Definition 6.2.31. Given a tensor $T \in (\mathbb{C}^d)^{\otimes t}$, we index its modes $1, 2, \ldots, t$. For sets $S_1, \ldots, S_k$ that partition $[t]$, we define $F_{S_1,\ldots,S_k}(T)$ to be the order-k tensor whose dimensions are $d^{|S_1|} \times \cdots \times d^{|S_k|}$ and are obtained by flattening the respective modes of T indexed by elements of $S_1, \ldots, S_k$ respectively.

Definition 6.2.32. For a vector $v \in \mathbb{C}^{d^t}$ and integer $1 \leqslant j \leqslant t$, we define $G_j(v)$ to be a $d^j \times d^j$ matrix defined as

$$G_j(v) = \sum_{S \subset [t], |S|=j} \left(F_{S,[t]\setminus S}(T(v))\right) \left(F_{S,[t]\setminus S}(T(v))\right)^\dagger \,.$$

Now we have the following relations.

Fact 6.2.33. *For any vector $v \in \mathbb{C}^{d^t}$, matrix $E \in \mathbb{C}^{d\times d}$ and integer $1 \leqslant k \leqslant t$,*

$$\langle \sum_{\text{sym}} E^{\otimes k} \otimes I_d^{\otimes t-k}, vv^\dagger \rangle = \langle E^{\otimes k}, G_k(v) \rangle \,.$$

Proof. The identity follows immediately from the definitions. □

Lemma 6.2.34. *Let $f_1 \geqslant \ldots \geqslant f_d \geqslant 0$ and $f_1 + \cdots + f_d = t$. Let $v_1, \ldots, v_d$ be an*

orthonormal basis for $\mathbb{C}^d$. Let

$$v = \sum_g w_g (v_{g(1)} \otimes \cdots \otimes v_{g(t)})$$

where $g : [t] \to [d]$ ranges over all functions such that exactly f_j distinct elements are mapped to j for each $j \in [d]$. Then

$$G_1(v) = \left(\sum_g w_g^2 \right) \cdot (f_1 v_1 v_1^\dagger + \cdots + f_d v_d v_d^\dagger) .$$

Proof. Note that for any two distinct functions g, g' with the specified frequencies, for any index j,

$$F_{\{j\},[t]\setminus\{j\}}(v_{g(1)} \otimes \cdots \otimes v_{g(t)}) F_{\{j\},[t]\setminus\{j\}}(v_{g'(1)} \otimes \cdots \otimes v_{g'(t)})^\dagger = v_{g(j)} v_{g'(j)}^\dagger \cdot \prod_{l \neq j} \langle v_{g(l)}, v_{g'(l)} \rangle = 0$$

because g, g' must differ on at least two inputs $l \in [t]$ (since their outputs have the same frequencies) so then $\langle v_{g(l)}, v_{g'(l)} \rangle = 0$ for some $l \in [t]\setminus\{j\}$. Next, using the same formula as above but with $g = g'$,

$$F_{\{j\},[t]\setminus\{j\}}(v_{g(1)} \otimes \cdots \otimes v_{g(t)}) F_{\{j\},[t]\setminus\{j\}}(v_{g(1)} \otimes \cdots \otimes v_{g(t)})^\dagger = v_{g(j)} v_{g(j)}^\dagger .$$

Writing out the definition of v and combining the above two identities completes the proof of the desired equality. □

As a consequence of Lemma 6.2.34, we have:

Corollary 6.2.35. *Let d be an integer and $\lambda \vdash t$ be a partition. Let $v_{\lambda,j}^d \in \mathbb{C}^{d^t}$ be one of the vectors as defined in Definition 6.2.16. Then*

$$G_1(v_{\lambda,j}^d) = \mathrm{diag}(\lambda_1, \ldots, \lambda_d) .$$

Furthermore, for any unitary $U \in \mathbb{C}^{d\times d}$,

$$G_1(U^{\otimes t} v^d_{\lambda,j}) = U \operatorname{diag}(\lambda_1, \ldots, \lambda_d) U^\dagger .$$

Proof. This statement follows immediately from Lemma 6.2.15 and Lemma 6.2.34. □

6.2.4 Additional Facts

In this section, we present a few additional facts that will be used later in the proof. First, we give an expression for the symmetric polynomials as a linear combination of products of power sums.

Lemma 6.2.36. *Let $x = (x_1, \ldots, x_d)$ be a d-tuple of variables and let $\lambda \vdash n$ be a partition of n into d parts (possibly including zeros). For each integer k, let $S_k = x_1^k + \cdots + x_d^k$. Then we can write*

$$\mathsf{count}(\lambda) \sum_{\text{sym}} x^\lambda = \sum_{\substack{\mu \vdash n \\ \mu \succeq \lambda}} c_\mu \prod_{\mu_i > 0} S_{\mu_i}$$

for some coefficients c_μ satisfying $|c_\mu| \leqslant (2n^2)^{(\ell(\lambda)-\ell(\mu))}$ (recall count *is defined in Definition 6.2.5).*

Proof. We prove the statement by induction on λ where we induct on the reverse "majorizing order". The base case is when $\lambda = (n, 0, \ldots, 0)$, which is the (unique) maximal partition according to the majorizing order. Now assume that we have proved the claim for all λ' that strictly majorize λ. Now note that we can write

$$\prod_{\lambda_i > 0} S_{\lambda_i} = \sum_{\substack{\mu \vdash n \\ \mu \succeq \lambda}} z_\mu \sum_{\text{sym}} x^\mu$$

for some coefficients z_μ that are nonnegative integers. Furthermore, the coefficient z_λ

is exactly equal to $\mathsf{count}(\lambda)$. Thus,

$$\mathsf{count}(\lambda) \sum_{\mathrm{sym}} x^{\lambda} - \prod_{\lambda_i > 0} S_{\lambda_i}$$

is a symmetric polynomial whose monomials strictly majorize λ and we can now apply the inductive hypothesis to write $\mathsf{count}(\lambda) \sum_{\mathrm{sym}} x^{\lambda}$ in the desired form. Using the inductive hypothesis, we can bound the coefficients. First $c_\lambda = 1$ by construction. Next, for $\mu \succ \lambda$, we can bound c_μ as

$$|c_\mu| \leqslant \sum_{\substack{\mu' \vdash n \\ \mu \succeq \mu' \succ \lambda}} \frac{(2n^2)^{\ell(\mu') - \ell(\mu)} z_{\mu'}}{\mathsf{count}(\mu')} \leqslant (2n^2)^{\ell(\lambda) - \ell(\mu)} \sum_{\substack{\mu' \vdash n \\ \mu' \succ \lambda}} \frac{z_{\mu'}}{\mathsf{count}(\mu')(2n^2)^{\ell(\lambda) - \ell(\mu')}} \,.$$

Let $\ell(\lambda) = k$ and fix some μ' with $\ell(\mu') = k'$. Now recall that $z_{\mu'}$ is the coefficient of $x^{\mu'}$ in the expansion $S_{\lambda_1} \cdots S_{\lambda_k}$. It is 0 unless $k' < k$ (since we are assuming $\mu' \succ \lambda$). Also, $z_{\mu'}/\mathsf{count}(\mu')$ is at most equal to the number of partitions of the set $\{\lambda_1, \ldots, \lambda_k\}$ into k' parts such that the sums of the parts of the partition are exactly $\mu'_1, \ldots, \mu'_{k'}$. Thus, $\sum_{\ell(\mu') = k'} z_{\mu'}/\mathsf{count}(\mu')$ is at most equal to the number of ways to partition $[k]$ into exactly k' disjoint subsets. This is at most equal to the number of labeled forests on k vertices with $k - k'$ edges which is at most $k^{2(k-k')} \leqslant (n^2)^{\ell(\lambda) - \ell(\mu')}$. Substituting this into the above sum completes the proof. □

We also have the following formulas for integrals over Haar random unitaries.

Claim 6.2.37. Let $v \in \mathbb{C}^d$ be a uniformly random unit vector. Let v_1 be its first entry and v_2 be its second entry. Then

$$\begin{aligned}
\mathbb{E}[|v_1|^2] &= \mathbb{E}[|v_2|^2] = \frac{1}{d} \\
\mathbb{E}[|v_1|^4] &= \frac{2}{d(d+1)} \\
\mathbb{E}[|v_1|^2 |v_2|^2] &= \frac{1}{d(d+1)} \,.
\end{aligned}$$

Proof. These follow from direct computation using formulas for Haar integrals (see

[107]). □

Claim 6.2.38. Let $X, Y \in \mathbb{C}^{d\times d}$ be Hermitian matrices. Then

$$\mathop{\mathbb{E}}_{U}[(U^\dagger XU)\langle U^\dagger XU, Y\rangle] = \frac{1}{d^2-1}\left(\|X\|_F^2 - \frac{\operatorname{tr}(X)^2}{d}\right)\left(Y - \frac{\operatorname{tr}(Y)I}{d}\right) + \frac{\operatorname{tr}(X)^2\operatorname{tr}(Y)I}{d^2}$$

where the expectation is over a Haar random unitary U.

Proof. First by Claim 6.2.37, for a random unit vector v,

$$\mathop{\mathbb{E}}_{v}[vv^\dagger\langle vv^\dagger, Y\rangle] = \frac{\operatorname{tr}(Y)I + Y}{d(d+1)}$$

and for Haar random orthonormal unit vectors v_1, v_2,

$$\mathop{\mathbb{E}}_{\substack{v_1,v_2\\ v_1\perp v_2}}[v_1v_1^\dagger\langle v_2v_2^\dagger, Y\rangle] = \mathop{\mathbb{E}}_{\substack{v_1,v_2\\ v_1\perp v_2}}\left[v_1v_1^\dagger\left\langle \frac{I - v_1v_1^\dagger}{d-1}, Y\right\rangle\right] = \frac{d\operatorname{tr}(Y)I - Y}{(d-1)d(d+1)}\,.$$

Now let the eigenvalues of X be $x_1, \ldots, x_d$ and the eigenvalues of Y be $y_1, \ldots y_d$. We have

$$\begin{aligned}
\mathop{\mathbb{E}}_{U}[(U^\dagger XU)\langle U^\dagger XU, Y\rangle] &= \mathop{\mathbb{E}}_{v}\left[\left(\sum_j x_j^2\right)vv^\dagger\langle vv^\dagger, Y\rangle\right] + \mathop{\mathbb{E}}_{\substack{v_1,v_2\\ v_1\perp v_2}}\left[\left(2\sum_{j_1<j_2}x_{j_1}x_{j_2}\right)v_1v_1^\dagger\langle v_2v_2^\dagger, Y\rangle\right]\\
&= \|X\|_F^2\cdot\frac{\operatorname{tr}(Y)I + Y}{d(d+1)} + (\operatorname{tr}(X)^2 - \|X\|_F^2)\frac{d\operatorname{tr}(Y)I - Y}{(d-1)d(d+1)}\\
&= \frac{1}{d^2-1}\left(\|X\|_F^2 - \frac{\operatorname{tr}(X)^2}{d}\right)\left(Y - \frac{\operatorname{tr}(Y)I}{d}\right) + \frac{\operatorname{tr}(X)^2\operatorname{tr}(Y)I}{d^2}\,.
\end{aligned}$$

□

6.3 Learning Balanced States

Now, we will present our learning algorithm for states that are close to maximally mixed. The main result that we prove in this section is as follows:

Theorem 6.3.1. *Let $\rho = \frac{I_d}{d} + E$ be an unknown quantum state in $\mathbb{C}^{d\times d}$. Let t be*

a parameter such that $t \leqslant d^2$ *and* $\|E\|_F \leqslant (0.01/t)^4$. *Then for any target accuracy* ε, *there is an algorithm (Algorithm 6.3.2) that takes* $O(d^2/(t^{1.5}\varepsilon^2))$ *copies of* $\rho^{\otimes t}$ *and returns* $\widehat{E} \in \mathbb{C}^{d\times d}$ *such that*

$$\mathbb{E}\left[\|E - \widehat{E}\|_F^2\right] \leqslant O(t^3\|E\|_F^4 + \varepsilon^2)\,.$$

Our algorithm will make use of Keyl's POVM (recall Definition 6.2.19).

Algorithm 6.3.2 (Algorithm for Learning Balanced States)**.**

Input: m copies of $\rho^{\otimes t}$ for some unknown quantum state $\rho \in \mathbb{C}^{d\times d}$;
for $j \in [m]$ **do**
    Measure $\rho^{\otimes t}$ according to Keyl's POVM ;
    Let $\lambda \vdash t$ be the partition and U be the unitary obtained from the measurement ;
    Set $D_j = U\,\mathrm{diag}(\lambda_1/t, \ldots, \lambda_d/t)U^\dagger$;
Compute $\theta = \mathbb{E}_{\lambda\sim \mathrm{SW}_d^t}[\sum_j \lambda_j^2] - (t^2/d)$;
Compute $\widehat{E} = \frac{t(d^2-1)}{d\theta}\left(\frac{D_1+\cdots+D_m}{m} - \frac{I_d}{d}\right)$;
Output: $\widehat{E}$;

We can actually use Theorem 6.3.1 to learn any state ρ with all eigenvalues at most $4/d$ * by first constructing an estimate σ for the "complement" $\frac{4}{3d}I_d - \frac{1}{3}\rho$ and then simulating measurement access to copies $(\rho + 3\sigma)/4$ and applying Theorem 6.3.1 to refine the error in the estimate σ. To initially estimate σ, we will use the standard algorithm for state tomography with unentangled measurements.

Theorem 6.3.3 ([101])**.** *For any* $\delta, \varepsilon < 1$ *and unknown state* $\rho \in \mathbb{C}^{d\times d}$, *there is an algorithm that makes unentangled measurements on* $O(d^2\log(1/\delta)/\varepsilon^2)$ *copies of* ρ *and with* $1-\delta$ *probability outputs a state* $\widehat{\rho}$ *such that* $\|\rho - \widehat{\rho}\|_F \leqslant \varepsilon$.

*the choice of the constant 4 is arbitrary

In this way, we can prove the following corollary of Theorem 6.3.1. It is actually slightly more general in that we can learn any state ρ that can be written as a sum $\sigma + \Delta$ where $\|\sigma\| \leqslant 4/d$ and $\|\Delta\|_F$ is small.

Corollary 6.3.4. *Let $\delta, \varepsilon < 1$ and $\rho \in \mathbb{C}^{d\times d}$ be an unknown state. Let $t \leqslant \min(d^2, c(1/\varepsilon)^{0.2})$ be some parameter, where c is a sufficiently small absolute constant. Assume that ρ can be written as $\rho = \rho' + \Delta$ where $\|\rho'\| \leqslant 4/d$ and $\|\Delta\|_F \leqslant \sqrt{\varepsilon}/t^2$. Then there is an algorithm that takes $O(d^2 \log(1/\delta)/(\sqrt{t}\varepsilon^2))$ total copies of ρ and measures them in batches of t-entangled copies and with probability $1-\delta$, outputs a state $\widehat{\rho}$ such that $\|\rho - \widehat{\rho}\|_F \leqslant \varepsilon$.*

Proof. In the first step, we run the algorithm in Theorem 6.3.3 with $O(d^2 \log(1/\delta)/(\sqrt{t}\varepsilon^2))$ samples to produce an estimate $\hat{\rho}$ satisfying $\|\hat{\rho} - \rho\|_F \leqslant O(\varepsilon t^{1/4})$. Project this (with respect to the Frobenius norm) to the convex set of density matrices with operator norm at most $4/d$ to produce a new estimate $\widetilde{\rho}$. Note that $\|\widetilde{\rho} - \rho'\|_F \leqslant \|\hat{\rho} - \rho'\|_F$, and

$$\begin{aligned}\|\widetilde{\rho} - \rho\|_F \leqslant \|\widetilde{\rho} - \rho'\|_F + \|\Delta\|_F \leqslant \|\widehat{\rho} - \rho'\|_F + \|\Delta\|_F \leqslant \|\widehat{\rho} - \rho\|_F + 2\|\Delta\|_F \\ \leqslant O(\varepsilon t^{1/4} + \sqrt{\varepsilon}/t^2) \leqslant O(\sqrt{\varepsilon}/t^2)\,,\end{aligned}$$

where in the last step we used the assumption that $t \leqslant c(1/\varepsilon)^{0.2}$. Because $\|\widetilde{\rho}\| \leqslant 4/d$ by design, the "complement" state $\sigma \triangleq \frac{4}{3d} I_d - \frac{1}{3}\widetilde{\rho}$ is a valid density matrix. We would like to apply Theorem 6.3.1 to the state $(\rho + 3\sigma)/4$. Note that

$$\frac{1}{4}(\rho + 3\sigma) = \frac{1}{d} I_d + \frac{1}{4}(\rho - \widetilde{\rho})\,, \tag{6.3}$$

and for $E \triangleq (\rho - \widetilde{\rho})/4$, we have $\|E\|_F \leqslant O(\sqrt{\varepsilon}/t^2) \leqslant (0.01/t)^4$ by the assumption that $t \leqslant c(1/\varepsilon)^{0.2}$, provided we take c sufficiently small. Theorem 6.3.1 thus implies that with $O(d^2/(t^{1.5}\varepsilon^2))$ copies of $((\rho + 3\sigma)/4)^{\otimes t}$, measured in batches of t-entangled copies, we can use Algorithm 6.3.2 to produce an estimate $\widehat{E}$ for which

$$\|E - \widehat{E}\|_F \leqslant O(t^{3/2}\|E\|_F^2 + \varepsilon) \leqslant O(\varepsilon) \tag{6.4}$$

with probability at least $1 - \delta$, where in the last step we used that $\|E\|_F^2 \leqslant O(\varepsilon/t^4)$. By Lemma 6.3.5, we can simulate t-entangled measurement access to $(\rho + 3\sigma)/4$ using just t-entangled measurement access to ρ.

Our estimate for ρ is given by taking $\widetilde{\rho} + 4\widehat{E}$ and projecting to the convex set of density matrices. As this projection can only decrease the Frobenius distance to ρ, the proceeding arguments imply that

$$\|\widetilde{\rho} + 4\widehat{E} - \rho\|_F \leqslant O(\varepsilon)$$

with constant probability with using $O(d^2/(\sqrt{t}\varepsilon^2))$ copies. To achieve failure probability $1 - \delta$, we can simply repeat this process $O(\log 1/\delta)$ times and take the geometric median of the outputs. Standard arguments then imply that this is $O(\varepsilon)$ close to ρ with probability $1 - \delta$. Rescaling ε by an appropriate constant factor completes the proof. □

Lemma 6.3.5. *Let $0 \leqslant \lambda \leqslant 1$. Given t copies of an unknown state ρ, and given a description of a density matrix σ, it is possible to simulate any measurement of $(\lambda\rho + (1-\lambda)\sigma)^{\otimes t}$ using a measurement of $\rho^{\otimes t}$.*

Proof. Let $\{M_z\}_{z \in \mathcal{Z}}$ be an arbitrary POVM. Then for any $z \in \mathcal{Z}$, observe that

$$\langle M_z, (\lambda\rho + (1-\lambda)\sigma)^{\otimes t} \rangle = \sum_{S \subseteq [t]} \lambda^{|S|}(1-\lambda)^{t-|S|} \cdot \langle M_z, \rho \otimes_S \sigma \rangle \,, \tag{6.5}$$

where $\rho \otimes_S \sigma$ denotes the product state such that for every $i \in S$, the i-th component of the state is a copy of ρ, and every other component is a copy of σ. To simulate measuring $(\lambda\rho + (1-\lambda)\sigma)$ with $\{M_z\}_{z \in \mathcal{Z}}$, we can simply sample S by including each $i \in [t]$ in S with probability λ, prepare the state $\rho \otimes_S \sigma$, and measure with $\{M_z\}$. □

The remainder of this section will be devoted to proving Theorem 6.3.1.

6.3.1 Approximation for Rotationally Invariant POVMs

Recall that our goal in this section is to learn ρ when $\rho = I_d/d + E$ for some E that is sufficiently small. One important insight in the analysis of Algorithm 6.3.2 is that because Keyl's POVM is invariant to rotating all of the copies simultaneously, we can replace $X = \rho^{\otimes t} = (I_d/d + E)^{\otimes t}$ with its first order approximation $X' = (I_d/d)^{\otimes t} + \sum_{\text{sym}} E \otimes (I_d/d)^{\otimes t-1}$ at the cost of some small error. We can then analyze the POVM applied to X' instead which is significantly simpler (although X' may not technically be a state).

First, we have the following lemma for bounding quantities that involve averaging over simultaneous Haar random rotations of all copies.

Lemma 6.3.6. *Let $v \in \mathbb{C}^{d^t}$ be any vector with $\|v\| = 1$. Let $E \in \mathbb{C}^{d \times d}$ be a Hermitian matrix with $\operatorname{tr}(E) = 0$. Then*

$$\left| \mathop{\mathbb{E}}_{U} [\langle (UEU^\dagger)^{\otimes t}, vv^\dagger \rangle] \right| \leqslant \frac{(4t)^{4t} \|E\|_F^t}{d^t}$$

where U is a Haar random unitary.

Proof. Let the eigenvalues of E be $\alpha_1, \ldots, \alpha_d$. Then

$$\mathop{\mathbb{E}}_{U} [\langle (U^\dagger EU)^{\otimes t}, vv^\dagger \rangle]$$

is a symmetric, degree t polynomial in the α_i since U is Haar random. Now consider any tuple of nonnegative integers $\lambda = (\lambda_1, \ldots, \lambda_d)$ with $\lambda_1 \geqslant \ldots \geqslant \lambda_d$ and $\lambda_1 + \cdots + \lambda_d = t$. For each integer k, define $S_k = \alpha_1^k + \cdots + \alpha_d^k$. Use Lemma 6.2.36 to write

$$\mathsf{count}(\lambda) \sum_{\text{sym}} \alpha^\lambda = \sum_{\substack{\mu \vdash t \\ \mu \succeq \lambda}} c_\mu \prod_{\mu_i > 0} S_{\mu_i} .$$

Next note that $S_1 = 0$ and for all $k \geqslant 2$, $S_k \leqslant (\alpha_1^2 + \cdots + \alpha_d^2)^{k/2}$. Thus,

$$\left| \mathsf{count}(\lambda) \sum_{\mathrm{sym}} \alpha^\lambda \right| \leqslant (\alpha_1^2 + \cdots + \alpha_d^2)^{t/2} (10t^2)^t$$

where we used the coefficient bound in Lemma 6.2.36 and Fact 6.2.4. Next, recall that by Schur Weyl duality, we can write

$$\mathop{\mathbb{E}}_U [(U^\dagger E U)^{\otimes t}] = \sum_{\lambda \vdash t} \frac{s_\lambda(\alpha_1, \ldots, \alpha_d)}{\dim(V_\lambda^t)} \Pi_\lambda^t \,.$$

All of the Schur subspaces V_λ^t have dimension at least $(d/t)^t$ (to see this, note that if $t < d$, then there are at least $\binom{d}{t}$ distinct SSYT of any given shape and otherwise the claim is vacuously true). Thus, since $\|v\| = 1$, we can write

$$\left| \mathop{\mathbb{E}}_U [\langle (U^\dagger E U)^{\otimes t}, v v^\dagger \rangle] \right| = \sum_\lambda z_\lambda s_\lambda(\alpha_1, \ldots, \alpha_d) \tag{6.6}$$

where the coefficients z_λ satisfy $\sum_\lambda |z_\lambda| \leqslant (t/d)^t$. Finally, note that the Schur polynomial $s_\lambda(\alpha_1, \ldots, \alpha_d)$ has positive coefficients and these are all dominated, monomial by monomial, by the coefficients in $(\alpha_1 + \cdots + \alpha_d)^t$ which are all at most t^t. Thus, overall we conclude by Eq. (6.6) that

$$\left| \mathop{\mathbb{E}}_U [\langle (U^\dagger E U)^{\otimes t}, v v^\dagger \rangle] \right| \leqslant |\{\lambda : \lambda \vdash t\}| \cdot t^t \cdot (t/d)^t \cdot (10t^2)^t (\alpha_1^2 + \cdots + \alpha_d^2)^{t/2} \leqslant \frac{(4t)^{4t} \|E\|_F^t}{d^t} \,.$$

□

We now define the family of POVMs in $\mathbb{C}^{d^t \times d^t}$ that are invariant to rotating all of the $d \times d$ copies simultaneously by the same unitary.

Definition 6.3.7. We say a POVM $\{M_z\}_{z \in \mathcal{Z}}$ in $\mathbb{C}^{d^t \times d^t}$ is copy-wise rotationally invariant if it is equivalent to

$$\{U^{\otimes t} M_z (U^\dagger)^{\otimes t} dU\}_{z \in \mathcal{Z}}$$

where $U \in \mathbb{C}^{d\times d}$ is a random unitary drawn from the Haar measure.

Next, we bound a χ^2-like distance between the outcome distributions from measuring X and X' with a copy-wise rotationally invariant POVM. Again, we emphasize that the low-degree truncation X' is not necessarily an actual density matrix; nevertheless, the expression on the left-hand side of Eq. (6.7) below is a well-defined quantity.

Lemma 6.3.8. *Let $\{M_z\}_{z\in\mathcal{Z}}$ be a POVM in $\mathbb{C}^{d^t\times d^t}$ that is copywise rotationally invariant. Let $E \in \mathbb{C}^{d\times d}$ be a matrix with $\operatorname{tr}(E) = 0$ and $\|E\|_F \leqslant \left(\frac{0.01}{t}\right)^4$. Let $X = (I_d/d + E)^{\otimes t}$ and let $X' = (I_d/d)^{\otimes t} + \sum_{\mathrm{sym}} E \otimes (I_d/d)^{\otimes t-1}$. Then*

$$\int_{\mathcal{Z}} \frac{d^t \langle X - X', M_z\rangle^2}{\operatorname{tr}(M_z)} \,\mathrm{d}z \leqslant (100t)^4 \|E\|_F^4 \,. \tag{6.7}$$

Proof. We write

$$X = X' + \sum_{j=2}^{t} \sum_{\mathrm{sym}} E^{\otimes j} \otimes (I_d/d)^{\otimes t-j} \,.$$

Now we will first bound

$$\begin{aligned}
&\int_{\mathcal{Z}} \frac{d^t \langle M_z, E^{\otimes j} \otimes (I_d/d)^{\otimes t-j}\rangle^2}{\operatorname{tr}(M_z)} \,\mathrm{d}z \\
&= \int_{\mathcal{Z}} \frac{d^t \langle M_z \otimes M_z, E^{\otimes j} \otimes (I_d/d)^{\otimes t-j} \otimes E^{\otimes j} \otimes (I_d/d)^{\otimes t-j}\rangle}{\operatorname{tr}(M_z)} \,\mathrm{d}z \\
&= \int_{\mathcal{Z}} \mathop{\mathbb{E}}_{U} \left[\frac{d^t \langle M_z \otimes M_z, U^{\otimes 2t}(E^{\otimes j} \otimes (I_d/d)^{\otimes t-j} \otimes E^{\otimes j} \otimes (I_d/d)^{\otimes t-j})(U^\dagger)^{\otimes 2t}\rangle}{\operatorname{tr}(M_z)}\right] \mathrm{d}z \,,
\end{aligned}$$

where the expectation is over Haar random unitaries $U \in \mathbb{C}^{d\times d}$ (the last step is valid because we assumed that the POVM is copywise rotationally invariant). Now we can use Lemma 6.3.6 to conclude that the above is at most

$$\int_{\mathcal{Z}} d^t \operatorname{tr}(M_z) \cdot \frac{(8j)^{8j}\|E\|_F^{2j}}{d^{2t}} \,\mathrm{d}z = (8j)^{8j}\|E\|_F^{2j} \,.$$

Thus, by Cauchy Schwarz we have

$$\int_{\mathcal{Z}} \frac{d^t \langle X - X', M_z\rangle^2}{\operatorname{tr}(M_z)} \,\mathrm{d}z$$

$$\leqslant \int_{\mathcal{Z}} \frac{d^t}{\operatorname{tr}(M_z)} \left(\sum_{S\subset[t],|S|\geqslant 2} \frac{1}{2^{|S|}\binom{t}{|S|}} \right) \left(\sum_{S\subset[t],|S|\geqslant 2} 2^{|S|} \binom{t}{|S|} \langle M_z, E^{\otimes S} \otimes (I_d/d)^{\otimes [t]\setminus S}\rangle^2 \right) \mathrm{d}z$$
$$\leqslant \sum_{j=2}^{t} 2^j \binom{t}{j}^2 (8j)^{8j} \|E\|_F^{2j} \leqslant (100t)^4 \|E\|_F^4 ,$$

where we use the condition that $\|E\|_F \leqslant (0.01/t)^4$ in the last step. □

Now given a copy-wise rotationally invariant POVM, we can consider a family of estimators $f : \mathbb{C}^{d^t\times d^t} \to \mathbb{C}^{d\times d}$ that are rotationally compatible with it.

Definition 6.3.9. Let $\{M_z\}_{z\in\mathcal{Z}}$ be a POVM in $\mathbb{C}^{d^t\times d^t}$ that is copywise rotationally invariant. We say a function $f : \{M_z\}_{z\in\mathcal{Z}} \to \mathbb{C}^{d\times d}$ is rotationally compatible with the POVM if

$$f(U^{\otimes t} M_z (U^\dagger)^{\otimes t}) = U f(M_z) U^\dagger$$

for all $z \in \mathcal{Z}$ and unitary U.

We can apply Lemma 6.3.8 to bound the error of a rotationally compatible estimator between measuring X' and X with some copy-wise rotationally invariant POVM. Note that in Algorithm 6.3.2, the POVM is rotationally invariant and our estimator is rotationally compatible with it.

Lemma 6.3.10. *Let $\{M_z\}_{z\in\mathcal{Z}}$ be a POVM in in $\mathbb{C}^{d^t\times d^t}$ that is copywise rotationally invariant. Let $f : \{M_z\}_{z\in\mathcal{Z}} \to \mathbb{C}^{d\times d}$ be a rotationally compatible estimator such that $\operatorname{tr}(f(M_z)) = 0$ for all $z \in \mathcal{Z}$. Let $X = (I_d/d + E)^{\otimes t}$ and let $X' = (I_d/d)^{\otimes t} + \sum_{\mathrm{sym}} E \otimes (I_d/d)^{\otimes t-1}$. Assume that $\|E\|_F \leqslant \left(\frac{0.01}{t}\right)^4$. Then*

$$\| \int_{\mathcal{Z}} f(M_z) \langle X - X', M_z\rangle \, \mathrm{d}z \|_F \leqslant \frac{10^5 t^2 \|E\|_F^2}{d} \sqrt{\int_{\mathcal{Z}} \frac{\|f(M_z)\|_F^2 \operatorname{tr}(M_z)}{d^t} \, \mathrm{d}z} .$$

Proof. We will upper bound the left-hand side by bounding $\langle A, \int_{\mathcal{Z}} f(M_z)\langle X - X', M_z\rangle \, \mathrm{d}z\rangle$ for all $\|A\|_F \leqslant 1$. Note that the integral on the left-hand side has trace 0, so it suffices to consider A with trace 0.

Consider a matrix $A \in \mathbb{C}^{d\times d}$ with $\operatorname{tr}(A) = 0$. Then

$$\int_{\mathcal{Z}} \langle A, f(M_z)\rangle \langle X - X', M_z\rangle \,\mathrm{d}z \leqslant \left(\int_{\mathcal{Z}} \langle A, f(M_z)\rangle^2 \frac{\operatorname{tr}(M_z)}{d^t}\,\mathrm{d}z\right)^{1/2} \left(\int_{\mathcal{Z}} \frac{d^t \langle X - X', M_z\rangle^2}{\operatorname{tr}(M_z)}\,\mathrm{d}z\right)^{1/2}.$$

To bound the first term above, by Claim 6.2.38, for any traceless Hermitian matrix $Y \in \mathbb{C}^{d\times d}$,

$$\mathop{\mathbb{E}}_{U}[\langle A, U^\dagger Y U\rangle^2] = \frac{\|A\|_F^2 \|Y\|_F^2}{d^2 - 1}$$

where the expectation is over the Haar measure. Thus, since the POVM $\{M_z\}_{z\in\mathcal{Z}}$ is copywise rotationally invariant, we have

$$\int_{\mathcal{Z}} \langle A, f(M_z)\rangle^2 \frac{\operatorname{tr}(M_z)}{d^t}\,\mathrm{d}z \leqslant \int_{\mathcal{Z}} \frac{2\|A\|_F^2 \|f(M_z)\|_F^2}{d^2} \cdot \frac{\operatorname{tr}(M_z)}{d^t}\,\mathrm{d}z\,.$$

Next, by Lemma 6.3.8

$$\int_{\mathcal{Z}} \frac{d^t \langle X - X', M_z\rangle^2}{\operatorname{tr}(M_z)}\,\mathrm{d}z \leqslant (100t)^4 \|E\|_F^4$$

and thus, taking the maximum over all A with $\|A\|_F \leqslant 1$, we get

$$\left\|\int_{\mathcal{Z}} f(M_z)\langle X - X', M_z\rangle \,\mathrm{d}z\right\|_F \leqslant \frac{10^5 t^2 \|E\|_F^2}{d} \sqrt{\int_{\mathcal{Z}} \frac{\|f(M_z)\|_F^2 \operatorname{tr}(M_z)}{d^t}\,\mathrm{d}z}\,. \qquad \square$$

6.3.2 Proof of Theorem 6.3.1

Recall that the high-level idea for proving Theorem 6.3.1 is to replace measurements of $X = \rho^{\otimes t}$ with measurements of X' and then analyze those measurements of X'. The next result allows us to compute the mean of the estimator in Algorithm 6.3.2 if we were able to measure X'.

Corollary 6.3.11. *Let $\{M_{\lambda,U}\}_{\lambda,U}$ be Keyl's POVM where λ ranges over partitions of t and U ranges over unitaries in $\mathbb{C}^{d\times d}$. Let $X' = (I_d/d)^{\otimes t} + \sum_{\mathrm{sym}} E \otimes (I_d/d)^{\otimes t-1}$.*

Then

$$\sum_{\lambda \vdash t} \int U \operatorname{diag}(\lambda_1/t, \ldots, \lambda_d/t) U^\dagger \cdot \langle M_{\lambda,U}, X' \rangle \, \mathrm{d}U = \frac{I_d}{d} + \frac{dE}{t(d^2-1)} \mathop{\mathbb{E}}_{\lambda \sim \mathrm{SW}_d^t} \Big[\sum_{j=1}^{d} \lambda_j^2 - (t^2/d) \Big].$$

Proof. Note that the actual POVM elements of Keyl's POVM are $\dim(V_\lambda^t) U^{\otimes t} M_\lambda (U^\dagger)^{\otimes t}$ where λ ranges over all partitions and U ranges over Haar random unitaries. Now we apply Fact 6.2.33 and Corollary 6.2.35 to the vectors $U^{\otimes t} v_{\lambda,j}$ as j ranges over all of the components of M_λ (recall Definition 6.2.16). Then

$$\begin{aligned} \langle M_{\lambda,U}, X' \rangle &= \sum_{j=1}^{\dim(\lambda)} \langle \dim(V_\lambda^t) U^{\otimes t} v_{\lambda,j} v_{\lambda,j}^\dagger (U^\dagger)^{\otimes t}, X' \rangle \\ &= \dim(\lambda) \dim(V_\lambda^t) \Big(\frac{1}{d^t} + \frac{1}{d^{t-1}} \langle U \operatorname{diag}(\lambda_1, \ldots, \lambda_d) U^\dagger, E \rangle \Big). \end{aligned}$$

So by taking $X = \operatorname{diag}(\lambda_1/t, \ldots, \lambda_d/t)$ and $Y = E$ in Claim 6.2.38, for any fixed λ we get

$$\mathop{\mathbb{E}}_U [U \operatorname{diag}(\lambda_1/t, \ldots, \lambda_d/t) U^\dagger \cdot \langle M_{\lambda,U}, X' \rangle] = \dim(\lambda) \dim(V_\lambda^t) \Big(\frac{I_d}{d^{t+1}} + \frac{\sum_j \lambda_j^2 - (t^2/d)}{t(d^2-1)d^{t-1}} E \Big).$$

Now summing over all λ, we conclude

$$\sum_{\lambda \vdash t} \int \operatorname{diag}(\lambda_1/t, \ldots, \lambda_d/t) U^\dagger \cdot \langle M_{\lambda,U}, X' \rangle \, \mathrm{d}U = \frac{I_d}{d} + \frac{dE}{t(d^2-1)} \mathop{\mathbb{E}}_{\lambda \sim \mathrm{SW}_d^t} \Big[\sum_j \lambda_j^2 - (t^2/d) \Big].$$

□

Now we can complete the proof of Theorem 6.3.1.

Proof of Theorem 6.3.1. Note that the POVM in Algorithm 6.3.2 is clearly copywise rotationally invariant and the estimator is rotationally compatible with it. Let us use the shorthand $\{M_z\}_{z \in \mathcal{Z}}$ to denote this POVM and for M_z corresponding to unitary U

and partition λ, we let $f(M_z) = U \operatorname{diag}(\lambda_1/t, \dots, \lambda_d/t)U^\dagger$. We have

$$\mathbb{E}\left[\frac{D_1 + \cdots + D_m}{m}\right] = \int_{\mathcal{Z}} f(M_z)\langle M_z, (I_d/d + E)^{\otimes t}\rangle \, \mathrm{d}z$$

where the expectation is over the randomness of the quantum measurement in Algorithm 6.3.2. We can make the estimator D_j have trace 0 by simply subtracting out I_d/d and adding it back at the end. Thus, by Lemma 6.3.10 and Corollary 6.3.11, recalling the definition of θ in Line 6 of Algorithm 6.3.2, we have

$$\begin{aligned}\|\mathbb{E}\left[\frac{D_1 + \cdots + D_m}{m}\right] - \frac{I_d}{d} - \frac{d\theta E}{t(d^2-1)}\|_F &\leqslant \frac{10^5 t^2 \|E\|_F^2}{d}\sqrt{\int_{\mathcal{Z}} \frac{\|f(M_z)\|_F^2 \operatorname{tr}(M_z)}{d^t}\,\mathrm{d}z} \\ &\leqslant \frac{10^5 t^2 \|E\|_F^2}{d}\,.\end{aligned}$$

Thus, if $\widehat{E}$ is the output of Algorithm 6.3.2, then

$$\|\mathbb{E}[\widehat{E}] - E\|_F \leqslant \frac{10^5 t^3 \|E\|_F^2}{\theta}\,.$$

Next, we compute the variance of the estimator $\widehat{E}$. We have

$$\mathbb{E}\left[\|\widehat{E} - \mathbb{E}[\widehat{E}]\|_F^2\right] \leqslant \frac{d^2t^2}{m\theta^2}\,\mathbb{E}\left[\|D_1 - \mathbb{E}[D_1]\|_F^2\right] \leqslant \frac{d^2t^2}{m\theta^2}\,\mathbb{E}\left[\|D_1\|_F^2\right] = \frac{d^2}{m\theta^2}\,\mathop{\mathbb{E}}_{\lambda\sim \mathrm{SW}^t(\rho)}\Big[\sum_j \lambda_j^2\Big]\,.$$

Now by Claim 6.2.28, we can upper bound $\mathbb{E}_{\lambda\sim \mathrm{SW}^t(\rho)}[\sum_j \lambda_j^2] \leqslant 2(\|\rho\|_F^2 t^2 + t^{1.5}) \leqslant 4t^{1.5}$, where in the last step we used the assumption that $\rho = \frac{I_d}{d} + E$ for $\|E\|_F \leqslant (0.01/t)^4$ and that $t \leqslant 0.01d^2$. While Theorem 6.3.1 is technically stated for $t \leqslant d^2$, for t in the range $0.01d^2 \leqslant t \leqslant d^2$, we can just use $0.01d^2$-entangled measurements instead and this loses at most a constant factor in the total copy complexity. Also by Claim 6.2.27, we have $\theta \geqslant t^{1.5}/4$. Thus, putting everything together, we conclude

$$\mathbb{E}\left[\|\widehat{E} - E\|_F^2\right] \leqslant 2 \cdot 10^{10} t^3 \|E\|_F^4 + \frac{(10d)^2}{mt^{1.5}}\,.$$

□

6.4 Reduction via Quantum Splitting

Now we demonstrate how to generalize our results in Section 6.3 to learn arbitrary states. The main idea will be to take an arbitrary state ρ and construct a state $\mathsf{Split}(\rho)$ that preserves information about ρ and we can simulate measurement access to, but also has bounded operator norm.

6.4.1 Construction of the Quantum Splitting

We formalize the splitting procedure below.

Definition 6.4.1. Let $b_1, \ldots, b_d \in \mathbb{Z}_{\geqslant 0}$. We define $\mathsf{Split}_{b_1,\ldots,b_d}$ to be a linear map that sends any $M \in \mathbb{C}^{d \times d}$ to a square matrix with dimension $2^{d_1} + \cdots + 2^{b_d}$ defined as follows. The rows and columns of $\mathsf{Split}_{b_1,\ldots,b_d}(M)$ are indexed by pairs (j, s) where $j \in [d]$ and $s \in \{0,1\}^{b_j}$ and these are sorted first by j and then lexicographically according to s. Now the entry indexed by row (j_1, s_1) and column (j_2, s_2) is defined as

- If $b_{j_1} \leqslant b_{j_2}$ then the entry is $M_{j_1 j_2}/2^{b_{j_2}}$ if s_1 is a prefix of s_2 and is 0 otherwise
- If $b_{j_1} > b_{j_2}$ then the entry is $M_{j_1 j_2}/2^{b_{j_1}}$ if s_2 is a prefix of s_1 and is 0 otherwise

Now we can define the natural inverse map to Split.

Definition 6.4.2. Given $b_1, \ldots, b_d \in \mathbb{Z}_{\geqslant 0}$ and a matrix $N \in \mathbb{C}^{k \times k}$ where $k = 2^{b_1} + \cdots + 2^{b_d}$, we define $\mathsf{Rec}_{b_1,\ldots,b_d}(N)$ to be a $d \times d$ matrix X obtained as follows. First index the rows of N by pairs (j, s) where $j \in [d]$ and $s \in \{0,1\}^{b_j}$ in sorted order (first by j and lexicographically by s). Now the entry $X_{j_1 j_2}$ is equal to

- If $b_{j_1} \leqslant b_{j_2}$ then it is $\sum_{s \in \{0,1\}^{b_{j_2}}} N_{(j_1, s[:b_{j_1}]),(j_2,s)}$ where $s[:b_{j_1}]$ denotes truncating s to its first b_{j_1} bits.
- If $b_{j_1} > b_{j_2}$ then it is $\sum_{s \in \{0,1\}^{b_{j_1}}} N_{(j_1,s),(j_2,s[:b_{j_2}])}$

The following basic facts about the splitting and its inverse are easily verified through direct computation.

Claim 6.4.3. Let $b_1, \ldots, b_d \in \mathbb{Z}_{\geqslant 0}$. We have the following statements (for any matrices M, N of appropriate dimensions):

- $\mathsf{Rec}_{b_1,\ldots,b_d}(\mathsf{Split}_{b_1,\ldots,b_d}(M)) = M$
- $\|\mathsf{Split}_{b_1,\ldots,b_d}(M)\|_F \leqslant \|M\|_F$
- $\|\mathsf{Rec}_{b_1,\ldots,b_d}(N)\|_F \leqslant 2^{\max(b_1,\ldots,b_d)/2}\|N\|_F$

Proof. The first statement is immediate from the definition since Rec sums up exactly the entries that are equal to $M_{j_1 j_2}/2^{\max(b_{j_1},b_{j_2})}$ in Split. The second statement holds because Split simply splits up each of the entries of $M_{j_1 j_2}$ evenly into multiple entries which can only decrease the Frobenius norm. The last statement holds because each element of $\mathsf{Rec}_{b_1,\ldots,b_d}(N)$ is equal to a sum of at most $2^{\max(b_1,\ldots,b_d)}$ elements of N. □

Now the key fact about the splitting that makes it a useful abstraction in our learning algorithm is that we can actually simulate measurements on t-entangled copies of $\mathsf{Split}_{b_1,\ldots,b_d}(\rho)$ with measurements on t-entangled copies of ρ.

Claim 6.4.4. Given measurement access to $\rho^{\otimes t}$ where $\rho \in \mathbb{C}^{d\times d}$ is a state, $\mathsf{Split}_{b_1,\ldots,b_d}(\rho)$ is a valid state and we can simulate measurement access to access to $\mathsf{Split}_{b_1,\ldots,b_d}(\rho)^{\otimes t}$.

Proof. Note that $\mathsf{Split}_{b_1,\ldots,b_d}(\rho)$ can be constructed by embedding ρ in various different principal submatrices of a $k \times k$ matrix (where $k = 2^{b_1} + \cdots + 2^{b_d}$) and averaging them. In particular, for a string $s \in \{0,1\}^{\max(b_1,\ldots,b_d)}$, we can imagine indexing the rows and columns of the $k \times k$ matrix as in Definition 6.4.1 and embedding ρ in the rows and columns indexed by $(1, s[:b_1]), \ldots, (d, s[:b_d])$ where $s[:b_j]$ denotes truncating s to its first b_j bits. Now averaging these embeddings over all $2^{\max(b_1,\ldots,b_d)}$ choices for s gives exactly $\mathsf{Split}_{b_1,\ldots,b_d}(\rho)$.

Thus, $\mathsf{Split}_{b_1,\ldots,b_d}(\rho)^{\otimes t}$ can be constructed by embedding $\rho^{\otimes t}$ in various different principal submatrices of a $k^t \times k^t$ matrix and averaging them. Thus when we measure $\mathsf{Split}_{b_1,\ldots,b_d}(\rho)^{\otimes t}$ with some POVM in $\mathbb{C}^{k^t\times k^t}$, this is equivalent to averaging the different

embeddings of $\rho^{\otimes t}$ and thus we can actually simulate this measurement by measuring $\rho^{\otimes t}$ with a single POVM. $\square$

6.4.2 Full Algorithm

In this section, we present our full learning algorithm. Note that the quantum splitting procedure requires knowledge of ρ, or at least an estimate of ρ, to be useful. We can obtain such a rough estimate for ρ via tomography with unentangled measurements. First, we show how to learn when we are given this estimate as a black-box. We will then put everything together to prove our main learning result, Theorem 6.4.8.

Lemma 6.4.5. *Let $d, t, \varepsilon, \delta, C$ be parameters and $\rho \in \mathbb{C}^{d\times d}$ be an unknown quantum state. Assume that $t \leqslant \min(d^2, (1/\varepsilon)^{0.2})$. Let ρ' be a quantum state whose description we know such that $\|\rho'\| \leqslant C/d$ and $\|\rho' - \rho\|_F \leqslant \sqrt{\varepsilon}/t^2$. Then there is an algorithm that takes $\widetilde{O}(d^2 \log(1/\delta)/(\sqrt{t}\varepsilon^2))$ copies of ρ and measures them in batches of t-entangled copies and with probability $1-\delta$, outputs a state $\widehat{\rho}$ such that $\|\rho - \widehat{\rho}\|_F \leqslant \sqrt{C}\varepsilon$.*

Proof. WLOG ρ' is diagonal with eigenvalues $\lambda_1 \geqslant \ldots \geqslant \lambda_d$ (otherwise, we apply the unitary that diagonalzies it and work in this basis instead). For each $j \in [d]$ let b_j be the smallest nonnegative integer such that $2^{b_j} \geqslant d\lambda_j$. Note that we must have $2^{b_1} + \cdots + 2^{b_d} \leqslant 4d$. Also, $\mathsf{Split}_{b_1,\ldots,b_d}(\rho')$ is a diagonal matrix with all entries at most $1/d$ so $\|\mathsf{Split}_{b_1,\ldots,b_d}(\rho')\| \leqslant 1/d$. By Claim 6.4.3 and linearity of Split,

$$\|\mathsf{Split}_{b_1,\ldots,b_d}(\rho') - \mathsf{Split}_{b_1,\ldots,b_d}(\rho)\|_F \leqslant \|\rho - \rho'\|_F \leqslant \frac{\sqrt{\varepsilon}}{t^2}\,.$$

Now, by Corollary 6.3.4 (with d replaced with $4d$), given $O(d^2 \log(1/\delta)/(\sqrt{t}\varepsilon^2))$ copies of the state $\mathsf{Split}_{b_1,\ldots,b_d}(\rho)$, we can make measurements entangled over at most t copies and learn a state σ such that

$$\|\sigma - \mathsf{Split}_{b_1,\ldots,b_d}(\rho)\|_F \leqslant 0.1\varepsilon\,.$$

However by Claim 6.4.4, we can simply simulate these measurements with copies

of ρ instead. Finally, once we obtain σ, we simply output $\widehat{\rho} = \mathsf{Rec}_{b_1,\ldots,b_d}(\sigma)$. By Claim 6.4.3,

$$\|\widehat{\rho} - \rho\|_F \leqslant 2^{\max(b_1,\ldots,b_d)/2}(0.1\varepsilon) \leqslant \sqrt{C}\varepsilon$$

by construction, so we are done. □

Sometimes the states that we work with will have large eigenvalues and we won't want to apply Lemma 6.4.5 directly as the $\sqrt{C}$ factor loss in the accuracy could be very large. Instead, we will try to project out the large eigenvalues and learn in the orthogonal complement. Thus, we have the following subroutine.

Corollary 6.4.6. *Let $d, t, \varepsilon, \delta, C$ be parameters and $\rho \in \mathbb{C}^{d\times d}$ be an unknown quantum state. Assume that $C \geqslant 1$ and $t \leqslant \min(d^2, (1/\varepsilon)^{0.2})$. Let ρ' be a quantum state and $P \in \mathbb{C}^{d\times d}$ be a projection matrix whose descriptions we know. Assume that $\mathrm{tr}(P\rho P^\dagger) \geqslant 0.1$, $\|P\rho' P^\dagger\| \leqslant C/d$ and $\|P(\rho' - \rho)P^\dagger\|_F \leqslant \sqrt{\varepsilon}/t^2$. Then there is an algorithm that takes $\widetilde{O}(d^2 \log(1/\delta)/(\sqrt{t}\varepsilon^2))$ copies of ρ, measures them in batches of t-entangled copies, and with probability $1 - \delta$ outputs a PSD matrix $\widehat{\rho}$ such that $\|P(\rho - \widehat{\rho})P^\dagger\|_F \leqslant O(\sqrt{C}\varepsilon)$.*

Proof. First estimate $\mathrm{tr}(P\rho P^\dagger)$ by making (unentangled) measurements on $O(\log(1/\delta)/\varepsilon^2)$ copies of ρ with the POVM $(P, I_d - P)$ and letting β be the fraction with outcome P. With probability $1 - 0.1\delta$, we have that $|\beta - \mathrm{tr}(P\rho P^\dagger)| \leqslant 0.01\varepsilon$ so in particular $\beta \geqslant 0.09$.

Now we can assume $t \geqslant 100$ as otherwise we can simply run the algorithm in Theorem 6.3.3 which uses unentangled measurements. Now set $t' = t\beta/2$ and $\varepsilon' = \varepsilon/\beta$. We apply Lemma 6.4.5 with parameters t', ε' to learn the state

$$\rho_0 = \frac{P\rho P^\dagger}{\mathrm{tr}(P\rho P^\dagger)}\,.$$

To do this, for each copy, of ρ, we first measure with the POVM $(P, I_d - P)$ and keep only those whose outcome is P. Now we can simulate measurement access to the state

ρ_0. Note that if we have t-entangled copies of ρ, then with probability at least $1/2$, when we measure them with $(P, I_d - P)$, at least t' of them will have outcome P so then we can make an entangled measurement on t' copies of ρ_0. In other words, if we have batches of the form $\rho^{\otimes t}$, then with at least half of the batches, we will be able to simulate measurements of $\rho_0^{\otimes t'}$.

By Lemma 6.4.5, with probability $1 - 0.1\delta$, we obtain a state $\widetilde{\rho}$ such that

$$\|\widetilde{\rho} - \frac{P\rho P^\dagger}{\mathrm{tr}(P\rho P^\dagger)}\|_F \leqslant 10^3\sqrt{C}\varepsilon\,.$$

Clearly, we can ensure that the state $\widetilde{\rho}$ lives entirely in the subspace given by P. Now we simply output $\widehat{\rho} = \beta\widetilde{\rho}$. We have that

$$\|P(\rho - \widehat{\rho})P^\dagger\|_F = \|\widehat{\rho} - P\rho P^\dagger\|_F \leqslant \|(\beta - \mathrm{tr}(P\rho P^\dagger))\widetilde{\rho}\|_F + \|\mathrm{tr}(P\rho P^\dagger)\widetilde{\rho} - P\rho P^\dagger\|_F \leqslant O(\sqrt{C}\varepsilon)$$

so we are done. □

Now we present our full learning algorithm. At a high level, we first replace ρ with $\sigma = (\rho + I_d/d)/2$ and learn a rough estimate $\widehat{\sigma}_0$ via unentangled measurements. Then we restrict to subspaces corresponding to eigenvalues of $\widehat{\sigma}_0$ at various scales and apply Corollary 6.4.6 to refine our estimate on each of these subspaces. Finally, we aggregate our estimates to obtain a refined estimate $\widehat{\sigma}$, from which we can recover an estimate of ρ. Note that previously ε was used to measure the accuracy in Frobenius norm but in Algorithm 6.4.7, it will be used for accuracy in trace norm. Our algorithm only guarantees recovery to trace norm ε with $d^3/(\sqrt{t}\varepsilon^2)$ samples (instead of a stronger Frobenius norm guarantee of $\varepsilon/\sqrt{d}$ which is what we would get in the balanced case) due to the $\sqrt{C}$ factor loss in Corollary 6.4.6.

Algorithm 6.4.7 (Full Learning Algorithm)**.**

1: **Input:** Parameters $d, t, \varepsilon, \delta$;

2: **Input:** copies of ρ for some unknown quantum state $\rho \in \mathbb{C}^{d\times d}$;

3: Set $n = \widetilde{O}\left(\frac{d^3 \log(1/\delta)}{\sqrt{t}\varepsilon^2}\right)$;

4: Set $\sigma = (\rho + I_d/d)/2$ in future steps, simulate measurements of σ using the given copies of ρ ;

5: Perform tomography with unentangled measurements on n copies of σ (Theorem 6.3.3) to learn estimate $\widehat{\sigma}_0$ for σ ;

6: Let U be the unitary that diagonalizes $\widehat{\sigma}_0$ so that

$$U\widehat{\sigma}_0 U^\dagger = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$$

with $\lambda_1 \geqslant \ldots \geqslant \lambda_d$;

7: **for** $j = 1, \ldots, \lfloor \log \sqrt{t} \rfloor$ **do**

8: Let P_j be the projection onto the eigenspace corresponding to $\lambda_i \leqslant \sqrt{t}/(2^{j-1}d)$;

9: Take n copies of σ and apply Corollary 6.4.6 with $\rho' \leftarrow \sigma_0, P \leftarrow P_j, C = \sqrt{t}/2^{j-1}, \varepsilon \leftarrow \varepsilon/\sqrt{d}$ to obtain estimate $\widehat{\sigma}_j$;

10: Set $\widehat{\sigma} = \sum_{j=0}^{\lfloor \log \sqrt{t} \rfloor} (P_j \widehat{\sigma}_j P_j^\dagger - P_{j+1} \widehat{\sigma}_j P_{j+1}^\dagger)$ where we define $P_0 = I_d$ and $P_{\lfloor \log \sqrt{t} \rfloor + 1} = 0$;

11: Set $\widehat{\rho} = \frac{\mathsf{trunc}(2\widehat{\sigma} - I_d/d)}{\mathrm{tr}(\mathsf{trunc}(2\widehat{\sigma} - I_d/d))}$ where trunc denotes zeroing out the negative eigenvalues ;

12: **Output:** $\widehat{\rho}$;

Our main theorem is stated below.

Theorem 6.4.8. *Let $d, t, \varepsilon, \delta$ be parameters and $\rho \in \mathbb{C}^{d \times d}$ be an unknown quantum state. If $t \leqslant \min(d^2, (\sqrt{d}/\varepsilon)^{0.2})$, then Algorithm 6.4.7 takes $\widetilde{O}(d^3 \log(1/\delta)/(\sqrt{t}\varepsilon^2))$ copies of ρ and measures them in batches of t-entangled copies and with probability $1 - \delta$, outputs a state $\widehat{\rho}$ such that $\|\rho - \widehat{\rho}\|_1 \leqslant \varepsilon$.*

Proof. By Theorem 6.3.3, with $1 - 0.1\delta$ probability, we have

$$\|\sigma - \widehat{\sigma}_0\|_F \leqslant \frac{\varepsilon t^{0.25}}{\sqrt{d}}\,.$$

Now we verify that the conditions of Corollary 6.4.6 hold whenever we apply it in Algorithm 6.4.7. The condition on t clearly holds. Next, note that all eigenvalues of σ are at least $1/(2d)$. For all $j \geqslant 1$, the dimension of the orthogonal complement of P_j is at most $d/2$ and thus, $\operatorname{tr}(P_j \sigma P_j^\dagger) \geqslant \frac{1}{4}$. Also, by definition, we have $\|P_j \widehat{\sigma}_0 P_j^\dagger\| \leqslant \sqrt{t}/(2^{j-1}d)$. Finally,

$$\|P(\widehat{\sigma}_0 - \sigma)P^\dagger\|_F \leqslant \|\sigma - \widehat{\sigma}_0\|_F \leqslant \frac{\varepsilon t^{0.25}}{\sqrt{d}} \leqslant \frac{\sqrt{\varepsilon/\sqrt{d}}}{t^2}$$

so we can ensure that with probability $1 - 0.1\delta/t$,

$$\|P_j(\widehat{\sigma}_j - \sigma)P_j^\dagger\|_F \leqslant O\left(\frac{t^{0.25}\varepsilon}{2^{j/2}\sqrt{d}}\right)\,.$$

Thus, since P_{j+1} is a projector onto a subspace of P_j, we also have that

$$\|P_j(\widehat{\sigma}_j - \sigma)P_j^\dagger - P_{j+1}(\widehat{\sigma}_j - \sigma)P_{j+1}^\dagger\|_F \leqslant O\left(\frac{t^{0.25}\varepsilon}{2^{j/2}\sqrt{d}}\right)\,.$$

Next, the dimension of the orthogonal complement of P_{j+1} is at most $2^{j+1}d/\sqrt{t}$, so

$$\|P_j(\widehat{\sigma}_j - \sigma)P_j^\dagger - P_{j+1}(\widehat{\sigma}_j - \sigma)P_{j+1}^\dagger\|_1 \leqslant O(\varepsilon)\,.$$

Also, similarly,

$$\|P_0(\widehat{\sigma}_0 - \sigma)P_0^\dagger - P_1(\widehat{\sigma}_0 - \sigma)P_1^\dagger\|_1 \leqslant O(\varepsilon)\,.$$

Putting these together, we get that $\|\sigma - \widehat{\sigma}\|_1 \leqslant O(\varepsilon \log t)$. This immediately implies

$$\|(2\widehat{\sigma} - I_d/d) - \rho\|_1 \leqslant O(\varepsilon \log t)\,.$$

Finally, the truncation and trace normalization to obtain $\widehat{\rho}$ increases the trace norm

distance by at most a constant factor so we conclude $\|\rho - \widehat{\rho}\|_1 \leqslant O(\varepsilon \log t)$. This completes the proof (note that we can simply redefine ε appropriately and absorb the logarithmic factors in the number of copies into the $\widetilde{O}(\cdot)$). □

6.5 Lower Bound Machinery

Recall Definition 6.2.32. The key ingredients in our lower bound involve understanding $G_j(v)$, particularly $G_1(v)$, in terms of properties of the vector $v \in \mathbb{C}^{d^t}$. The next lemma relates $\|G_1(v)\|_F^2$ to the projections of v onto various Schur subspaces of $\mathbb{C}^{d^t}$.

Lemma 6.5.1. *Let $v \in \mathbb{C}^{d^t}$ be a vector with $\|v\| = 1$. Then*

$$\|G_1(v)\|_F^2 \leqslant \sum_{\lambda \vdash t} \|\Pi_\lambda v\|^2 (\lambda_1^2 + \cdots + \lambda_d^2)\,.$$

Proof. Let $A \in \mathbb{C}^{d \times d}$ be an arbitrary Hermitian matrix with $\|A\|_F = 1$. Let its eigenvectors be $v_1, \ldots, v_d$. Now we write v in the basis given by $v_{s_1} \otimes \cdots \otimes v_{s_t}$ i.e.

$$v = \sum_{s \in [d]^t} c_s (v_{s_1} \otimes \cdots \otimes v_{s_t})\,.$$

For each $s \in [d]^t$ and $k \in [d]$, let $f_s(k)$ denote the number of occurrences of k in s. Note that

$$\langle G_1(v), A \rangle = \left\langle \sum_{s \in [d]^t} c_s^2 G_1(v_{s_1} \otimes \cdots \otimes v_{s_t}), A \right\rangle = \left\langle \sum_{s \in [d]^t} c_s^2 \left(\sum_{k=1}^{d} f_s(k) v_k v_k^\dagger \right), A \right\rangle \tag{6.8}$$

where the first step holds because A is diagonal in the basis $v_1, \ldots, v_d$ and all of the cross terms that appear when we expand $G_1(v)$ are off-diagonal. Now the above is at

most

$$\begin{aligned}\|\sum_{s\in[d]^t} c_s^2 \left(\sum_{k=1}^{d} f_s(k) v_k v_k^\dagger\right)\|_F &\leqslant \sqrt{\sum_{s\in[d]^t} c_s^2} \cdot \sqrt{\sum_{s\in[d]^t} c_s^2 \|\sum_{k=1}^{d} f_s(k) v_k v_k^\dagger\|_F^2} \\ &= \sqrt{\sum_{s\in[d]^t} c_s^2 \left(\sum_{k=1}^{d} f_s(k)^2\right)}\end{aligned} \tag{6.9}$$

since by assumption, $\|v\|^2 = \sum_{s\in[d]^t} c_s^2 = 1$. For each s, let f_s denote the partition corresponding to $(f_s(1), \ldots, f_s(d))$ (in sorted order). We have that

$$\sum_{\lambda\vdash t} \|\Pi_\lambda v\|^2 \cdot (\lambda_1^2 + \cdots + \lambda_d^2) = \sum_{k=1}^{t^2} \left(\sum_{\substack{\lambda\vdash t \\ \lambda_1^2+\cdots+\lambda_d^2 \geqslant k}} \|\Pi_\lambda v\|^2\right) \geqslant \sum_{k=1}^{t^2} \left(\sum_{\substack{s\in[d]^t \\ f_s(1)^2+\cdots+f_s(d)^2 \geqslant k}} c_s^2\right)$$

where the last inequality holds because by Lemma 6.2.17, we know that for any s with $f_s(1)^2 + \cdots + f_s(d)^2 \geqslant k$, the vector $v_{s_1} \otimes \cdots \otimes v_{s_t}$ is contained in the space

$$\bigoplus_{\substack{\lambda\vdash t \\ \lambda\trianglerighteq f_s}} \Pi_\lambda \subseteq \bigoplus_{\substack{\lambda\vdash t \\ \lambda_1^2+\cdots+\lambda_d^2 \geqslant k}} \Pi_\lambda \,.$$

However, we have that

$$\sum_{k=1}^{t^2} \left(\sum_{\substack{s\in[d]^t \\ f_s(1)^2+\cdots+f_s(d)^2 \geqslant k}} c_s^2\right) = \sum_{s\in[d]^t} c_s^2 \left(\sum_{k=1}^{d} f_s(k)^2\right)$$

so we get

$$\sum_{\lambda\vdash t} \|\Pi_\lambda v\|^2 \cdot (\lambda_1^2 + \cdots + \lambda_d^2) \geqslant \sum_{s\in[d]^t} c_s^2 \left(\sum_{k=1}^{d} f_s(k)^2\right) . \tag{6.10}$$

Putting (6.8) (6.9), (6.10) together, and taking the maximum over all choices of A, we conclude

$$\|G_1(v)\|_F^2 \leqslant \sum_{\lambda\vdash t} \|\Pi_\lambda v\|^2 (\lambda_1^2 + \cdots + \lambda_d^2)$$

as desired. □

Now we will use Lemma 6.5.1 to bound the "likelihood ratio" $\frac{x^\dagger \rho^{\otimes t} x}{x^\dagger \rho_0^{\otimes t} x}$ for different quantum states ρ and ρ_0 and vectors $x \in \mathbb{C}^{d^t}$. We explain the lower bound framework in detail and why this quantity is meaningful in Section 6.6.

Lemma 6.5.2. *Let $0 < \varepsilon < 1$ be some parameter. Let $\rho_0 = (I_d + Z)/d \in \mathbb{C}^{d\times d}$ be a quantum state with $\|Z\| \leqslant \varepsilon$. Let μ be a distribution on matrices in $\mathbb{C}^{d\times d}$ that is rotationally symmetric i.e. invariant under rotation by any unitary U. Also assume that any Δ in the support of μ has $\operatorname{tr}(\Delta) = 0$ and $\|\Delta\| \leqslant \varepsilon/d$. Let t be an integer with $t \leqslant 0.01/\varepsilon^{0.1}$. Then for any vector $x \in \mathbb{C}^{d^t}$ with $\|x\| = 1$,*

$$\mathop{\mathbb{E}}_{\Delta\sim\mu}\left[\log \frac{x^\dagger(\rho_0+\Delta)^{\otimes t}x}{x^\dagger\rho_0^{\otimes t}x}\right] \geqslant \frac{\frac{1}{d^{t-2}}\mathbb{E}_{\Delta\sim\mu}\left[\langle G_2(x), \Delta\otimes\Delta\rangle\right]}{x^\dagger\rho_0^{\otimes t}x} - \frac{\frac{2}{d^{t-2}}\mathbb{E}_{\Delta\sim\mu}\left[\langle G_1(x),\Delta\rangle^2\right]}{x^\dagger\rho_0^{\otimes t}x} - \frac{30t^3\varepsilon^3}{d}\,.$$

Proof. Using the condition on t, we have that all eigenvalues of $\rho_0^{\otimes t}$ and $(\rho_0+\Delta)^{\otimes t}$ are between $0.98/d^t$ and $1.02/d^t$ so $\frac{x^\dagger(\rho_0+\Delta)^{\otimes t}x}{x^\dagger\rho_0^{\otimes t}x} \in [0.9, 1.1]$. Thus,

$$\log \frac{x^\dagger(\rho_0+\Delta)^{\otimes t}x}{x^\dagger\rho_0^{\otimes t}x} \geqslant \frac{x^\dagger((\rho_0+\Delta)^{\otimes t}-\rho_0^{\otimes t})x}{x^\dagger\rho_0^{\otimes t}x} - \frac{2}{3}\left(\frac{x^\dagger((\rho_0+\Delta)^{\otimes t}-\rho_0^{\otimes t})x}{x^\dagger\rho_0^{\otimes t}x}\right)^2\,. \quad (6.11)$$

For the first term, we can write

$$\begin{aligned}(\rho_0+\Delta)^{\otimes t} - \rho_0^{\otimes t} &= \sum_{j=1}^{t}\sum_{\text{sym}} \Delta^{\otimes j}\otimes\rho_0^{\otimes t-j} \\ &= \sum_{\text{sym}}\Delta\otimes\rho_0^{\otimes t-1} + \sum_{j=2}^{t}\sum_{k=0}^{t-j}\sum_{\text{sym}}\Delta^{\otimes j}\otimes(Z/d)^{\otimes k}\otimes(I_d/d)^{\otimes t-j-k}\,.\end{aligned} \quad (6.12)$$

Let

$$T(\Delta) = \sum_{\text{sym}}\Delta^{\otimes 2}\otimes(I_d/d)^{\otimes t-2}\,.$$

Since μ is supported on traceless matrices and is rotationally invariant, $\mathbb{E}_{\Delta\sim\mu}[\Delta] = 0$. Now by Lemma 6.3.6,

$$\|\mathop{\mathbb{E}}_{\Delta\sim\mu}[\Delta^{\otimes j}]\| \leqslant \mathop{\mathbb{E}}_{\Delta\sim\mu}\left[\frac{(4j)^{4j}\|\Delta\|_F^j}{d^j}\right] \leqslant \frac{(4j)^{4j}\varepsilon^j}{d^{1.5j}}\,. \quad (6.13)$$

Thus, the first sum in the RHS of (6.12) is 0 and aside from T, the remaining terms all contain a product of at least two copies of Δ and one copy of Z or at least three copies of Δ. We can upper bound the norm of the expectation of these terms using (6.13). Combining (6.12), (6.13) and using the condition $t \leqslant 0.01/\varepsilon^{0.1}$, we deduce

$$\mathop{\mathbb{E}}_{\Delta\sim\mu}[x^\dagger((\rho_0+\Delta)^{\otimes t}-\rho_0^{\otimes t})x] \geqslant x^\dagger\left(\mathop{\mathbb{E}}_{\Delta\sim\mu}[T(\Delta)]\right)x - \frac{10t^3\varepsilon^3}{d^{t+1}}\,. \tag{6.14}$$

Next, we lower bound the quadratic term. Further expanding (6.12), we can write

$$\begin{aligned}
&(\rho_0+\Delta)^{\otimes t}-\rho_0^{\otimes t}\\
&=\underbrace{\sum_{\text{sym}}\Delta\otimes(I_d/d)^{\otimes t-1}}_{S_1(\Delta)}+\underbrace{\sum_{k=1}^{t-1}\sum_{\text{sym}}\Delta\otimes(Z/d)^{\otimes k}\otimes(I_d/d)^{\otimes t-1-k}}_{S_2(\Delta)}\\
&\quad+\underbrace{\sum_{j=2}^{t}\sum_{k=0}^{t-j}\sum_{\text{sym}}\Delta^{\otimes j}\otimes(Z/d)^{\otimes k}\otimes(I_d/d)^{\otimes t-j-k}}_{S_3(\Delta)}
\end{aligned}$$

and we label the three sums above as $S_1(\Delta), S_2(\Delta), S_3(\Delta)$. Now let us consider expanding $((\rho_0+\Delta)^{\otimes t}-\rho_0^{\otimes t})\otimes((\rho_0+\Delta)^{\otimes t}-\rho_0^{\otimes t})$. The key is that in the expansion, aside from $S_1(\Delta)\otimes S_1(\Delta)$, all other terms have a product of at least two copies of Δ and one copy of Z or at least three copies of Δ so we can apply (6.13) to upper bound the norms of all of these terms by ε^3/d^{2t+1}. Formally, we get

$$\|\mathop{\mathbb{E}}_{\Delta\sim\mu}\left[\left((\rho_0+\Delta)^{\otimes t}-\rho_0^{\otimes t}\right)\otimes\left((\rho_0+\Delta)^{\otimes t}-\rho_0^{\otimes t}\right)-S_1(\Delta)\otimes S_1(\Delta)\right]\| \leqslant \frac{10t^3\varepsilon^3}{d^{2t+1}}\,.$$

The above implies that

$$\mathop{\mathbb{E}}_{\Delta\sim\mu}\left[(x^\dagger((\rho_0+\Delta)^{\otimes t}-\rho_0^{\otimes t})x)^2\right] \leqslant \mathop{\mathbb{E}}_{\Delta\sim\mu}[(x^\dagger S_1(\Delta)x)^2]+\frac{10t^3\varepsilon^3}{d^{2t+1}}\,. \tag{6.15}$$

Finally, observe that by Fact 6.2.33, $x^\dagger S_1(\Delta)x = \frac{1}{d^{t-1}}\langle G_1(x),\Delta\rangle$ and $x^\dagger T(\Delta)x =$

$\frac{1}{d^{t-2}}\langle G_2(X), \Delta\otimes\Delta\rangle$. Thus, combining (6.11), (6.14), (6.15), we get

$$\mathop{\mathbb{E}}_{\Delta\sim\mu}\left[\log\frac{x^\dagger(\rho_0+\Delta)^{\otimes t}x}{x^\dagger\rho_0^{\otimes t}x}\right] \geqslant \frac{\frac{1}{d^{t-2}}\mathbb{E}_{\Delta\sim\mu}\left[\langle G_2(x),\Delta\otimes\Delta\rangle\right]}{x^\dagger\rho_0^{\otimes t}x} - \frac{\frac{2}{d^{t-2}}\mathbb{E}_{\Delta\sim\mu}\left[\langle G_1(x),\Delta\rangle^2\right]}{x^\dagger\rho_0^{\otimes t}x} - \frac{30t^3\varepsilon^3}{d}$$

where we also used that by the assumption on t, $\rho_0^{\otimes t}$ has all eigenvalues at least $0.9/d^t$. This completes the proof. □

We can aggregate Lemma 6.5.2 over sequences of vectors $x_1,\dots,x_m$ and apply Jensen's inequality to get the following bound on the product of a sequence of likelihood ratios.

Lemma 6.5.3. *Let $0<\varepsilon<1$ be some parameter. Let $\rho_0=(I_d+Z)/d\in\mathbb{C}^{d\times d}$ be a quantum state with $\|Z\|\leqslant\varepsilon$. Let μ be a distribution on matrices in $\mathbb{C}^{d\times d}$ that is rotationally symmetric i.e. invariant under rotation by any unitary U. Also assume that any Δ in the support of μ has $\operatorname{tr}(\Delta)=0$ and $\|\Delta\|\leqslant\varepsilon/d$. Let t be an integer with $t\leqslant 0.01/\varepsilon^{0.1}$. Let $x_1,\dots,x_m\in\mathbb{C}^{d^t}$ be vectors such that*

- $\sum_{\lambda\vdash t}\sum_{j=1}^m \frac{\|\Pi_\lambda x_j\|^2}{d^t x_j^\dagger\rho_0^{\otimes t}x_j}(\lambda_1^2+\cdots+\lambda_d^2)\leqslant A$
- $\|-\binom{t}{2}mI_{d^2}+\frac{1}{d^{t-2}}\sum_{j=1}^m\frac{G_2(x_j)}{x_j^\dagger\rho_0^{\otimes}x_j}\|\leqslant B$

for some parameter β. Then

$$\mathop{\mathbb{E}}_{\Delta\sim\mu}\left[\prod_{j=1}^m\frac{x_j^\dagger(\rho_0+\Delta)^{\otimes t}x_j}{x_j\rho_0^{\otimes t}x_j}\right]\geqslant\exp\left(-\frac{2\varepsilon^2A}{d}-\frac{10^8\varepsilon^2B}{d}-\frac{30mt^3\varepsilon^3}{d}\right).$$

Proof. By Jensen's inequality and Lemma 6.5.2,

$$\mathop{\mathbb{E}}_{\Delta\sim\mu}\left[\prod_{j=1}^m\frac{x_j^\dagger(\rho_0+\Delta)^{\otimes t}x_j}{x_j\rho_0^{\otimes t}x_j}\right]\geqslant \exp\left(\sum_{j=1}^m\left(\frac{\frac{1}{d^{t-2}}\mathbb{E}_{\Delta\sim\mu}\left[\langle G_2(x_j),\Delta\otimes\Delta\rangle\right]}{x_j^\dagger\rho_0^{\otimes t}x_j}-\frac{\frac{2}{d^{t-2}}\mathbb{E}_{\Delta\sim\mu}\left[\langle G_1(x_j),\Delta\rangle^2\right]}{x_j^\dagger\rho_0^{\otimes t}x_j}-\frac{30t^3\varepsilon^3}{d}\right)\right). \tag{6.16}$$

Now by Lemma 6.3.6, $\|\mathbb{E}_{\Delta\sim\mu}[\Delta\otimes\Delta]\|\leqslant\frac{10^8\varepsilon^2}{d^3}$ so combining this with the assumption

about $x_1, \ldots, x_m$ and that $\mathrm{tr}(\Delta) = 0$, we get

$$\sum_{j=1}^{m} \frac{\frac{1}{d^{t-2}} \mathbb{E}_{\Delta\sim\mu} \left[\langle G_2(x_j), \Delta \otimes \Delta\rangle\right]}{x_j^\dagger \rho_0^{\otimes t} x_j} \geqslant -\frac{10^8\varepsilon^2}{d^3}(Bd^2) = \frac{-10^8\varepsilon^2 B}{d}\,.$$

Next, by Claim 6.2.38 and Lemma 6.5.1,

$$\begin{aligned}
\sum_{j=1}^{m} \frac{\frac{2}{d^{t-2}} \mathbb{E}_{\Delta\sim\mu} \left[\langle G_1(x_j), \Delta\rangle^2\right]}{x_j^\dagger \rho_0^{\otimes t} x_j} &\leqslant \frac{2}{d^{t-2}} \sum_{j=1}^{m} \frac{\varepsilon^2}{d^3} \frac{\|G_1(x_j)\|_F^2}{x_j^\dagger \rho_0^{\otimes t} x_j} \\
&\leqslant \frac{2\varepsilon^2}{d^{t+1}} \sum_{j=1}^{m} \sum_{\lambda\vdash t} \frac{\|\Pi_\lambda x_j\|^2(\lambda_1^2 + \cdots + \lambda_d^2)}{x_j^\dagger \rho_0^{\otimes t} x_j} \\
&\leqslant \frac{2A\varepsilon^2}{d}\,.
\end{aligned}$$

Plugging the above two inequalities into (6.16), we get

$$\mathop{\mathbb{E}}_{\Delta\sim\mu} \left[\prod_{j=1}^{m} \frac{x_j^\dagger(\rho_0 + \Delta)^{\otimes t} x_j}{x_j \rho_0^{\otimes t} x_j}\right] \geqslant \exp\left(-\frac{2\varepsilon^2 A}{d} - \frac{10^8\varepsilon^2 B}{d} - \frac{30mt^3\varepsilon^3}{d}\right)$$

as desired. □

6.6 Proof of Lower Bound

6.6.1 Lower Bound Framework

The remainder of the proof of the lower bound will closely follow the framework in Chapter 4 ([62]) with Lemma 6.5.3 as the main new ingredient. Recall that the learner measures $m = n/t$ copies of $\rho^{\otimes t}$ in sequence with POVMs in $\mathbb{C}^{d^t \times d^t}$ possibly chosen adaptively. It is a standard fact that without loss of generality (see e.g. [66, Lemma 4.8]) we may assume that all POVMs used are rank-1 and we will work with this assumption for the rest of the lower bound. We will sometimes represent a sequence of m measurement outcomes by $\boldsymbol{x} = (x_1, \ldots, x_m)$ to denote that in the i-th step, the outcome that was observed corresponds to a POVM element which is a scalar multiple of $x_i x_i^\dagger$.

Next, we have the following standard formalism for representing an adaptive algorithm as a tree which generalizes Definition 4.2.1 to the t-entangled setting.

Definition 6.6.1 (Tree representation, see e.g. [66])**.** Fix an unknown d-dimensional mixed state ρ. An algorithm for state tomography that only uses m batches of t-entangled copies of ρ can be expressed as a pair $(\mathcal{T}, \mathcal{A})$, where $\mathcal{T}$ is a rooted tree $\mathcal{T}$ of depth m satisfying the following properties:

- Each node is labeled by a string of vectors $\boldsymbol{x} = (x_1, \ldots, x_k)$, where each x_i corresponds to measurement outcome observed in the i-th step.

- Each node $\boldsymbol{x}$ is associated with a probability $p^{\rho^{\otimes t}}(\boldsymbol{x})$ corresponding to the probability of observing $\boldsymbol{x}$ over the course of the algorithm. The probability for the root is 1.

- At each non-leaf node, we measure $\rho^{\otimes t}$ using a rank-1 POVM $\left\{\omega_x d^t \cdot xx^\dagger\right\}_x$ to obtain classical outcome (that is a unit vector) $x \in \mathbb{C}^{d^t}$. The children of $\boldsymbol{x}$ consist of all strings $\boldsymbol{x}' = (x_1, \ldots, x_k, x)$ for which x is a possible POVM outcome.

- If $\boldsymbol{x}' = (x_1, \ldots, x_k, x)$ is a child of $\boldsymbol{x}$, then

$$p^{\rho^{\otimes t}}(\boldsymbol{x}') = p^{\rho^{\otimes t}}(\boldsymbol{x}) \cdot \omega_x d^t \cdot x^\dagger \rho^{\otimes t} x\,. \tag{6.17}$$

- Every root-to-leaf path is length-m. Note that $\mathcal{T}$ and ρ induce a distribution over the leaves of $\mathcal{T}$.

$\mathcal{A}$ is a randomized algorithm that takes as input any leaf $\boldsymbol{x}$ of $\mathcal{T}$ and outputs a state $\mathcal{A}(\boldsymbol{x})$. The output of $(\mathcal{T}, \mathcal{A})$ upon measuring m copies of a state ρ in t-entangled batches is the random variable $\mathcal{A}(\boldsymbol{x})$, where $\boldsymbol{x}$ is sampled from the aforementioned distribution over the leaves of $\mathcal{T}$.

We also recall the definition of the Gaussian Unitary Ensemble (GUE) and define a trace-centered variant, which will be the basis of our hard distribution.

Definition 6.6.2 (GUE, Trace-centered GUE)**.** A sample $G \sim \mathrm{GUE}(d)$ is a Hermitian matrix with independent Gaussians on and above the diagonal, with $G_{j,j} \sim \mathcal{N}(0, 2/d)$ and $G_{j_1,j_2} \sim \mathcal{N}(0, 1/d) + i\mathcal{N}(0, 1/d)$ for $j_1 < j_2$. A sample $G' \sim \mathrm{GUE}^*(d)$ is sampled by $G' = G - \mathrm{tr}(G)I_d/d$ where $G \sim \mathrm{GUE}(d)$.

We recall the following standard fact about extremal eigenvalues of the GUE matrix.

Fact 6.6.3. *If $G \sim \mathrm{GUE}^*(d)$, then $\|G\| \leqslant 3$ with probability $1 - e^{-\Omega(d)}$.*

6.6.2 Construction of Hard Distribution

We construct the following hard distribution μ over quantum states. Let $U \subseteq \mathbb{C}^{d\times d}$ be the subspace of Hermitian matrices with trace 1 and $U_0 \subseteq \mathbb{C}^{d\times d}$ be the subspace of Hermitian matrices with trace 0. These spaces inherit the inner product of $\mathbb{C}^{d\times d}$, which defines Lebesgue measures Leb_U and Leb_{U_0} on them. Let ε be the target accuracy. We can assume ε is sufficiently small and let $\sigma = C\varepsilon$ for some constant $C > 1$ to be chosen later. A sample $\rho \sim \mu$ is generated by

$$\rho = \frac{1}{d}(I_d + \sigma G),$$

where G is a sample from $\mathrm{GUE}^*(d)$ conditioned on $\|G\| \leqslant 4$. Note that such matrices are clearly valid quantum states. Concretely, μ has density (with respect to Leb_U)

$$\mu(\rho) = \frac{1}{Z}\exp\left(-\frac{d^3}{4\sigma^2}\|\rho - \frac{1}{d}I_d\|_F^2\right)\mathbb{1}\{\rho \in S_{\mathrm{supp}}\}, \qquad S_{\mathrm{supp}} = \left\{\rho \in U : \|\rho - \frac{1}{d}I_d\| \leqslant \frac{4\sigma}{d}\right\}.$$

where Z is a normalizing constant. Further define a set of "good" states

$$S_{\mathsf{good}} = \left\{\rho \in U : \|\rho - \frac{1}{d}I_d\| \leqslant \frac{3\sigma}{d}\right\},$$

which corresponds to the event $\|G\| \leqslant 3$. Due to Fact 6.6.3, $\mu(S_{\mathsf{good}}) \geqslant 1 - e^{-\Omega(d)}$.

In the below proof, we will show that all $\rho_0 \in S_{\mathsf{good}}$ are hard to learn. The important property of S_{good} is that it is far from the boundary of $\mathrm{supp}(\mu) = S_{\mathrm{supp}}$; this ensures

that we can choose a suitable sub-sampling of μ in a neighborhood of ρ_0, which is rotationally symmetric around ρ_0.

Finally, we record the following straightforward fact.

Lemma 6.6.4. *For all $\rho, \rho' \in S_{\text{supp}}$, $\exp(-4d^2) \leqslant \mu(\rho)/\mu(\rho') \leqslant \exp(4d^2)$.*

Proof. For all $\rho \in S_{\text{supp}}$,

$$0 \leqslant \frac{d^3}{4\sigma^2}\|\rho - \frac{1}{d}I_d\|_F^2 \leqslant \frac{d^4}{4\sigma^2}\|\rho - \frac{1}{d}I_d\|^2 \leqslant 4d^2. \qquad \square$$

6.6.3 Anticoncentration of Posterior Distribution

Fix a tomography algorithm $(\mathcal{T}, \mathcal{A})$ as in Definition 6.6.1, and let $\mathcal{T}_\rho$ denote the distribution over observation sequences $\boldsymbol{x} = (x_1, \ldots, x_m)$ when $\mathcal{T}$ is run on state ρ. Note that for any states ρ, ρ' in the support of μ, the likelihood ratio

$$\frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho'}}(\boldsymbol{x}) = \prod_{i=1}^{m} \frac{x_i^\dagger \rho^{\otimes t} x_i}{x_i^\dagger \rho'^{\otimes t} x_i}$$

is well defined, since μ is supported on full-rank matrices. Let $\nu_{\boldsymbol{x}}$ denote the posterior distribution of ρ given observations $\boldsymbol{x}$. The density ratio of any $\rho, \rho' \in S_{\text{supp}}$ under $\nu_{\boldsymbol{x}}$ is given by Bayes' rule, and equals

$$\frac{\nu_{\boldsymbol{x}}(\rho)}{\nu_{\boldsymbol{x}}(\rho')} = \frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho'}}(\boldsymbol{x}) \cdot \frac{\mu(\rho)}{\mu(\rho')}.$$

So, for an arbitrary reference state $\rho' \in S_{\text{supp}}$ (below we take $\rho' = \rho_0$, the unknown true state) the density of $\nu_{\boldsymbol{x}}$ is

$$\nu_{\boldsymbol{x}}(\rho) = \frac{1}{Z_{\boldsymbol{x}}} \frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho'}}(\boldsymbol{x})\mu(\rho), \qquad Z_{\boldsymbol{x}} = \int_U \frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho'}}(\boldsymbol{x})\mu(\rho)\ \mathrm{d}\mathsf{Leb}_U(\rho).$$

The main technical component of the proof is the following anti-concentration result for $\nu_{\boldsymbol{x}}$.

Definition 6.6.5. For $\rho \in U$, let $B(\rho, \varepsilon)$ denote the ball $\{\rho' \in U : \|\rho' - \rho\|_1 \leqslant \varepsilon\}$. Similarly for $\rho \in U_0$, let $B(\rho, \varepsilon) = \{\rho' \in U_0 : \|\rho' - \rho\|_1 \leqslant \varepsilon\}$.

Theorem 6.6.6. *Suppose $d \gg 1$, $\varepsilon \leqslant \varepsilon_0$ for an absolute constant ε_0 and t satisfies $t \leqslant 1/(C\varepsilon^{0.1})$ for some absolute constant C. Also assume $m \ll d^3/(t^{1.5}\varepsilon^2)$. If $\rho_0 \in S_{\mathsf{good}}$ and $\boldsymbol{x} \sim \mathcal{T}_{\rho_0}$, there is an event $S_{\rho_0} \in \sigma(\boldsymbol{x})$, with $\mathbb{P}(\boldsymbol{x} \in S_{\rho_0}) \geqslant 1 - o(1)$, on which $\nu_{\boldsymbol{x}}(B(\rho_0, \varepsilon)) \ll 1$.*

Let C be a large constant we will set later. The starting point of the proof of Theorem 6.6.6 is the estimate

$$\frac{\nu_{\boldsymbol{x}}(B(\rho_0, C\varepsilon))}{\nu_{\boldsymbol{x}}(B(\rho_0, \varepsilon))} = \frac{\int_{B(\rho_0, C\varepsilon)} \frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x})\mu(\rho)\ \mathrm{d}\mathsf{Leb}_U(\rho)}{\int_{B(\rho_0, \varepsilon)} \frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x})\mu(\rho)\ \mathrm{d}\mathsf{Leb}_U(\rho)} \tag{6.18}$$

$$\geqslant \exp(-4d^2)\frac{\int_{B(\rho_0, C\varepsilon)} \frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x})\mathbb{1}\{\rho \in S_{\mathrm{supp}}\}\ \mathrm{d}\mathsf{Leb}_U(\rho)}{\int_{B(\rho_0, \varepsilon)} \frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x})\mathbb{1}\{\rho \in S_{\mathrm{supp}}\}\ \mathrm{d}\mathsf{Leb}_U(\rho)}, \tag{6.19}$$

where the second line uses Lemma 6.6.4. Applying Lemma 6.6.4 in this way amounts to replacing μ in the numerator of (6.18) with a measure that sub-samples it, and in the denominator with a measure that upper bounds it. Define the volumes

$$V_1 = \int_{B(0,1)} \mathrm{d}\mathsf{Leb}_{U_0}(\rho), \qquad V_2 = \int_{B(0,1)} \mathbb{1}\left\{\|\rho\| \leqslant 1/d\right\}\ \mathrm{d}\mathsf{Leb}_{U_0}(\rho).$$

We now separately bound the numerator and denominator of (6.19) in terms of these volumes, beginning with the denominator.

Lemma 6.6.7. *If $\rho_0 \in S_{\mathsf{good}}$, there is an event $S_{\rho_0} \in \sigma(\boldsymbol{x})$, with $\mathbf{Pr}(\boldsymbol{x} \in S_{\rho_0}) \geqslant 1 - \exp(-d^2)$, on which*

$$\int_{B(\rho_0, \varepsilon)} \frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x})\mathbb{1}\{\rho \in S_{\mathrm{supp}}\}\ \mathrm{d}\mathsf{Leb}_U(\rho) \leqslant \exp(d^2)\varepsilon^{d^2-1}V_1\,.$$

Proof. Note that $\mathbb{E}_{\boldsymbol{x} \sim \mathcal{T}_{\rho_0}} \frac{\mathrm{d}T_\rho}{\mathrm{d}T_{\rho_0}}(\boldsymbol{x}) = 1$. So

$$\mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{T}_{\rho_0}} \int_{B(\rho_0, \varepsilon)} \frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x})\mathbb{1}\{\rho \in S_{\mathrm{supp}}\}\ \mathrm{d}\mathsf{Leb}_U(\rho) = \int_{B(\rho_0, \varepsilon)} \mathbb{1}\{\rho \in S_{\mathrm{supp}}\}\ \mathrm{d}\mathsf{Leb}_U(\rho)$$

$$\leqslant \int_{B(\rho_0,\varepsilon)} \mathrm{d}\mathsf{Leb}_U(\rho) = \varepsilon^{d^2-1} V_1.$$

The exponent d^2-1 comes from the fact that the space of complex Hermitian matrices has dimension d^2, so U has dimension d^2-1. The result follows from Markov's inequality. □

Before bounding the numerator of (6.19), we define the set

$$N_*(\rho_0) = \left\{\rho \in U : \|\rho - \rho_0\| \leqslant \frac{C\varepsilon}{d}, \|\rho - \rho_0\|_1 \leqslant C\varepsilon\right\},$$

which is a rotationally symmetric neighborhood of ρ_0. Let $\gamma(\rho_0)$ denote the uniform distribution on $N_*(\rho_0)$ (w.r.t. Leb_U). That is, for bounded measurable test function $f : U \to \mathbb{R}$,

$$\mathop{\mathbb{E}}_{\rho\sim\gamma(\rho_0)} f(\rho) = \frac{\int_U f(\rho)\mathbb{1}\{\rho \in N_*(\rho_0)\}\mathrm{d}\mathsf{Leb}_U(\rho)}{\int_U \mathbb{1}\{\rho \in N_*(\rho_0)\}\mathrm{d}\mathsf{Leb}_U(\rho)}.$$

Lemma 6.6.8. *If $\rho_0 \in S_{\mathsf{good}}$ and $\varepsilon \leqslant \varepsilon_0$ for an absolute constant ε_0, then*

$$\int_{B(\rho_0,C\varepsilon)} \frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x})\mathbb{1}\{\rho \in S_{\mathrm{supp}}\}\ \mathrm{d}\mathsf{Leb}_U(\rho) \geqslant (C\varepsilon)^{d^2-1} \mathop{\mathbb{E}}_{\rho\sim\gamma(\rho_0)} \left[\frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x})\right] V_2. \tag{6.20}$$

Proof. Because $\rho_0 \in S_{\mathsf{good}}$, we have $\|\rho_0 - \frac{1}{d}I_d\| \leqslant \frac{3\sigma}{d}$. Because $\varepsilon \leqslant \sigma/C$, if ρ satisfies $\|\rho - \rho_0\| \leqslant \frac{C\varepsilon}{d}$ we have the implication chain

$$\|\rho - \rho_0\| \leqslant \frac{C\varepsilon}{d} \Rightarrow \|\rho - \rho_0\| \leqslant \frac{\sigma}{d} \Rightarrow \|\rho - \frac{1}{d}I_d\| \leqslant \frac{4\sigma}{d} \Rightarrow \rho \in S_{\mathrm{supp}}.$$

So, letting X denote the left-hand side of (6.20), we have

$$X \geqslant \int_{B(\rho_0,C\varepsilon)} \frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x})\mathbb{1}\left\{\|\rho - \rho_0\| \leqslant \frac{C\varepsilon}{d}\right\}\ \mathrm{d}\mathsf{Leb}_U(\rho) = \int_U \frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x})\mathbb{1}\{\rho \in N_*(\rho_0)\}\ \mathrm{d}\mathsf{Leb}_U(\rho)$$
$$= \mathop{\mathbb{E}}_{\rho\sim\gamma(\rho_0)} \left[\frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x})\right] \int_U \mathbb{1}\{\rho \in N_*(\rho_0)\}\ \mathrm{d}\mathsf{Leb}_U(\rho).$$

Finally, note that

$$\int_U \mathbb{1}\{\rho \in N_*(\rho_0)\}\, \mathrm{d}\mathsf{Leb}_U(\rho) = \int_{B(\rho_0, C\varepsilon)} \mathbb{1}\left\{\|\rho - \rho_0\| \leqslant \frac{C\varepsilon}{d}\right\}\, \mathrm{d}\mathsf{Leb}_U(\rho) = (C\varepsilon)^{d^2-1} V_2,$$

which conludes the proof. □

It remains to control $\mathbb{E}_{\rho\sim\gamma(\rho_0)}\left[\frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x})\right]$ and the volume ratio V_2/V_1. We focus on the former first, the latter can be bounded using formulas from random matrix theory as in [62]. We will need to condition on a "good" event for the observations $\boldsymbol{x}$. Note that this complication does not show up in [62] but shows up here due to the more complicated nature of the likelihood ratio expressions with t-entangled measurements. The event that we condition on is defined below.

Definition 6.6.9. We say that a sequence of vectors $\boldsymbol{x} = (x_1, \ldots, x_m)$ where $x_j \in \mathbb{C}^{d^t}$ is well-balanced if

- $\sum_{\lambda \vdash t} \sum_{j=1}^m \frac{\|\Pi_\lambda x_j\|^2}{d^t x_j^\dagger \rho_0^{\otimes t} x_j}(\lambda_1^2 + \cdots + \lambda_d^2) \leqslant 10(mt^{1.5} + \log(d/\varepsilon)\sqrt{m}t^2)$
- $\|-\binom{t}{2} m I_{d^2} + \frac{1}{d^{t-2}} \sum_{j=1}^m \frac{G_2(x_j)}{x_j^\dagger \rho_0^{\otimes} x_j}\| \leqslant 10t^2 d\sqrt{m}\log(d/\varepsilon)$

We first prove that the set of observations $\boldsymbol{x}$ is indeed well-balanced with high probability.

Lemma 6.6.10. *Let ε, d be parameters such that $\varepsilon \leqslant \varepsilon_0$ for some sufficiently small absolute constant ε_0. Let t, m be parameters such that $t \leqslant \frac{1}{10C\varepsilon}$ where C is some constant. Let $\rho_0 \in \mathbb{C}^{d\times d}$ be a state with $\|\rho_0 - I_d/d\| \leqslant \frac{4C\varepsilon}{d}$. Measure the state $\rho_0^{\otimes t}$ sequentially m times with arbitrary rank-1 POVMs (possibly chosen adaptively) and let the outcomes be $\boldsymbol{x} = (x_1, \ldots, x_m)$ where x_j is a unit vector in $\mathbb{C}^{d^t}$. Then with probability $1 - (\varepsilon/d)^{10}$, the collection $\boldsymbol{x}$ is well-balanced.*

Proof. For the first claim, note that for any POVM $\mathcal{M} = \{\omega_x d^t x x^\dagger\}_x$, if we measure

$\rho_0^{\otimes t}$ with this POVM, then

$$\mathop{\mathbb{E}}_{\mathcal{M}}\left[\frac{xx^\dagger}{d^t x^\dagger \rho_0^{\otimes t} x}\right] = \sum_{x\in\mathcal{M}} \omega_x xx^\dagger = \frac{I_{d^t}}{d^t} \tag{6.21}$$

where the expectation is over the outcome from $\mathcal{M}$. Thus,

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{M}}\left[\sum_{\lambda\vdash t} \frac{\|\Pi_\lambda x\|^2}{d^t x^\dagger \rho_0^{\otimes t} x}(\lambda_1^2+\cdots+\lambda_d^2)\right] &= \sum_{\lambda\vdash t} \frac{\dim(\lambda)\dim(V_\lambda^d)}{d^t}(\lambda_1^2+\cdots+\lambda_d^2)\\ &= \mathop{\mathbb{E}}_{\lambda\sim SW_d^t}[\lambda_1^2+\cdots+\lambda_d^2]\\ &\leqslant 3t^{1.5}\end{aligned}$$

where the last step uses Claim 6.2.28. Also note that the quantity inside the expectation is at most $2t^2$ always since ρ_0 has all eigenvalues at least $\frac{1}{2d^t}$ by the assumption on t. Thus, since the above holds for any POVM, we can apply Azuma's inequality and get that with probability at least $1-(\varepsilon/d)^{20}$,

$$\sum_{\lambda\vdash t}\sum_{j=1}^m \frac{\|\Pi_\lambda x_j\|^2}{d^t x_j^\dagger \rho_0^{\otimes t} x_j}(\lambda_1^2+\cdots+\lambda_d^2) \leqslant 10(mt^{1.5}+\log(d/\varepsilon)\sqrt{m}t^2)$$

and thus the first condition of well-balanced is satisfied. Now we analyze the second condition. By (6.21), we also have

$$\mathop{\mathbb{E}}_{\mathcal{M}}\left[\frac{1}{d^{t-2}}\frac{G_2(x)}{x^\dagger\rho_0^{\otimes t}x}\right] = \binom{t}{2} I_{d^2}\,.$$

Also, the quantity inside the expectation is PSD and has trace at most $2t^2d^2$ always. This means its variance has operator norm at most $2t^4d^2$. Since the above holds for any POVMs, we can apply matrix Azuma and get that with probability at least $1-(\varepsilon/d)^{20}$,

$$\|-\binom{t}{2} mI_{d^2} + \frac{1}{d^{t-2}}\sum_{j=1}^m \frac{G_2(x_j)}{x_j^\dagger \rho_0^{\otimes} x_j}\| \leqslant 10t^2 d\sqrt{m}\log(d/\varepsilon)$$

which matches the second condition for well-balanced, as desired. □

Now we can bound the average likelihood ratio whenever the sequence of observations is well-balanced.

Lemma 6.6.11. *Assume that $\varepsilon \leqslant \varepsilon_0$ and $t \leqslant 1/(C\varepsilon^{0.1})$ for some absolute constants C, ε_0 and $m \ll d^3/(t^{1.5}\varepsilon^2)$. If $\rho_0 \in S_{\mathsf{good}}$, then for any well-balanced sequence of unit vectors $\boldsymbol{x} = (x_1, \ldots, x_n)$,*

$$\mathop{\mathbb{E}}_{\rho \sim \gamma(\rho_0)} \left[\frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x}) \right] \geqslant \exp(-d^2).$$

Proof. Note that the distribution $\gamma(\rho_0)$ is rotationally symmetric around ρ. Thus, we can apply Lemma 6.5.3 with $\varepsilon \leftarrow C\varepsilon$ and

$$A = 10(mt^{1.5} + \log(d/\varepsilon)\sqrt{m}t^2)$$
$$B = 10t^2 d\sqrt{m}\log(d/\varepsilon)\,.$$

Plugging in the assumptions about the parameters m, t, we get

$$\mathop{\mathbb{E}}_{\rho \sim \gamma(\rho_0)} \left[\frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x}) \right] \geqslant \exp(-d^2)$$

as desired. □

The volume ratio V_2/V_1 is bounded by the following lemma from [62] [†].

Lemma 6.6.12 (See [62])**.** *We have that $V_2/V_1 \geqslant \exp(-10d^2)$.*

We can now combine the above lemmas together to prove Theorem 6.6.6.

Proof of Theorem 6.6.6. Let S_{ρ_0} be the intersection of the event from Lemma 6.6.7 and that the observations $\boldsymbol{x}$ are well-balanced. By Lemma 6.6.10, this event occurs with $1 - o(1)$ probability. Next, by the calculation (6.19) and Lemmas 6.6.7 and 6.6.8,

[†]Technically [62] works with real symmetric matrices instead of Hermitian matrices but the proofs are exactly the same.

for all $\boldsymbol{x} \in S_{\rho_0}$,

$$\frac{\nu_{\boldsymbol{x}}(B(\rho_0, C\varepsilon))}{\nu_{\boldsymbol{x}}(B(\rho_0, \varepsilon))} \geqslant \exp(-5d^2) C^{d^2-1} \mathop{\mathbb{E}}_{\rho \sim \gamma(\rho_0)} \left[\frac{\mathrm{d}\mathcal{T}_\rho}{\mathrm{d}\mathcal{T}_{\rho_0}}(\boldsymbol{x}) \right] \cdot \frac{V_2}{V_1}.$$

Lemmas 6.6.11 and 6.6.12 bound the remaining factors, giving

$$\frac{\nu_{\boldsymbol{x}}(B(\rho_0, C\varepsilon))}{\nu_{\boldsymbol{x}}(B(\rho_0, \varepsilon))} \geqslant \exp(-20d^2) C^{d^2-1}.$$

Taking $C = e^{30}$ gives $\frac{\nu_{\boldsymbol{x}}(B(\rho_0, C\varepsilon))}{\nu_{\boldsymbol{x}}(B(\rho_0, \varepsilon))} \gg 1$. Since $\nu_{\boldsymbol{x}}(B(\rho_0, C\varepsilon)) \leqslant 1$, this implies $\nu_{\boldsymbol{x}}(B(\rho_0, \varepsilon)) \ll 1$. □

6.6.4 Completing the Lower Bound

We can now complete the proof of our full lower bound for tomography with t-entangled measurements, which we state formally below:

Theorem 6.6.13. *There exist absolute constants $\varepsilon_0 > 0$ and $d_0 \in \mathbb{N}$ such that for any $0 < \varepsilon < \varepsilon_0$ and any integer $d \geqslant d_0$ and parameter $t \leqslant 1/\varepsilon^{0.1}$, the following holds. If $n = o(d^3/(\sqrt{t}\varepsilon^2))$, then for any algorithm for state tomography $(\mathcal{T}, \mathcal{A})$ that uses n total copies of ρ in (at most) t-entangled batches, its output $\widehat{\rho}$ after making all of its measurements satisfies $\|\rho - \widehat{\rho}\|_1 > \varepsilon$ with probability $1 - o(1)$.*

Proof. First, note that we can reduce to the case where the algorithm always makes exactly t-entangled measurements because for any algorithm that makes measurements with lesser entanglement in sequence, we can simply combine them into one larger entangled measurement (this comes at the cost of wasting up to half of the copies of ρ when we make all of the measurements exactly t-entangled).

Let $S \in \sigma(\rho, \boldsymbol{x})$ be the event that $\rho \sim \mu$ lies in S_{good} and $\boldsymbol{x} \sim \mathcal{T}_\rho$ lies in S_ρ. In this proof we will abuse notation and use $\mathcal{A}$ to also denote the internal randomness used

by $\mathcal{A}$. It suffices to show

$$\mathbf{Pr}_{\mathcal{A},\rho\sim\mu,\boldsymbol{x}\sim\mathcal{T}_\rho}[\|\mathcal{A}(\boldsymbol{x})-\rho\|_1 \leqslant \varepsilon] = o(1)\,.$$

First note that

$$\begin{aligned}\mathbf{Pr}_{\mathcal{A},\rho,\boldsymbol{x}}\left[\|\mathcal{A}(\boldsymbol{x})-\rho\|_1 \leqslant \varepsilon\right] &= \mathop{\mathbb{E}}_{\mathcal{A},\boldsymbol{x}} \mathop{\mathbb{E}}_{\rho\sim\nu_{\boldsymbol{x}}}\left[\mathbf{1}\|\mathcal{A}(\boldsymbol{x})-\rho\|_1 \leqslant \varepsilon\right] \\ &\leqslant \mathop{\mathbb{E}}_{\mathcal{A},\boldsymbol{x}} \mathop{\mathbb{E}}_{\rho\sim\nu_{\boldsymbol{x}}}\left[\mathbf{1}\|\mathcal{A}(\boldsymbol{x})-\rho\|_1 \leqslant \varepsilon \text{ and } (\rho,\boldsymbol{x})\in S\right] + o(1)\end{aligned} \tag{6.22}$$

where the second step follows by a union bound and the fact that $\mathbf{Pr}[(\rho,\boldsymbol{x})\notin S] = o(1)$ by Theorem 6.6.6.

For any choice of internal randomness for $\mathcal{A}$ and any transcript $\boldsymbol{x}$, let $\rho_{\boldsymbol{x}}^{\mathcal{A}}$ denote an arbitrary state for which $(\rho_{\boldsymbol{x}}^{\mathcal{A}},\boldsymbol{x})\in S$ and $\|\mathcal{A}(\boldsymbol{x})-\rho_{\boldsymbol{x}}^{\mathcal{A}}\|_1 \leqslant \varepsilon$, if such a state exists. Denote by $\mathcal{E}$ the event that such a state exists. Then under $\mathcal{E}$, for any state ρ for which $\|\mathcal{A}(\boldsymbol{x})-\rho\|_1 \leqslant \varepsilon$, we have $\|\rho_{\boldsymbol{x}}^{\mathcal{A}}-\rho\|_1 \leqslant 2\varepsilon$. If $\mathcal{E}$ does not occur for some choice of internal randomness for $\mathcal{A}$ and some $\boldsymbol{x}$, note that the corresponding inner expectation in (6.22) is zero. We can thus upper bound the double expectation in (6.22) by

$$\mathop{\mathbb{E}}_{\mathcal{A},\boldsymbol{x}|\mathcal{E}} \mathop{\mathbb{E}}_{\rho\sim\nu_{\boldsymbol{x}}}\left[\mathbf{1}\|\rho_{\boldsymbol{x}}^{\mathcal{A}}-\rho\|_1 \leqslant 2\varepsilon \text{ and } (\rho,\boldsymbol{x})\in S\right] \leqslant \mathop{\mathbb{E}}_{\mathcal{A},\boldsymbol{x}|\mathcal{E}} \mathbf{Pr}_{\rho\sim\nu_{\boldsymbol{x}}}\left[\|\rho_{\boldsymbol{x}}^{\mathcal{A}}-\rho\|_1 \leqslant 2\varepsilon\right] = o(1),$$

where in the last step we used the fact that under $\mathcal{E}$ we have $(\rho_{\boldsymbol{x}}^{\mathcal{A}},\boldsymbol{x})\in S$, so by Theorem 6.6.6 the posterior measure $\nu_{\boldsymbol{x}}$ places $o(1)$ mass on the trace norm ε-ball around $\rho_{\boldsymbol{x}}^{\mathcal{A}}$. □

Appendix A

Omitted Proofs from Chapter 2

A.1 Cluster Expansion

We give bounds on the Taylor series expansion of the log-partition function of a local Hamiltonian, closely following the cluster expansion argument of Mann and Helmuth [160, 141]. We reproduce this for clarity, increased generality, and completeness, when merged this with results on computing the cluster expansion [103]. We work with the dual interaction graph, as opposed to the interaction hypergraph as in [160], to give slightly tighter bounds. The arguments here straightforwardly extend to qudit systems, but for the sake of simplicity we present only the qubit setting.

Notation. For this section, we will be working with multisets. For a base set S, a multiset of S, $\boldsymbol{T}\colon S \to \mathbb{N}$, is defined in terms of its multiplicity function. The support of $\boldsymbol{T}$ is the set of $s \in S$ with $\boldsymbol{T}(s) > 0$, and the size of $\boldsymbol{T}$ is denoted $|\boldsymbol{T}| = \sum_{s\in S} \boldsymbol{T}(s)$. We can take unions of multisets, $\boldsymbol{T}_1 \cup \boldsymbol{T}_2$, in the usual way, where its multiplicity function is the sum of the two individual multiplicities. We use the factorial notation $\boldsymbol{T}! = \prod_{s\in S}(\boldsymbol{T}(s)!)$. For a collection of complex numbers indexed by S, $\{x_s\}_{s\in S}$, $x^{\boldsymbol{T}} = \prod_{s\in S} x_s^{\boldsymbol{T}(s)}$ is the product of variables associated with $\boldsymbol{T}$ with multiplicity.

We use the Iverson bracket: for a proposition P, $[\![P]\!]$ equals one when P is true,

and zero otherwise. We denote the normalized trace by $\overline{\mathrm{tr}}(A) = \mathrm{tr}(A)/\mathrm{tr}(I)$. The symmetric group with order $\mathfrak{n}!$ is denoted $\mathrm{S}_{\mathfrak{n}}$.

A.1.1 Abstract Polymer Model

We first describe cluster expansion for abstract polymer models, along with the Kotecký–Preiss condition for convergence of the expansion [141]. This is fully standard; we follow the formulation of Friedli and Velenik [88], which in turn follows the analysis of Ueltschi [203].

Definition A.1.1 (Abstract polymer model)**.** An *abstract polymer model* is a collection of objects K which we refer to as *polymers*, along with associated weights $\{w_\gamma\}_{\gamma\in K}$ in $\mathbb{C}$ and a symmetric relation $\sim$ on K such that, for all $\gamma \in K$, $\gamma \nsim \gamma$. Here, $\sim$ denotes compatibility: for example, $\gamma \nsim \gamma$ means that a polymer is incompatible with itself.

For an ordered list of polymers $\Gamma = (\gamma_1, \ldots, \gamma_\ell)$, its *incompatibility graph* H_Γ has ℓ vertices indexed by Γ and an edge between γ_i and γ_j if and only if $\gamma_i \nsim \gamma_j$. Γ is a *cluster* if its incompatibility graph is connected.*

Definition A.1.2 (Partition function of a polymer model, [88, Definition 5.2])**.** A polymer model has an associated *partition function*.

$$Z = \sum_{S\subseteq K} \Big(\prod_{\gamma\in S} w_\gamma\Big) [\![S \text{ is pairwise compatible}]\!].$$

We say that a subset $S \subseteq K$ is pairwise compatible if, for all distinct $\gamma, \gamma' \in S$, $\gamma \sim \gamma'$.

Definition A.1.3 (Ursell function)**.** The *Ursell function* φ is a function mapping graphs to real numbers. For $H = (V, E)$,

$$\varphi(H) = \frac{1}{|V|!} \sum_{\substack{A\subseteq E \\ \text{spanning} \\ \text{connected}}} (-1)^{|A|}$$

*The definition of a cluster is different from [103], but this is the definition that is consistent with the rest of the cluster expansion literature.

Proposition A.1.4 (Formal expansion for the log-partition function, [88, Proposition 5.3]). *The following formal equality holds.*

$$\log Z = \sum_{k=1}^{\infty} \sum_{(\gamma_1,\dots,\gamma_k)\in K^k} \varphi(H_{(\gamma_1,\dots,\gamma_k)}) \prod_{i\in[k]} w_{\gamma_i}$$

Theorem A.1.5 (Kotecký–Preiss convergence condition, [88, Theorem 5.4]). *Suppose that there exists a function $\xi\colon K \to \mathbb{R}_{>0}$ and weights $\{\widehat{w}_\gamma\}_{\gamma\in K}$ such that, for each $\gamma_* \in K$,*

$$\sum_{\gamma\in K} |\widehat{w}_\gamma| e^{\xi(\gamma)} [\![\gamma \nsim \gamma_*]\!] \leqslant \xi(\gamma_*). \tag{A.1}$$

Then, for all $\gamma_1 \in K$,

$$1 + \sum_{k\geqslant 2} k \sum_{\gamma_2,\dots,\gamma_k} |\varphi(H_{\gamma_1,\dots,\gamma_k})| \prod_{j=2}^{k} |\widehat{w}_{\gamma_j}| \leqslant e^{\xi(\gamma_1)}. \tag{A.2}$$

If $\widehat{w} = w$, then this implies convergence of the cluster expansion series (provided the set of polymers is finite). We will take $\widehat{w}_\gamma = w_\gamma e^{b(\gamma)}$ for some function b to get a stronger bound on the convergence of the series.

A.1.2 Quantum Spin System as an Abstract Polymer Model

This section follows the work of Mann and Helmuth [160], which in turn follows Netočný and Redig [167].

Lemma A.1.6 (Following [160, Lemma 2]). *Let $H = \sum_{a=1}^{m} \lambda_a E_a$ be a low-intersection Hamiltonian with dual interaction graph $\mathfrak{G}$ (see Definition 1.4.14). Then $\overline{\mathrm{tr}} \exp(-\beta H)$ equals the partition function of the polymer model where K is the space of non-empty multisets of $[m]$ whose support is connected on $\mathfrak{G}$; $\gamma \sim \gamma'$ if and only if their distance*

on $\mathfrak{G}$ is at least two; and the weight function is, for a polymer γ with $|\gamma|$ elements,

$$w_\gamma = \frac{(-\beta)^{|\gamma|}}{|\gamma|!\gamma!}\,\overline{\mathrm{tr}}\Big(\sum_{\sigma\in \mathrm{S}_{|\gamma|}}\prod_{i=1}^{|\gamma|} E_{\gamma_{\sigma(i)}}\Big)\lambda^\gamma$$

Combining this with the formal equality in Proposition A.1.4, we deduce the following expression for the log-partition function of the Hamiltonian,

$$\begin{aligned}
&\log \overline{\mathrm{tr}}\exp(-\beta H)\\
&= \sum_{k=1}^{\infty}\sum_{(\gamma_1,\dots,\gamma_k)\in K^k} \varphi(H_{(\gamma_1,\dots,\gamma_k)})\prod_{i\in[k]} w_{\gamma_i}\\
&= \sum_{k=1}^{\infty}\sum_{(\gamma_1,\dots,\gamma_k)\in K^k} \varphi(H_{(\gamma_1,\dots,\gamma_k)})\prod_{i\in[k]}\Big(\frac{(-\beta)^{|\gamma_i|}}{|\gamma_i|!\gamma_i!}\,\overline{\mathrm{tr}}\Big(\sum_{\sigma\in \mathrm{S}_{|\gamma_i|}}\prod_{j=1}^{|\gamma_i|} E_{(\gamma_i)_{\sigma(j)}}\Big)\lambda^{\gamma_i}\Big)\\
&= \sum_{k=1}^{\infty}\sum_{(\gamma_1,\dots,\gamma_k)\in K^k} (-\beta\lambda)^{\gamma_1\cup\dots\cup\gamma_k}\varphi(H_{(\gamma_1,\dots,\gamma_k)})\prod_{i\in[k]}\Big(\frac{1}{|\gamma_i|!\gamma_i!}\,\overline{\mathrm{tr}}\Big(\sum_{\sigma\in \mathrm{S}_{|\gamma_i|}}\prod_{j=1}^{|\gamma_i|} E_{(\gamma_i)_{\sigma(j)}}\Big)\Big).
\end{aligned}$$

This cluster expansion series is closely related to the multivariate Taylor series expansion of $\log \overline{\mathrm{tr}}\exp(-\beta H)$: if we group together all the terms with the same monomial λ^γ, we get the Taylor series expansion around $\lambda = (0,\dots,0)$.

Proof sketch. We can write

$$\begin{aligned}
\overline{\mathrm{tr}}\exp(-\beta H) &= \sum_{k\geqslant 0}\frac{1}{k!}\,\overline{\mathrm{tr}}((-\beta H)^k)\\
&= \sum_{k\geqslant 0}\frac{(-\beta)^k}{k!}\sum_{a_1,\dots,a_k}\overline{\mathrm{tr}}(E_{a_1}\dots E_{a_k})\lambda_{a_1}\dots\lambda_{a_k}.
\end{aligned}$$

We want to show that this is equal to the expression for the partition function from Definition A.1.2:

$$\sum_{S\subseteq K}[\![S \text{ is pairwise compatible}]\!]\prod_{\gamma\in S}\frac{(-\beta)^{|\gamma|}}{|\gamma|!\gamma!}\,\overline{\mathrm{tr}}\Big(\sum_{\sigma\in \mathrm{S}_{|\gamma|}}\prod_{i=1}^{|\gamma|} E_{\gamma_{\sigma(i)}}\Big)\lambda^\gamma.$$

Since $\overline{\mathrm{tr}}(XY) = \overline{\mathrm{tr}}(X)\,\overline{\mathrm{tr}}(Y)$ when X and Y have disjoint supports, we can split up $\overline{\mathrm{tr}}(E_{a_1} \dots E_{a_k})$ into a product of vertex-disjoint parts, each of which corresponds to a polymer γ. Because these parts are vertex-disjoint, their distance on $\mathfrak{G}$ is at least two, and so these polymers are compatible. There is still an underlying ordering $a_1, \dots, a_k$, but after grouping terms appropriately and counting, we can get a sum over "unordered" polymers which take the desired form. □

Quantum many-body systems satisfy the Kotecký–Preiss condition at high temperature.†

Fact A.1.7. *By [103, Eqs. 29, 36], there are at most*

$$\begin{aligned}
&\sum_{v=1}^{w} \binom{v(\mathfrak{d}-1)+1}{v-1} \frac{\mathfrak{d}}{v(\mathfrak{d}-1)+1} \binom{w-1}{v-1} \\
&= \sum_{v=1}^{w} \binom{v(\mathfrak{d}-1)}{v-1} \frac{\mathfrak{d}}{v(\mathfrak{d}-2)+3} \binom{w-1}{v-1} \\
&\leqslant \sum_{v=1}^{w} \binom{v(\mathfrak{d}-1)}{v-1} \binom{w-1}{v-1} \\
&\leqslant e \sum_{v=1}^{w} (e(\mathfrak{d}-1))^{v-1} \binom{w-1}{v-1} \\
&= e(1+e(\mathfrak{d}-1))^{w-1}
\end{aligned}$$

clusters γ such that $a \in \gamma$ and $|\gamma| = w$.

Lemma A.1.8 ([160, Lemma 4])**.** *The polymer model in Lemma A.1.6 satisfies the condition in Theorem A.1.5 for $\widehat{w}_\gamma = w_\gamma (\frac{\beta_c}{\beta})^{|\gamma|}$ and $\xi(\gamma) = |\gamma|$, provided $\beta < \beta_c = \frac{1}{e(e+1)(1+e(\mathfrak{d}-1))}$.*

$$\sum_{\gamma \in K} |\widehat{w}_\gamma| e^{\xi(\gamma)} [\![\gamma \nsim \gamma_*]\!] \leqslant \xi(\gamma_*). \tag{A.3}$$

†We note here a small discrepancy in the proof of [160, Lemma 4]: there, they use that the number of vertices in a polymer γ is at most $|\gamma| + 1$, but this fact is only true for two-local Hamiltonians. The analysis presented here accounts for this, getting a critical temperature which is slightly weaker (by a factor of about $\mathfrak{K}$) than what is claimed in [160].

This means that, for all $\gamma_1 \in K$,

$$1 + \sum_{k \geqslant 2} k \sum_{\gamma_2, \ldots, \gamma_k} |\varphi(H_{\gamma_1, \ldots, \gamma_k})| \prod_{j=2}^{k} |w_{\gamma_j}| \big(\frac{\beta_c}{\beta}\big)^{|\gamma_j|} \leqslant e^{|\gamma_1|}. \tag{A.4}$$

Proof. First, we bound the weight w_γ of a polymer γ.

$$|w_\gamma| = \Big| \frac{(-\beta)^{|\gamma|}}{|\gamma|!\gamma!} \operatorname{tr}\Big(\sum_{\sigma \in S_{|\gamma|}} \prod_{i=1}^{|\gamma|} E_{\gamma_{\sigma(i)}} \Big) \lambda^\gamma \Big| \leqslant \frac{\beta^{|\gamma|}}{|\gamma|!\gamma!} |\gamma|! |\lambda^\gamma| \leqslant \frac{\beta^{|\gamma|}}{\gamma!} \leqslant \beta^{|\gamma|} \tag{A.5}$$

This bound implies that $|\widehat{w}_\gamma| = |w_\gamma| (\frac{\beta_c}{\beta})^{|\gamma|} \leqslant \beta_c^{|\gamma|}$. Now, we bound the expression in the lemma statement. First consider a polymer $\gamma_{\{a\}}$ which corresponds to the multiset containing a single term $a \in [m]$.

$$\sum_{\gamma \in K} |\widehat{w}_\gamma| e^{\xi(\gamma)} [\![\gamma \nsim \gamma_{\{a\}}]\!] \leqslant \sum_{\gamma \in K} (e\beta_c)^{|\gamma|} [\![\gamma \nsim \gamma_{\{a\}}]\!]$$

The set of polymers which are incompatible with $\gamma_{\{a\}}$ are the set of connected multisets of $[m]$ which are adjacent to a. By Fact A.1.7, there are at most $e(1 + e(\mathfrak{d} - 1))^{w-1}$ polymers γ such that $a \in \gamma$ and $|\gamma| = w$, and this is equal to the number of polymers γ' with $|\gamma'| = w - 1$ and which are adjacent to a (provided $w > 1$). So,

$$\begin{aligned} \sum_{\gamma \in K} |\widehat{w}_\gamma| e^{\xi(\gamma)} [\![\gamma \nsim \gamma_{\{a\}}]\!] &\leqslant \sum_{w=2}^{\infty} (e\beta_c)^{w-1} e (1 + e(\mathfrak{d} - 1))^{w-1} \\ &= e \sum_{w=1}^{\infty} (e\beta_c (1 + e(\mathfrak{d} - 1)))^w \\ &= e \sum_{w=1}^{\infty} \frac{1}{(e+1)^w} = 1 \end{aligned}$$

where the last line uses that $\beta_c = \frac{1}{e(e+1)(1+e(\mathfrak{d}-1))}$. Consequently, if we take a general γ_*, then we get the desired bound:

$$\sum_{\gamma \in K} |\widehat{w}_\gamma| e^{\xi(\gamma)} [\![\gamma \nsim \gamma_*]\!] \leqslant \sum_{a \in \gamma_*} \sum_{\gamma \in K} |\widehat{w}_\gamma| e^{\xi(\gamma)} [\![\gamma \nsim \gamma_{\{a\}}]\!] \leqslant |\gamma_*|. \qquad \square$$

Corollary A.1.9. *When $\beta < \beta_c = \frac{1}{e(e+1)(1+e(\mathfrak{d}-1))}$, the cluster expansion of the polymer model in Lemma A.1.6 can be truncated. Then the error term from truncation at degree d is bounded,*

$$\Big|\sum_{\ell=d+1}^{\infty}\sum_{k=1}^{\infty}\sum_{\substack{(\gamma_1,\dots,\gamma_k)\in K^k\\ |\gamma_1|+\cdots+|\gamma_k|=\ell}}\varphi(H_{(\gamma_1,\dots,\gamma_k)})\prod_{i\in[k]}w_{\gamma_i}\Big| \leqslant \mathfrak{n}\frac{(\beta/\beta_c)^{d+1}}{1-\beta/\beta_c}.$$

Proof. This follows from both parts of Lemma A.1.8. First, let $T \subseteq [m]$ be a set of terms such that $|T| \leqslant \mathfrak{n}$ and every term is a neighbor of some element of T in $\mathfrak{G}$.

$$\begin{aligned}
&\Big|\sum_{\ell=d+1}^{\infty}\sum_{k=1}^{\infty}\sum_{\substack{(\gamma_1,\dots,\gamma_k)\in K^k\\ |\gamma_1|+\cdots+|\gamma_k|=\ell}}\varphi(H_{(\gamma_1,\dots,\gamma_k)})\prod_{i\in[k]}w_{\gamma_i}\Big| \\
&\leqslant \sum_{\ell=d+1}^{\infty}\sum_{k=1}^{\infty}\sum_{\substack{(\gamma_1,\dots,\gamma_k)\in K^k\\ |\gamma_1|+\cdots+|\gamma_k|=\ell}}|\varphi(H_{(\gamma_1,\dots,\gamma_k)})|\prod_{i\in[k]}|w_{\gamma_i}| \\
&= \sum_{\ell=d+1}^{\infty}\sum_{\substack{\gamma_1\in K\\ |\gamma_1|\leqslant\ell}}|w_{\gamma_1}|\big(\frac{\beta_c}{\beta}\big)^{-\ell+|\gamma_1|}\sum_{k=1}^{\infty}\sum_{\substack{(\gamma_2,\dots,\gamma_k)\in K^{k-1}\\ |\gamma_1|+\cdots+|\gamma_k|=\ell}}|\varphi(H_{(\gamma_1,\dots,\gamma_k)})|\prod_{i=2}^{k}|w_{\gamma_i}|\big(\frac{\beta_c}{\beta}\big)^{|\gamma_i|} \\
&\leqslant \sum_{\ell=d+1}^{\infty}\sum_{\substack{\gamma_1\in K\\ |\gamma_1|\leqslant\ell}}|w_{\gamma_1}|\big(\frac{\beta_c}{\beta}\big)^{-\ell+|\gamma_1|}e^{|\gamma_1|} && \text{by Eq. (A.4)} \\
&= \sum_{\ell=d+1}^{\infty}\big(\frac{\beta_c}{\beta}\big)^{-\ell}\sum_{\substack{\gamma_1\in K\\ |\gamma_1|\leqslant\ell}}|\widehat{w}_{\gamma_1}|e^{|\gamma_1|} \\
&\leqslant \sum_{\ell=d+1}^{\infty}\big(\frac{\beta_c}{\beta}\big)^{-\ell}\sum_{a\in T}\sum_{\substack{\gamma_1\in K\\ |\gamma_1|\leqslant\ell}}|\widehat{w}_{\gamma_1}|e^{|\gamma_1|}[\![\gamma_1 \nsim \gamma_{\{a\}}]\!] && \text{by property of } T \\
&\leqslant \sum_{\ell=d+1}^{\infty}\big(\frac{\beta_c}{\beta}\big)^{-\ell}|T| && \text{by Eq. (A.3)} \\
&= |T|\frac{(\beta/\beta_c)^{d+1}}{1-\beta/\beta_c} \leqslant \mathfrak{n}\frac{(\beta/\beta_c)^{d+1}}{1-\beta/\beta_c} && \text{by property of } T
\end{aligned}$$

□

A.1.3 Computation of the Cluster Expansion

Lemma A.1.10 (Version of [103, Theorem 3.1] for the log-partition function)**.** *Let $H = H(\lambda) = \sum_{a=1}^{m} \lambda_a E_a$ be a Hamiltonian on $\mathfrak{n}$ qubits with Pauli terms, locality $\mathfrak{K}$, and dual interaction graph $\mathfrak{G}$ with max degree $\mathfrak{d}$ (Definitions 1.4.13 and 1.4.14). Let $0 \leqslant \beta < \beta_c = 1/(e(e+1)(1+e(\mathfrak{d}-1)))$. Then the log-partition function of H can be expressed as a power series in β,*

$$\mathcal{L} \coloneqq \log(\mathrm{tr}(e^{-\beta H})) = \sum_{\ell \geqslant 0} \beta^{\ell} p_{\ell}(\lambda),$$

where p_ℓ is a degree-ℓ homogeneous polynomial in λ with the following properties:

1. *p_ℓ consists of at most $\mathfrak{n}(e\mathfrak{d})^{\ell}$ monomials.*

2. *The series decays as $|\sum_{\ell \geqslant d} \beta^{\ell} p_{\ell}(\lambda)| \leqslant \mathfrak{n} \frac{(\beta/\beta_c)^d}{1-\beta/\beta_c}$.*

Further, after $\mathcal{O}(\mathfrak{K} m \mathfrak{d} \log \mathfrak{d})$ pre-processing time, we have the following form of access to p_ℓ:

A. *The list of monomials that appear in p_ℓ can be enumerated in time $\mathcal{O}(\ell \mathfrak{d} \mu)$, where μ is the number of monomials.*

B. *The coefficient of any monomial in p_ℓ can be computed exactly in $\mathcal{O}(\mathfrak{K}\ell^3 + 8^{\ell}\ell^5 \log^2 \ell) = (8^{\ell} + \mathfrak{K}) \,\mathrm{poly}(\ell)$ time.*

Proof. Most of this directly follows from the results in Section 3 of [103], with property 2. following from Corollary A.1.9. We direct the reader to [103] for the proofs of those results. Here, we only identify where our statements come from.

First, we observe that the $\mathcal{L} = \log(\mathrm{tr}(e^{-\beta H}))$ has a formal multivariate Taylor series

expansion around $\lambda = (0, \ldots, 0)$ [103, Eqs. 24 and 25],

$$\mathcal{L} = \sum_{\ell \geqslant 0} \underbrace{\sum_{\boldsymbol{V} : |\boldsymbol{V}| = \ell} \frac{\lambda^{\boldsymbol{V}}}{\boldsymbol{V}!} \mathcal{D}_{\boldsymbol{V}} \mathcal{L}}_{\beta^\ell p_\ell(\lambda)},$$

where $\boldsymbol{V}$ denotes a multiset over terms $[m]$; $\lambda^{\boldsymbol{V}} = \prod_{a \in \boldsymbol{V}} \lambda^a$ is the product of all coefficients associated to the terms in $\boldsymbol{V}$ with multiplicity; and $\mathcal{D}_{\boldsymbol{V}} \mathcal{L} = \prod_{a \in \boldsymbol{V}} \frac{\partial}{\partial \lambda_a} \mathcal{L}|_{\lambda = (0, \ldots, 0)}$ is the log-partition function, with derivatives taken for every λ_a with $a \in \boldsymbol{V}$ with multiplicity, evaluated at $\lambda = (0, \ldots, 0)$. Note that $\mathcal{D}_{\boldsymbol{V}} \mathcal{L}$ is a constant in λ. This formal expression becomes a true equality whenever the right-hand side series converges.

The coefficient, $\mathcal{D}_{\boldsymbol{V}} \mathcal{L}$, is only non-zero when $\boldsymbol{V}$ is connected [103, Proposition 3.5]. Because of the degree bound $\mathfrak{d}$, the number of such "clusters" (connected $\boldsymbol{V}$) of size ℓ is merely exponential, bounded by $\mathfrak{n} e \mathfrak{d} (1 + e(\mathfrak{d} - 1))^{\ell - 1} \leqslant \mathfrak{n} (e \mathfrak{d})^\ell$ [103, Proposition 3.6]. This gives the monomial bound.

As discussed in the previous section, the cluster expansion of

$$\log(\overline{\mathrm{tr}}(e^{-\beta H})) = \log(\mathrm{tr}(e^{-\beta H})) - \log(2^{\mathfrak{n}})$$

is equal to its multivariate Taylor series expansion when terms associated to a particular monomial are collected together. In particular, we have that, in the notation from Lemma A.1.6,

$$\sum_{\ell \geqslant d+1} \beta^\ell p_\ell(\lambda) = \sum_{\ell \geqslant d+1} \sum_{k=1}^{\infty} \sum_{\substack{(\gamma_1, \ldots, \gamma_k) \in K^k \\ |\gamma_1| + \cdots + |\gamma_k| = \ell}} \varphi(H_{(\gamma_1, \ldots, \gamma_k)}) \prod_{i \in [k]} w_{\gamma_i}.$$

Thus, we can apply the bound in Corollary A.1.9 to get property 2.

As for the running time statements, enumerating monomials amounts to enumerating clusters, which is done in [103, Section 3.4]. Computing a coefficient amounts to computing $\mathcal{D}_{\boldsymbol{V}} \mathcal{L}$, which is done in [103, Proposition 3.13]. □

Appendix B

Omitted Proofs from Chapter 3

B.1 Proof of Theorem 3.3.7

We recall the setup from Theorem 3.3.7. We have H, a $\mathfrak{K}$-local Hamiltonian with dual interaction graph $\mathfrak{G}$ with max degree $\mathfrak{d}$, along with $A = \sum_b \sigma_b P_b$, a $\mathfrak{K}$'-local operator where P_b are products of Pauli matrices and $-1 \leqslant \sigma_b \leqslant 1$ and whose dual interaction graph has max degree $\mathfrak{d}'$. For $\beta > 0$, ρ is the corresponding Gibbs state of H. We wish to show a lower bound on $\langle A^2 \rangle$.

$$\langle A^2 \rangle = \operatorname{tr}(A^2 \rho) \geqslant \max_{i \in [\mathfrak{n}]} \Big(c \operatorname{tr}(A_{(i)}^2 / N) \Big)^{6 + c'\beta}.$$

Remark B.1.1. [15, Theorem 33] proves that

$$\langle A^2 \rangle = \operatorname{tr}(A^2 \rho) \geqslant \max_{i \in \Lambda} \operatorname{tr}(A_{(i)}^2 / N)^{\beta^{\Omega(1)}}$$

for *quasi-local* operators.

B.1.1 Reproving Lemmas

First, we show that this quantity is "protected" by local unitary operations, meaning it doesn't change too much when we apply a local unitary to A. How this works in AAKS is as follows. There are two claims: Claim 37 and Claim 38. Claim 37 only uses AKL for local operators, while Claim 38 uses AKL for quasi-local operators (and relies on Claim 37).

Lemma B.1.2 ([15, Claim 37])**.** *Let U be a unitary supported on $S \subset [\mathfrak{n}]$. Then, for any M with $\|M\| \leqslant 1$,*

$$\langle (U^\dagger M U)^2 \rangle \leqslant \widetilde{\mathcal{O}}\Big((\mathfrak{d}+1)\mathfrak{K}|S| \big(4e^{\frac{2|S|}{\mathfrak{K}}}\big)^{\frac{2(\mathfrak{d}+1)\mathfrak{K}\beta}{1+2(\mathfrak{d}+1)\mathfrak{K}\beta}} \langle M^2 \rangle^{\frac{1}{1+2(\mathfrak{d}+1)\mathfrak{K}\beta}} \Big).$$

Thinking of $\mathfrak{d}$ and $\mathfrak{K}$ as being constants and choosing c, c' appropriately, the inequality becomes

$$\langle (U^\dagger M U)^2 \rangle \leqslant e^{c|S|} \langle M^2 \rangle^{\frac{1}{1+c'\beta}}.$$

The $\widetilde{\mathcal{O}}(\cdots)$ gets folded into the values of c, c'.

Proof. The overview of the proof is as follows. We want to bound the influence of U on the expression $\mathrm{tr}(UM^2U\rho)$, so we can consider it in terms of the eigenbasis of H. The off-diagonal pieces are bounded by Corollary 3.3.10 (applying it with $g = \mathfrak{d}+1$ and $R = (\mathfrak{d}+1)|S|$):

$$\|\Pi^{(H)}_{[\sigma+\Delta,\infty]} U \Pi^{(H)}_{[-\infty,\sigma]}\| \leqslant \|U\| e^{-\frac{1}{4(\mathfrak{d}+1)\mathfrak{K}}(\Delta - 4(\mathfrak{d}+1)|S|)} = e^{-\frac{\Delta}{4(\mathfrak{d}+1)\mathfrak{K}}} e^{\frac{|S|}{\mathfrak{K}}}. \tag{B.1}$$

We split our expression into the on-diagonal and off-diagonal pieces.

$\langle (U^\dagger M U)^2 \rangle$

$= \mathrm{tr}((U^\dagger M U)^2 \rho)$

$$
\begin{aligned}
&= \sum_i \operatorname{tr}((U^\dagger M U)^2 \Pi^{(H)}_{[i,i+1)} \rho) \\
&= \sum_i \|U^\dagger M U \Pi^{(H)}_{[i,i+1)} \sqrt{\rho}\|_F^2 \\
&= \sum_i \|M U \Pi^{(H)}_{[i,i+1)} \sqrt{\rho}\|_F^2 \\
&= \sum_i \|M(\Pi^{(H)}_{(-\infty,i-\Delta)} + \Pi^{(H)}_{[i-\Delta,i+1+\Delta)} + \Pi^{(H)}_{[i+1+\Delta,\infty)}) U \Pi^{(H)}_{[i,i+1)} \sqrt{\rho}\|_F^2 \\
&\leqslant 2 \sum_i \Big(\underbrace{\|M(\Pi^{(H)}_{(-\infty,i-\Delta)} + \Pi^{(H)}_{[i+1+\Delta,\infty)}) U \Pi^{(H)}_{[i,i+1)} \sqrt{\rho}\|_F^2}_{(\mathrm{OFF}_i)} + \underbrace{\|M \Pi^{(H)}_{[i-\Delta,i+1+\Delta)} U \Pi^{(H)}_{[i,i+1)} \sqrt{\rho}\|_F^2}_{(\mathrm{ON}_i)} \Big)
\end{aligned}
$$

We first bound (ON_i).

$$
\begin{aligned}
(\mathrm{ON}_i) &= \|M \Pi^{(H)}_{[i-\Delta,i+1+\Delta)} U \Pi^{(H)}_{[i,i+1)} \sqrt{\rho}\|_F^2 \\
&= \|M \Pi^{(H)}_{[i-\Delta,i+1+\Delta)}\|_F^2 \|U \Pi^{(H)}_{[i,i+1)} \sqrt{\rho}\|^2 \\
&\leqslant \|M \Pi^{(H)}_{[i-\Delta,i+1+\Delta)}\|_F^2 e^{-\beta i} \\
&\leqslant \|M \Pi^{(H)}_{[i-\Delta,i+1+\Delta)} \sqrt{\rho}\|_F^2 e^{\beta \Delta}
\end{aligned}
$$

Next, we bound (OFF_i). Here, we use that $\|M\| \leqslant 1$ and Eq. (B.1).

$$
\begin{aligned}
(\mathrm{OFF}_i) &= \|M(\Pi^{(H)}_{(-\infty,i-\Delta)} + \Pi^{(H)}_{[i+1+\Delta,\infty)}) U \Pi^{(H)}_{[i,i+1)} \sqrt{\rho}\|_F^2 \\
&\leqslant \|M\|^2 \|(\Pi^{(H)}_{(-\infty,i-\Delta)} + \Pi^{(H)}_{[i+1+\Delta,\infty)}) U \Pi^{(H)}_{[i,i+1)}\|^2 \|\Pi^{(H)}_{[i,i+1)} \sqrt{\rho}\|_F^2 \\
&\leqslant \|(\Pi^{(H)}_{(-\infty,i-\Delta)} + \Pi^{(H)}_{[i+1+\Delta,\infty)}) U \Pi^{(H)}_{[i,i+1)}\|^2 \|\Pi^{(H)}_{[i,i+1)} \sqrt{\rho}\|_F^2 \\
&\leqslant 2(e^{-\frac{\Delta}{4(\mathfrak{d}+1)\mathfrak{K}}} e^{\frac{|S|}{\mathfrak{K}}})^2 \|\Pi^{(H)}_{[i,i+1)} \sqrt{\rho}\|_F^2 \\
&= 2 e^{\frac{2|S|}{\mathfrak{K}}} e^{-\frac{\Delta}{2(\mathfrak{d}+1)\mathfrak{K}}} \|\Pi^{(H)}_{[i,i+1)} \sqrt{\rho}\|_F^2 \\
&= 2 e^{\frac{2|S|}{\mathfrak{K}}} e^{-\frac{\Delta}{2(\mathfrak{d}+1)\mathfrak{K}}} \|\Pi^{(H)}_{[i,i+1)} \sqrt{\rho}\|_F^2
\end{aligned}
$$

The remaining terms can be summed up nicely.

$$
\begin{aligned}
\langle (U^\dagger M U)^2 \rangle &\leqslant 2 \sum_i ((\mathrm{OFF}_i) + (\mathrm{ON}_i)) \\
&\leqslant 2 \sum_i (2 e^{\frac{2|S|}{\mathfrak{K}}} e^{-\frac{\Delta}{2(\mathfrak{d}+1)\mathfrak{K}}} \|\Pi^{(H)}_{[i,i+1)} \sqrt{\rho}\|_F^2 + \|M \Pi^{(H)}_{[i-\Delta,i+1+\Delta)} \sqrt{\rho}\|_F^2 e^{\beta \Delta})
\end{aligned}
$$

$$
\begin{aligned}
&= 4e^{\frac{2|S|}{\mathfrak{K}}} e^{-\frac{\Delta}{2(\mathfrak{d}+1)\mathfrak{K}}} \|\sqrt{\rho}\|_F^2 + (2\Delta+1)\|M\sqrt{\rho}\|_F^2 e^{\beta\Delta} \\
&= 4e^{\frac{2|S|}{\mathfrak{K}}} e^{-\frac{\Delta}{2(\mathfrak{d}+1)\mathfrak{K}}} + (2\Delta+1)e^{\beta\Delta}\langle M^2\rangle \\
&= \langle M^2\rangle e^{\beta\Delta}\Big(\frac{4e^{\frac{2|S|}{\mathfrak{K}}}}{\langle M^2\rangle} e^{-\Delta(\frac{1}{2(\mathfrak{d}+1)\mathfrak{K}}+\beta)} + 2\Delta + 1\Big)
\end{aligned}
$$

This holds for every $\Delta \geqslant 0$. We choose

$$
\Delta = \frac{1}{\frac{1}{2(\mathfrak{d}+1)\mathfrak{K}} + \beta} \log \frac{4e^{\frac{2|S|}{\mathfrak{K}}}}{\langle M^2\rangle}.
$$

Since $\langle M^2\rangle \leqslant 1$, Δ is indeed non-negative. With this choice of Δ, we have

$$
\begin{aligned}
\langle (U^\dagger M U)^2\rangle &\leqslant \langle M^2\rangle e^{\beta\Delta}\Big(\frac{4e^{\frac{2|S|}{\mathfrak{K}}}}{\langle M^2\rangle} e^{-\Delta(\frac{1}{2(\mathfrak{d}+1)\mathfrak{K}}+\beta)} + (2\Delta+1)\Big) \\
&= \langle M^2\rangle e^{\beta\Delta}(2\Delta+2) \\
&= \langle M^2\rangle (2\Delta+2)\Big(\frac{4e^{\frac{2|S|}{\mathfrak{K}}}}{\langle M^2\rangle}\Big)^{\frac{\beta}{\frac{1}{2(\mathfrak{d}+1)\mathfrak{K}}+\beta}} \\
&= (2\Delta+2)\big(4e^{\frac{2|S|}{\mathfrak{K}}}\big)^{\frac{2(\mathfrak{d}+1)\mathfrak{K}\beta}{1+2(\mathfrak{d}+1)\mathfrak{K}\beta}} \langle M^2\rangle^{\frac{1}{1+2(\mathfrak{d}+1)\mathfrak{K}\beta}} \\
&= \widetilde{\mathcal{O}}\Big((\mathfrak{d}+1)\mathfrak{K}|S|\big(4e^{\frac{2|S|}{\mathfrak{K}}}\big)^{\frac{2(\mathfrak{d}+1)\mathfrak{K}\beta}{1+2(\mathfrak{d}+1)\mathfrak{K}\beta}} \langle M^2\rangle^{\frac{1}{1+2(\mathfrak{d}+1)\mathfrak{K}\beta}}\Big).
\end{aligned}
$$

In the last line above, we are careful to pull out factors of $(\mathfrak{d}+1)\mathfrak{K}$ and $|S|$ to deal with regimes where $|S|/k$ and β are $\ll 1$ and $\gg 1$. □

Lemma B.1.3 ([15, Claim 38 + Corollary 39])**.** *Let U be a unitary supported on $S \subset [\mathfrak{n}]$ and let A be the local operator as defined previously. Then, for all $\gamma > 0$,*

$$
\langle (U^\dagger A U)^2\rangle \leqslant \gamma^2 + e^{c|S|}\Big(\langle \Pi^{(A)}_{(-\infty,-\gamma)\cup(\gamma,\infty)}\rangle^{\frac{1}{2(1+c'\beta)}} + \frac{\gamma+1}{\gamma^3}\langle A^2\rangle\Big)
$$

where c, c' are constants that depend on $\mathfrak{d}, \mathfrak{K}, \mathfrak{d}', \mathfrak{K}'$.

Our eventual goal is to get an upper bound in terms of $\langle A^2\rangle^{\Theta(\frac{1}{1+\beta})}$; since

$$
\langle \Pi^{(A)}_{(-\infty,-\gamma)\cup(\gamma,\infty)}\rangle \leqslant \gamma^{-2}\langle A^2\rangle, \tag{B.2}
$$

by setting γ appropriately we can do this.

Remark B.1.4. The original result proves this result for *quasi-local* operators, instead of the local operator case under consideration for us. They attain

$$\langle (U^\dagger A U)^2 \rangle \leqslant \gamma^2 + \frac{1}{\gamma} e^{\mathcal{O}(|S|)} \langle \Pi^{(A)}_{(-\infty,-\gamma)\cup(\gamma,\infty)} \rangle^{\mathcal{O}(1/\beta)} + \mathcal{O}(|S|^6 \frac{1}{\gamma^4} \langle A^2 \rangle) \tag{B.3}$$

The final $\frac{1}{\gamma^4}$ is not present in the statement, but is incurred in [15, Eq. 121].

Proof. Throughout, c's denote positive constants that depend on $k, k', \mathfrak{d}, \mathfrak{d}'$. We use Corollary 3.3.10 to conclude the exact same bound as used in Eq. (B.1), but this time for A.

$$\|\Pi^{(A)}_{[\sigma+\Delta,\infty]} U \Pi^{(A)}_{[-\infty,\sigma]}\| \leqslant \|U\| e^{-\frac{1}{4(\mathfrak{d}'+1)\mathfrak{K}'}(\Delta - 4(\mathfrak{d}'+1)|S|)} = e^{-\frac{\Delta}{4(\mathfrak{d}'+1)\mathfrak{K}'}} e^{\frac{|S|}{\mathfrak{K}'}}. \tag{B.4}$$

We split up our expression in pieces based on the eigenbasis of A. Let γ be a parameter that we choose later, and let $\Pi^{(A)}_i = \Pi^{(A)}_{[(-i-1)\gamma,-i\gamma)\cup[i\gamma,(i+1)\gamma)}$.

$$\begin{aligned}
\langle (U^\dagger A U)^2 \rangle &= \operatorname{tr}(U^\dagger A^2 U \rho) \\
&= \sum_{i \geqslant 0} \operatorname{tr}(U^\dagger A \Pi^{(A)}_i A U \rho) \\
&\leqslant \sum_{i \geqslant 0} (i+1)^2 \gamma^2 \operatorname{tr}(U^\dagger \Pi^{(A)}_i U \rho) \\
&= \sum_{i \geqslant 0} (i+1)^2 \gamma^2 \|\Pi^{(A)}_i U \sqrt{\rho}\|_F^2 \\
&\leqslant \gamma^2 + \gamma^2 \sum_{i \geqslant 1} (i+1)^2 \|\Pi^{(A)}_i U \sqrt{\rho}\|_F^2 \\
&\leqslant \gamma^2 + \gamma^2 \sum_{i \geqslant 1} (i+1)^2 \underbrace{\Big(\sum_{j \geqslant 0} \|\Pi^{(A)}_i U \Pi^{(A)}_j \sqrt{\rho}\|_F \Big)^2}_{(\text{TERM}_i)}
\end{aligned}$$

By Eq. (B.4), we can conclude

$$\|\Pi^{(A)}_i U \Pi^{(A)}_j \sqrt{\rho}\|_F \leqslant \|\Pi^{(A)}_i U \Pi^{(A)}_j\| \|\Pi^{(A)}_j \sqrt{\rho}\|_F \leqslant 16 e^{-\frac{\gamma|j-i|}{4(\mathfrak{d}'+1)\mathfrak{K}'}} e^{\frac{|S|}{\mathfrak{K}'}} \|\Pi^{(A)}_j \sqrt{\rho}\|_F \tag{B.5}$$

$$\leqslant e^{-c_0\gamma|j-i|}e^{c_1|S|}\|\Pi_j^{(A)}\sqrt{\rho}\|_F \tag{B.6}$$

We don't want to incur dependence on $\|\Pi_0^{(A)}\sqrt{\rho}\|_F^2 = \langle\Pi_0^{(A)}\rangle$. When $j = 0$, we can use Lemma B.1.2 to get a bound that still depends on $\langle I - \Pi_0^{(A)}\rangle$.

$$\begin{aligned}
\|\Pi_i^{(A)}U\Pi_0^{(A)}\sqrt{\rho}\|_F &\leqslant \|\Pi_i^{(A)}U\sqrt{\rho}\|_F + \|\Pi_i^{(A)}U(I-\Pi_0^{(A)})\sqrt{\rho}\|_F \\
&\leqslant \sqrt{e^{c_2|S|}\langle\Pi_i^{(A)}\rangle^{\frac{1}{1+c_3\beta}}} + \|(I-\Pi_0^{(A)})\sqrt{\rho}\|_F \\
&\leqslant e^{c_2|S|/2}\langle I-\Pi_0^{(A)}\rangle^{\frac{1}{2(1+c_3\beta)}}
\end{aligned} \tag{B.7}$$

With both of these bounds in hand, we can now bound (TERM_i). We use Eqs. (B.6) and (B.7) to get

$$\begin{aligned}
(\text{TERM}_i) &= \sum_{j\geqslant 0}\|\Pi_i^{(A)}U\Pi_j^{(A)}\sqrt{\rho}\|_F \\
&\leqslant \|\Pi_i^{(A)}U\Pi_0^{(A)}\sqrt{\rho}\|_F + \sum_{j\geqslant 1} e^{-c_0\gamma|j-i|}e^{c_1|S|}\|\Pi_j^{(A)}\sqrt{\rho}\|_F && \text{by Eq. (B.6)} \\
&\leqslant \sqrt{e^{c_2|S|/2}\langle I-\Pi_0^{(A)}\rangle^{\frac{1}{2(1+c_3\beta)}}e^{-c_0\gamma|0-i|}e^{c_1|S|}\|\Pi_0^{(A)}\sqrt{\rho}\|_F} + \sum_{j\geqslant 1} e^{-c_0\gamma|j-i|}e^{c_1|S|}\|\Pi_j^{(A)}\sqrt{\rho}\|_F \\
& && \text{by Eqs. (B.6) and (B.7)} \\
&\leqslant e^{(\frac{c_1}{2}+\frac{c_2}{4})|S|}\langle I-\Pi_0^{(A)}\rangle^{\frac{1}{4(1+c_3\beta)}}e^{-\frac{c_0}{2}\gamma i} + \sum_{j\geqslant 1} e^{-c_0\gamma|j-i|}e^{c_1|S|}\|\Pi_j^{(A)}\sqrt{\rho}\|_F \\
& && \text{by } \|\Pi_0^{(A)}\sqrt{\rho}\|_F \leqslant 1 \\
&\leqslant e^{c_4|S|}\left(\langle I-\Pi_0^{(A)}\rangle^{\frac{1}{4(1+c_3\beta)}}e^{-\frac{c_0}{2}\gamma i} + \Big(\sum_{j'\geqslant 1} e^{-c_0\gamma|j'-i|}\Big)^{\frac{1}{2}}\Big(\sum_{j\geqslant 1} e^{-c_0\gamma|j-i|}\|\Pi_j^{(A)}\sqrt{\rho}\|_F^2\Big)^{\frac{1}{2}}\right) \\
& && \text{by Cauchy-Schwarz and taking } c_4 = \max(\tfrac{c_1}{2}+\tfrac{c_2}{4}, c_1) \\
&\leqslant e^{c_4|S|}\Big(\langle I-\Pi_0^{(A)}\rangle^{\frac{1}{4(1+c_3\beta)}}e^{-\frac{c_0}{2}\gamma i} + c_5\sqrt{1+\frac{1}{\gamma}}\Big(\sum_{j\geqslant 1} e^{-c_0\gamma|j-i|}\|\Pi_j^{(A)}\sqrt{\rho}\|_F^2\Big)^{\frac{1}{2}}\Big) \\
& && \text{taking } c_5 \text{ appropriately} \\
(\text{TERM}_i)^2 &\leqslant 2e^{2c_4|S|}\Big(\langle I-\Pi_0^{(A)}\rangle^{\frac{1}{2(1+c_3\beta)}}e^{-c_0\gamma i} + c_5^2\Big(1+\frac{1}{\gamma}\Big)\Big(\sum_{j\geqslant 1} e^{-c_0\gamma|j-i|}\|\Pi_j^{(A)}\sqrt{\rho}\|_F^2\Big)\Big)
\end{aligned}$$

$$\leqslant e^{c_6|S|}\Big(\langle I-\Pi_0^{(A)}\rangle^{\frac{1}{2(1+c_3\beta)}}e^{-c_0\gamma i}+\Big(1+\frac{1}{\gamma}\Big)\Big(\sum_{j\geqslant 1}e^{-c_0\gamma|j-i|}\|\Pi_j^{(A)}\sqrt{\rho}\|_F^2\Big)\Big)$$

taking c_6 appropriately

Returning to what we originally wanted to bound,

$$\begin{aligned}
&\langle (U^\dagger AU)^2\rangle\\
&\leqslant \gamma^2+\gamma^2\sum_{i\geqslant 1}(i+1)^2(\mathrm{TERM}_i)^2\\
&\leqslant \gamma^2+\gamma^2 e^{c_6|S|}\sum_{i\geqslant 1}(i+1)^2\Big(\langle I-\Pi_0^{(A)}\rangle^{\frac{1}{2(1+c_3\beta)}}e^{-c_0\gamma i}+\Big(1+\frac{1}{\gamma}\Big)\Big(\sum_{j\geqslant 1}e^{-c_0\gamma|j-i|}\|\Pi_j^{(A)}\sqrt{\rho}\|_F^2\Big)\Big)\\
&= \gamma^2+e^{c_6|S|}\langle I-\Pi_0^{(A)}\rangle^{\frac{1}{2(1+c_3\beta)}}\sum_{i\geqslant 1}\gamma^2(i+1)^2e^{-c_0\gamma i}\\
&\quad+e^{c_6|S|}\frac{\gamma+1}{\gamma}\sum_{j\geqslant 1}j^2\|\Pi_j^{(A)}\sqrt{\rho}\|_F^2\sum_{i\geqslant 1}\gamma^2(\tfrac{i+1}{j})^2e^{-c_0\gamma|j-i|}\\
&\lesssim \gamma^2+c_7e^{c_6|S|}\langle I-\Pi_0^{(A)}\rangle^{\frac{1}{2(1+c_3\beta)}}+c_7e^{c_6|S|}\frac{\gamma+1}{\gamma}\sum_{j\geqslant 1}j^2\|\Pi_j^{(A)}\sqrt{\rho}\|_F^2\\
&\qquad\qquad\text{taking } c_7 \text{ such that } c_7\geqslant\textstyle\sum_{i\geqslant 1}\gamma^2(i+1)^2e^{-c_0\gamma|i-j|} \text{ for all } j\\
&= \gamma^2+e^{c_8|S|}\Big(\langle I-\Pi_0^{(A)}\rangle^{\frac{1}{2(1+c_3\beta)}}+\frac{\gamma+1}{\gamma^3}\langle A^2\rangle\Big)\qquad\text{by } \textstyle\sum_{j\geqslant 1}(\gamma j)^2\|\Pi_j^{(A)}\sqrt{\rho}\|_F^2\leqslant\langle A^2\rangle
\end{aligned}$$

□

B.1.2 Applying the Lemma

Lemma B.1.5 ([15, Claim 36])**.** *There exists a unitary U supported on* $\mathrm{supp}(A_{(i)})$ *such that*

$$\mathrm{tr}(A_{(i)}^2/N)\leqslant\mathrm{tr}((U^\dagger A_{(i)}U)^2\rho)=\langle(U^\dagger A_{(i)}U)^2\rangle$$

Proof. Take U to be the unitary that sends the ath largest eigenvector of $A_{(i)}^2$ to the ath largest eigenvector of $\mathrm{tr}_{[\mathrm{n}]\setminus\mathrm{supp}(A_{(i)})}\rho$.

□

Proof of Theorem 3.3.7. We consider $A_{(i)}$, and let U_* be the unitary whose existence is shown by Lemma B.1.5. Since $A_{(i)}$ is supported on $(\mathfrak{d}'+1)\mathfrak{K}'$ qubits (Eq. (3.6)), so is U_*.

$$\begin{aligned}
\operatorname{tr}(A_{(i)}^2/N) &\leqslant \langle (U_*^\dagger A_{(i)} U_*)^2 \rangle && \text{by Lemma B.1.5} \\
&= \|U_*^\dagger A_{(i)} U_* \sqrt{\rho}\|_F^2 \\
&= \Big\| U_*^\dagger \Big(A - \int \mathrm{d}\mu_i(U) U^\dagger A U \Big) U_* \sqrt{\rho} \Big\|_F^2 \\
&\leqslant 2\|U_*^\dagger A U_* \sqrt{\rho}\|_F^2 + 2\Big(\int \mathrm{d}\mu_i(U) \|U_*^\dagger U^\dagger A_{(i)} U U_* \sqrt{\rho}\|_F \Big)^2 \\
&\leqslant 4\gamma^2 + 4e^{c(\mathfrak{d}'+1)\mathfrak{K}'} \Big(\langle \Pi^{(A)}_{(-\infty,-\gamma)\cup(\gamma,\infty)} \rangle^{\frac{1}{2(1+c'\beta)}} + \frac{\gamma+1}{\gamma^3} \langle A^2 \rangle \Big) \\
& && \text{by Lemma B.1.3} \\
&\leqslant 4\gamma^2 + 4e^{c(\mathfrak{d}'+1)\mathfrak{K}'} \Big(\gamma^{-\frac{1}{1+c'\beta}} \langle A^2 \rangle^{\frac{1}{2(1+c'\beta)}} + \frac{\gamma+1}{\gamma^3} \langle A^2 \rangle \Big) && \text{by Eq. (B.2)}
\end{aligned}$$

We set $\gamma^2 = \operatorname{tr}(A_{(i)}^2/N)/8$ and $c'' = 8e^{c(\mathfrak{d}'+1)\mathfrak{K}'}$ so that this implies

$$\begin{aligned}
\operatorname{tr}(A_{(i)}^2/N) = 8\gamma^2 &\leqslant 4\gamma^2 + c'' \Big(\gamma^{-\frac{1}{1+c'\beta}} \langle A^2 \rangle^{\frac{1}{2(1+c'\beta)}} + \frac{\gamma+1}{\gamma^3} \langle A^2 \rangle \Big) \\
\frac{4}{c''}\gamma^2 &\leqslant \gamma^{-\frac{1}{1+c'\beta}} \langle A^2 \rangle^{\frac{1}{2(1+c'\beta)}} + \frac{\gamma+1}{\gamma^3} \langle A^2 \rangle
\end{aligned}$$

So, either

$$\begin{aligned}
\tfrac{2}{c''}\gamma^2 &\leqslant \gamma^{-\frac{1}{1+c'\beta}} \langle A^2 \rangle^{\frac{1}{2(1+c'\beta)}} \\
(\tfrac{2}{c''})^{2(1+c'\beta)} \gamma^{2+4(1+c'\beta)} &\leqslant \langle A^2 \rangle
\end{aligned} \tag{B.8}$$

or

$$\begin{aligned}
\frac{2}{c''}\gamma^2 &\leqslant \frac{\gamma+1}{\gamma^3} \langle A^2 \rangle \\
c'''\gamma^5 &\leqslant \langle A^2 \rangle
\end{aligned}$$

Above, we used that $\gamma = \operatorname{tr}(A_{(i)}^2/N)/8 \leqslant (\mathfrak{d}+1)/8$, so it can be folded into the constant. Eq. (B.8) gives us our final bound. □

B.2 Proof of Theorem 3.5.6

Here we prove Theorem 3.5.6. As the proof is quite long and computational, we break it into several manageable steps. First, we show that the even and odd truncations $s_{2\ell-1}(x), s_{2\ell}(x)$ can be related as follows:

Lemma B.2.1 (Even truncations are bounded)**.** *For all $\ell \in \mathbb{N}$, for all $x \in \mathbb{R}$,*

$$|s_{2\ell-1}(x)| < 99 s_{2\ell}(x).$$

Proof. First, note that whenever $s_{2\ell-1}(x) \geqslant 0$, we clearly have $s_{2\ell}(x) \geqslant |s_{2\ell-1}(x)|$ and the desired inequalities clearly hold. Thus, it suffices to consider when $s_{2\ell-1}(x) < 0$ which only happens when $x \leqslant -0.1\ell$. It remains to show that when $x \leqslant -0.1\ell$, $s_{2\ell}(x) + \frac{s_{2\ell-1}(x)}{99} \geqslant 0$. Let $f(x) = 0.99 s_{2\ell}(x) + 0.01 s_{2\ell-1}(x) - e^x$. Note that $f(0) = 0$ and $f^{(k)}(0) = 0$ for $k \leqslant 2\ell - 1$ and $f^{(2\ell)}(x) = 0.99 - e^x$ where $f^{(k)}(x)$ denotes the kth derivative of f. Now we can use the fundamental theorem of calculus to write

$$f(x) = \int_0^x f^{(2\ell)}(y) \frac{(x-y)^{2\ell-1}}{(2\ell-1)!}\,\mathrm{d}y = \int_0^x (0.99 - e^y) \frac{(x-y)^{2\ell-1}}{(2\ell-1)!}\,\mathrm{d}y\,.$$

We can rearrange the above as

$$f(-x) = \int_0^x (0.99 - e^{-y}) \frac{(x-y)^{2\ell-1}}{(2\ell-1)!}\,\mathrm{d}y\,.$$

It suffices to prove that for $x \geqslant 0.1\ell$ that $f(-x) \geqslant 0$ since then we get that $0.99 s_{2\ell}(-x) + 0.01 s_{2\ell-1}(-x) \geqslant e^{-x} \geqslant 0$. Now, we analyze the integral on the RHS of the above. We have

$$\begin{aligned}
&\int_0^x (0.99 - e^{-y})(x-y)^{2\ell-1}\,\mathrm{d}y \\
&\geqslant \int_{0.05}^{0.1} (x-y)^{2\ell-1}(0.99 - e^{-y})\,\mathrm{d}y - \int_0^{0.02} (e^{-y} - 0.99)(x-y)^{2\ell-1}\,\mathrm{d}y \\
&\geqslant 0.0019(x-0.1)^{2\ell-1} - 0.0002 x^{2\ell-1}
\end{aligned}$$

but since $x \geqslant 0.1\ell$,

$$\frac{(x-0.1)^{2\ell-1}}{x^{2\ell-1}} = \left(1 - \frac{0.1}{x}\right)^{2\ell-1} > \frac{1}{9}$$

and thus we conclude $f(-x) \geqslant 0$ and we are done. □

Next, we present a few basic facts about polynomials that allow us to construct bounded coefficient sum of squares representations.

Claim B.2.2. Let $p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0$ be a polynomial. Let $C > 0$ be a constant such that for all $i \in [n]$,

$$|a_i| \geqslant \frac{|a_{i-1}|}{C}\,.$$

Then all of the (complex) roots of p have magnitude at most $2C$.

Proof. Let $z \in \mathbb{C}$ have $|z| \geqslant 2C$. Then for all i,

$$|a_i z^i| \leqslant \frac{C^{n-i}}{|z|^{n-i}} \cdot |a_n z^n| \leqslant \frac{|a_n z^n|}{2^{n-i}}$$

and thus,

$$|a_n z^n| > |a_{n-1}z^{n-1}| + \cdots + |a_0|$$

so we cannot have

$$p(z) = a_n z^n + a_{n-1}z^{n-1} + \cdots + a_0 = 0\,.$$

Thus, it is impossible for z to be a root of p. □

As a straight-forward corollary, we show that each of the following admit a bounded coefficient sum-of-squares decomposition.

Corollary B.2.3 (Bounded coefficient Polynomials)**.** *Let ℓ be a positive integer and*

$-0.01 \leqslant c \leqslant 0.01$ *be a real number. Then, the polynomial*

$$s_{2\ell}(x) + cs_{2\ell-1}(x)$$

is a $(2^\ell, \ell, 10^\ell)$*-bounded SoS polynomial in* x *(see Definition 3.3.16).*

Proof. Note that all complex roots of $s_{2\ell}(x) + cs_{2\ell-1}(x)$ have magnitude at most 5ℓ by Claim B.2.2. Also, by Claim B.2.1, none of the roots are real so by the Fundamental Theorem of Algebra, they come in conjugate pairs $z_1, \overline{z_1}, \ldots, z_\ell, \overline{z_\ell}$. Then, we can write

$$s_{2\ell}(x) + c\ s_{2\ell-1}(x)99 = \frac{1}{(2\ell)!}\prod_{j\in[\ell]}(x - z_j)(x - \overline{z_j}) = \frac{1}{(2\ell)!}\prod_{j\in[\ell]}((x - \mathsf{Re}(z_j))^2 + \mathsf{Im}(z_j)^2) \tag{B.9}$$

and now we can expand out the product above to get a sum of 2^ℓ squares of polynomials of degree at most ℓ. Now we bound the coefficients of each of them. For each monomial x^k, its coefficient has magnitude at most

$$\binom{\ell}{k}\frac{(5\ell)^{\ell-k}}{\sqrt{(2l)!}} \leqslant \frac{\ell^k}{k!}\frac{(10\ell)^{\ell-k}}{\ell^\ell} = \frac{10^\ell}{k!} \tag{B.10}$$

and thus, we conclude that $s_{2\ell}(x) + cs_{2\ell-1}(x)$ is a $(2^\ell, \ell, 10^\ell)$-bounded sum-of-squares polynomial. □

Claim B.2.4. Let $p(x, y, t)$ be a polynomial such that for all $t \in [0, 1]$, it is a (k, d, C)-bounded SoS polynomial in x, y (after plugging in a real value for t). Then the polynomial

$$r(x, y) = \int_0^1 p(x, y, t)\,\mathrm{d}t$$

is a $(3d^2, d, \sqrt{k}C)$-bounded SoS polynomial in x, y.

Proof. We for each $t \in [0, 1]$, there are some (d, C)-bounded polynomials $q_{1,t}(x, y), \ldots, q_{k,t}(x, y)$ so that we can write

$$p(x, y, t) = q_{1,t}(x, y)^2 + \cdots + q_{k,t}(x, y)^2\,.$$

Let $v(x, y)$ be the vector of monomials

$$v(x,y) = \left(1, x, y, \frac{x^2}{2!}, \frac{xy}{2!}, \frac{y^2}{2!}, \ldots, \frac{x^d}{d!}, \frac{x^{d-1}y}{d!}, \ldots, \frac{y^d}{d!}\right).$$

Then we can associate each $q_{i,t}$ with a vector $u_{i,t}$ such that $q_{i,t}(x, y) = v(x, y)^\top u_{i,t}$. Since the $q_{i,t}$ are (d, C)-bounded, we know that all entries of $u_{i,t}$ are at most C. Define the matrix

$$M(t) = \sum_{i=1}^{k} u_{i,t} u_{i,t}^\top .$$

Then $M(t)$ is PSD and

$$p(x, y, t) = v(x, y)^\top M(t) v(x, y) .$$

Now we can write

$$r(x,y) = \int_0^1 p(x, y, t)\,\mathrm{d}t = v(x, y)^\top \left(\int_0^1 M(t)\,\mathrm{d}t\right) v(x, y) .$$

We know that all of the entries of $R(t) = \int_0^1 M(t)\,\mathrm{d}t$ are at most kC^2 and also it is a $\binom{d+2}{2} \times \binom{d+2}{2}$ matrix so we can write it as

$$R(t) = \sum_{i=1}^{\binom{d+2}{2}} u_i u_i^\top$$

for some vectors u_i whose entries are at most $\sqrt{k}C$. Now we can write

$$r(x,y) = \sum_{i=1}^{\binom{d+2}{2}} (v(x, y)^\top u_i)^2$$

and each of the $v(x, y)^\top u_i$ are $(d, \sqrt{k}C)$-bounded polynomials in x, y. Thus, $r(x, y)$ is a $(3d^2, d, \sqrt{k}C)$-bounded sum-of-squares polynomial, as desired. □

Now we move onto the main proof of Theorem 3.5.6. We have the following basic identities.

Fact B.2.5 (Integration Identities)**.** *For a function $p : \mathbb{R} \to \mathbb{R}$,*

$$p(z+a) - p(z-a) = a \int_0^1 (p'(z+ta) + p'(z-ta))\,\mathrm{d}t$$

and

$$p(z+a) + p(z-a) = 2p(z) + a \int_0^1 (p'(z+ta) + p'(z-ta))\,\mathrm{d}t$$

Proof. Both of the equations follow immediately from the Fundamental Theorem of Calculus. □

Recall that the key inequality that we need to prove is the following.

Lemma B.2.6 (Modified Gradient Identity with bounded coeffcieints)**.** *For all positive integers k, ℓ and real numbers x, y,*

$$0.5(x-y)(1+0.25(x-y)^2)(q_{k,\ell}(x) - q_{k,\ell}(y)) - 0.00025(x-y)^2 p_{k,\ell}(x) \geqslant 0\,.$$

Furthermore, the LHS is a $(10^{2^k\ell}, 2^k\ell + 10, 200^{2^k\ell})$-bounded SoS polynomial (in the variables x, y).

First, we prove the following algebraic identity which will be used to rewrite the expression in Lemma B.2.6 in a form for which it is easy to prove non-negativity.

Lemma B.2.7 (Relating q to first and second derivative)**.** *Let k, ℓ be positive integers. Let*

$$r(x,y) = 0.5(x-y)(1+0.25(x-y)^2)(q_{k,\ell}(x)-q_{k,\ell}(y)) - 0.00025(x-y)^2(p_{k,\ell}(x)+p_{k,\ell}(y))\,.$$

Let $z = (x+y)/2$ and $a = (x-y)/2$. We have the equality

$$\begin{aligned} r(x,y) = &\underbrace{\int_0^1 \left(\left(0.998a^2 + a^4\right)p_{k,\ell}(z+ta) - 0.001a^3(2-t)p'_{k,\ell}(z+ta)\right)\mathrm{d}t}_{\text{(B.11).(1)}} \\ &+ \underbrace{\int_0^1 \left(\left(0.998a^2 + a^4\right)p_{k,\ell}(z-ta) - 0.001a^3(2-t)p'_{k,\ell}(z-ta)\right)\mathrm{d}t}_{\text{(B.11).(2)}} \end{aligned} \tag{B.11}$$

Proof. Recalling that $p_{k,\ell}$ is the derivative of $q_{k,\ell}$, we can write

$$\begin{aligned} r(x,y) &= (a+a^3)(q_{k,\ell}(z+a) - q_{k,\ell}(z-a)) - 0.001a^2(p_{k,\ell}(z+a) + p_{k,\ell}(z-a)) \\ &= (a^2+a^4)\int_0^1 p_{k,\ell}(z+ta) + p_{k,\ell}(z-ta)\,\mathrm{d}t \\ &\quad - 0.002a^2 p_{k,\ell}(z) - 0.001a^3\int_0^1 p'_{k,\ell}(z+ta) - p'_{k,\ell}(z-ta)\,\mathrm{d}t, \end{aligned} \tag{B.12}$$

where p' is the derivative of p and the equality follows from applying Fact B.2.5 to $q_{k,\ell}$ and $p_{k,\ell}$. Next, observe

$$\begin{aligned} &2p_{k,\ell}(z) = 2\int_0^1 p_{k,\ell}(z)\,\mathrm{d}t \\ &= 2\int_0^1 p_{k,\ell}(z)\,\mathrm{d}t - \int_0^1 (p_{k,\ell}(z+ta) + p_{k,\ell}(z-ta)) + \int_0^1 (p_{k,\ell}(z+ta) + p_{k,\ell}(z-ta)) \\ &= \int_0^1 (p_{k,\ell}(z+ta) - p_{k,\ell}(z))\,\mathrm{d}t - \int_0^1 (p_{k,\ell}(z) - p_{k,\ell}(z-ta))\,\mathrm{d}t + \int_0^1 (p_{k,\ell}(z+ta) + p_{k,\ell}(z-ta))\,\mathrm{d}t \\ &= \int_0^1 a\left(\int_0^t \left(p'_{k,\ell}(z+sa) - p'_{k,\ell}(z-sa)\right)\mathrm{d}s\right)\mathrm{d}t + \int_0^1 (p_{k,\ell}(z+ta) + p_{k,\ell}(z-ta))\,\mathrm{d}t \end{aligned} \tag{B.13}$$

Substituting the expanded expression for $2p_{k,\ell}$ back into Equation (B.12),

$$\begin{aligned}
r(x,y) &= \big(0.998a^2 + a^4\big) \int_0^1 (p_{k,\ell}(z+ta) + p_{k,\ell}(z-ta))\,\mathrm{d}t \\
&\quad - 0.001a^2 \int_0^1 a\bigg(\int_0^t \big(p'_{k,\ell}(z+sa) - p'_{k,\ell}(z-sa)\big)\,\mathrm{d}s\bigg)\,\mathrm{d}t \\
&\quad - 0.001a^3 \int_0^1 p'_{k,\ell}(z+ta) - p'_{k,\ell}(z-ta)\,\mathrm{d}t \\
&= \big(0.998a^2 + a^4\big) \int_0^1 (p_{k,\ell}(z+ta) + p_{k,\ell}(z-ta))\,\mathrm{d}t \\
&\quad - 0.001a^3 \int_0^1 (1-s)\big(p'_{k,\ell}(z+sa) - p'_{k,\ell}(z-sa)\big)\,\mathrm{d}s \\
&\quad - 0.001a^3 \int_0^1 \big(p'_{k,\ell}(z+ta) - p'_{k,\ell}(z-ta)\big)\,\mathrm{d}t \\
&= \big(0.998a^2 + a^4\big) \int_0^1 (p_{k,\ell}(z+ta) + p_{k,\ell}(z-ta))\,\mathrm{d}t \\
&\quad - 0.001a^3 \int_0^1 (2-t)\big(p'_{k,\ell}(z+ta) - p'_{k,\ell}(z-ta)\big)\,\mathrm{d}t, \\
&= \underbrace{\int_0^1 \Big(\big(0.998a^2 + a^4\big)p_{k,\ell}(z+ta) - 0.001a^3(2-t)p'_{k,\ell}(z+ta)\Big)\,\mathrm{d}t}_{\text{(B.11).(1)}} \\
&\quad + \underbrace{\int_0^1 \Big(\big(0.998a^2 + a^4\big)p_{k,\ell}(z-ta) - 0.001a^3(2-t)p'_{k,\ell}(z-ta)\Big)\,\mathrm{d}t}_{\text{(B.11).(2)}}
\end{aligned}$$

where the second equality follows from switching the order of the integral and observing that the coefficient for a fixed s is simply $1-s$. □

In light of Lemma B.2.7, it remains to prove that the expressions (B.11).(1) and (B.11).(2) are nonnegative and can be written as bounded SoS polynomials in the variables x, y.

Lemma B.2.8. *Let $z = (x+y)/2$ and $a = (x-y)/2$. Then the expressions* (B.11)*.(1) and* (B.11)*.(2) are nonnegative and can be written as a $(10 \cdot (2^k\ell)^2, 2^k\ell + 10, 150^{2^k\ell})$-bounded SoS polynomials in the variables x, y.*

Proof. We focus on the expression (B.11).(1). The argument for (B.11).(2) is exactly

the same. First, we show that expression (B.11).(1) is non-negative. Using $1 + a^2 = \frac{(1+a)^2+(1-a)^2}{2}$ and $a = \frac{(1+a)^2-(1-a)^2}{4}$,

$$\begin{aligned}
(\text{B.11}).(1) &= a^2 \cdot \int_0^1 \Big(\big(0.998 + a^2\big) p_{k,\ell}(z+ta) - 0.001a(2-t) p'_{k,\ell}(z+ta) \Big)\, \mathrm{d}t \\
&= a^2 \cdot \int_0^1 (2-t) \left(\frac{(1+a^2)}{8} p_{k,\ell}(z+ta) - \frac{0.001((1+a)^2-(1-a)^2)}{4} p'_{k,\ell}(z+ta) \right) \mathrm{d}t \\
&\quad + a^2 \cdot \int_0^1 \left((0.998 + a^2) - \frac{(2-t)(1+a^2)}{8} p_{k,\ell}(z+ta) \right) \mathrm{d}t \\
&= a^2 \cdot \int_0^1 \frac{(2-t)(1+a)^2}{16} \big(p_{k,\ell}(z+ta) - 0.004 p'_{k,\ell}(z+ta) \big)\, \mathrm{d}t \\
&\quad + a^2 \cdot \int_0^1 \frac{(2-t)(1-a)^2}{16} \big(p_{k,\ell}(z+ta) + 0.004 p'_{k,\ell}(z+ta) \big)\, \mathrm{d}t, \\
&\quad + a^2 \cdot \int_0^1 \left((0.998 + a^2) - \frac{(2-t)(1+a^2)}{8} \right) p_{k,\ell}(z+ta)\, \mathrm{d}t.
\end{aligned} \tag{B.14}$$

To prove nonnegativitiy, it suffices to show that $a_{k,\ell}(x) = p_{k,\ell}(x) - 0.004 p'_{k,\ell}(x)$ and $b_{k,\ell}(x) = p_{k,\ell}(x) + 0.004 p'_{k,\ell}(x)$ are non-negative for all $x \in \mathbb{R}$. Recall, $p_{k,\ell}(x) = \prod_{j\in[k]} s_{2^j\ell}(x/k)$. Therefore, by product rule,

$$\begin{aligned}
p'_{k,\ell}(x) &= \sum_{j\in[k]} \frac{s'_{2^j\ell}(x/k)}{k} \cdot \prod_{j'\neq j\in[k]} s_{2^{j'}\ell}(x/k) \\
&= \sum_{j\in[k]} \frac{s_{(2^j\ell)-1}(x/k)}{k} \cdot \prod_{j'\neq j\in[k]} s_{2^{j'}\ell}(x/k).
\end{aligned}$$

Therefore,

$$a_{k,\ell}(x) = \sum_{j\in[k]} \left(\frac{s_{2^j\ell}(x/k) - 0.004 \cdot s_{(2^j\ell)-1}(x/k)}{k} \right) \prod_{j'\neq j\in[k]} s_{2^{j'}\ell}(x/k) \geqslant 0 \tag{B.15}$$

since for all $j \in [k]$, $s_{2^j\ell}(x/k) - 0.004 s_{(2^j\ell)-1}(x/k) \geqslant 0$ by Corollary B.2.3 and

$s_{2^{j'}\ell}(x/k) \geqslant 0$ by Corollary 3.5.9. Similarly,

$$b_{k,\ell}(x) = \sum_{j\in[k]} \left(\frac{s_{2^j\ell}(x/k) + 0.004 \cdot s_{(2^j\ell)-1}(x/k)}{k} \right) \prod_{j'\neq j\in[k]} s_{2^{j'}\ell}(x/k) \geqslant 0 \tag{B.16}$$

since for all $j \in [k]$, $s_{2^j\ell}(x/k) + 0.004 s_{(2^j\ell)-1}(x/k) \geqslant 0$ by Corollary B.2.3.

Now we have proved that the expression is nonnegative, we show that it is a SoS polynomial with bounded coefficients. By Corollary B.2.3 the polynomials

$$s_{2^j\ell}(x/k) \pm 0.004 s_{(2^j\ell)-1}(x/k), s_{2^j\ell}(x/k)$$

are all $(2^{2^{j-1}\ell}, 2^{j-1}\ell, 10^{2^{j-1}\ell})$-bounded SoS polynomials. Thus, by Claim 3.3.17, $a_{k,\ell}(x)$ and $b_{k,\ell}(x)$ are both $(k \cdot 2^{2^k\ell}, 2^k\ell, 90^{2^k\ell})$-bounded SoS polynomials. Since $z + ta = \frac{1+t}{2}x + \frac{1-t}{2}y$, by Claim 3.3.17 again, we get that for any real number $t \in [0, 1]$,

$$\frac{(2-t)(1+a)^2}{16}\big(p_{k,\ell}(z+ta) - 0.004 p'_{k,\ell}(z+ta)\big)$$

is a $(k\cdot 2^{2^k\ell}, 2^k\ell, 100^{2^k\ell})$-bounded SoS polynomial in the variables x, y after substituting $z = (x + y)/2, a = (x - y)/2$. We can use a similar argument for the other terms in the expression in (B.14). Then we can use Claim B.2.4 to bound the integral over t and deduce that the expression (B.11).(1) is a $(10 \cdot (2^k\ell)^2, 2^k\ell + 10, 150^{2^k\ell})$-bounded SoS polynomial in the variables x, y.

□

Now we can complete the proof of Lemma B.2.6.

Proof of Lemma B.2.6. Note that the expression in Lemma B.2.6 is equal to

$$r(x, y) + 0.00025(x - y)^2 p_{k,\ell}(y)\,.$$

Now by the definition of $p_{k,\ell}$, Corollary B.2.3, and Claim 3.3.17, we get that $p_{k,\ell}(y)$

is is a $(2^{2^k\ell}, 2^k\ell, 90^{2^k\ell})$-bounded SoS polynomial. Lemma B.2.8 and Lemma B.2.7 imply that $r(x, y)$ is a $(20 \cdot (2^k\ell)^2, 2^k\ell + 10, 150^{2^k\ell})$-bounded SoS polynomial so overall, $r(x, y) + 0.00025(x - y)^2 p_{k,\ell}(y)$ is a $(10^{2^k\ell}, 2^k\ell + 10, 200^{2^k\ell})$-bounded SoS polynomial (and thus also nonnegative) and we are done. □

Finally, this completes the proof of Theorem 3.5.6.

Proof of Theorem 3.5.6. The desired result follows immediately from Lemma B.2.6. □

Appendix C

Omitted Proofs from Chapter 5

C.1 Multi-Block Distinguishing Task

The proofs from the preceding sections imply a slightly weaker version of Theorem 5.5.1 where the lower bound involves removing $\varepsilon \log(d/\varepsilon)$ mass from σ. Avoiding the extra log factor requires working with a slightly more involved instance than the one from Section 5.3 in which diagonal blocks of σ at *all* scales are perturbed.

To that end, here we analyze the following more general distinguishing task.

$$H_0 : \rho = \frac{1}{d} \begin{pmatrix} A_1 & & \\ & \ddots & \\ & & A_m \end{pmatrix} \qquad \text{and} \qquad H_1 : \rho = \frac{1}{d} \begin{pmatrix} A_1 + \varepsilon_1 \overline{M}_1 & & \\ & \ddots & \\ & & A_m + \varepsilon_m \overline{M}_m \end{pmatrix}. \tag{C.1}$$

Here, each $A_\nu \in \boldsymbol{R}^{d_\nu \times d_\nu}$ is a diagonal matrix. There exist numbers $j_1, \ldots, j_m$ such that each A_ν has diagonal entries in the interval $[d \cdot 2^{-j_\nu}, d \cdot 2^{-j_\nu+1}]$, and $\sum_\nu \operatorname{tr}(A_\nu) = d$. Furthermore, for every $\nu \in [m]$, $\overline{M}_\nu \sim \mathrm{GOE}^*_{U_\nu}(d_\nu)$ for the events $U_\nu \triangleq U_{d_\nu}$ given by Lemma C.1.1 below.

We will refer to the set of d_ν row/column indices of ρ which correspond to A_ν as $\mathcal{B}_\nu$. Let $\mathbb{S}_\nu$ denote the set of unit vectors in $\mathbb{C}^{d-1}$ with entries supported on $\mathcal{B}_\nu$.

The following lemma follows easily from the proof of Lemma 5.3.2.

Lemma C.1.1. *There is an absolute constant $a^* > 0$ such that for any integer $d' \geqslant a^*$, there exists $U_{d'} \subseteq \boldsymbol{R}^{d' \times d'}$ such that if $M \sim \mathrm{GOE}^*(d')$, then $\mathbf{Pr}\, M \notin U_{d'} \leqslant o(1/m)$ and on the event $M \in U_{d'}$, we have $\|M\|_{\mathrm{op}} \leqslant 3 + \Theta(\sqrt{\log(m)/d'})$ and $\|M\|_1 \geqslant d'/12$.*

Our main result for the distinguishing task (C.1) is the following.

Theorem C.1.2. *Let $\varepsilon \triangleq \frac{1}{d}\sum_\nu d_\nu \varepsilon_\nu$ and $N \triangleq \frac{1}{m}\min_{\nu\in[m]} \frac{d_\nu^{1/2} d^2}{\varepsilon_\nu^2 2^{j_\nu}}$. For a^* from Lemma C.1.1, if $N \geqslant 1/\varepsilon$, $d_\nu \geqslant a^*$ for all ν, and*

$$\varepsilon_\nu \leqslant d \cdot 2^{-j_\nu}/(12 + \Theta(\sqrt{\log(m)/d_\nu})) \qquad \text{and} \qquad d_\nu/2^{j_\nu} \geqslant 2\varepsilon/\log(d/\varepsilon) \quad \forall\ \nu \in [m] \tag{C.2}$$

then the copy complexity of distinguishing between H_0 and H_1 with incoherent measurements with success probability at least $2/3$ is $\widetilde{\Omega}(N)$.

Note that the bounds in Lemma C.1.1 and the first part of (C.2) ensure that under H_1, ρ is psd (and thus a valid quantum state) and has trace distance at least $\Omega(\frac{1}{d}\sum_\nu d_\nu \varepsilon_\nu)$ to the null hypothesis.

Block structure of POVMs. Take any learning tree $\mathcal{T}$ corresponding to an algorithm for this task that uses n incoherent measurements. The following lemma shows that we can assume without loss of generality that every POVM respects the block structure in the distinguishing task, that is, it consists of $\omega_x d \cdot xx^\dagger$ for which each $x \in \mathbb{S}_\nu$ for some $\nu \in [m]$.

Lemma C.1.3. *Given an arbitrary d-dimensional POVM $\{E_x\}$, there is a corresponding rank-1 POVM $\{E'_y\}$ satisfying the following. Let p, p' be the distributions over measurement outcomes from measuring a state ρ with these POVMs respectively. Then:*

- *For every E'_y, there is some ν such that E'_y is zero outside the principal submatrix indexed by $\mathcal{B}_\nu$.*

- *There is an explicit function f mapping outcomes x of the former POVM to outcomes of the latter for which the pushforward of p' under f is p.*

Proof. This immediately follows from [65, Lemma 5.6] and the fact that we can always assume without loss of generality that POVMs are rank-1. □

Given any $x \in \mathbb{S}_\nu$, we use the notation $\nu(x)$ to denote the $\nu \in [m]$ for which $x \in \mathbb{S}_\nu$. Observe that for any $\nu \in [m]$,

$$\sum_{x:\nu(x)=\nu} \omega_x(xx^\dagger - I_{d_\nu}) = 0 \tag{C.3}$$

The fact that we can assume every POVM respects the block structure will allow our proof to proceed along very similar lines to that of Theorem 5.3.3, the key distinction being that instead of tracking the overall likelihood ratio, we track one likelihood ratio for each set of coordinates $\mathcal{B}_\nu$.

Recalling the terminology from Definition 4.2.1, we let p_0 and p_1 denote the distributions over leaves of $\mathcal{T}$ induced by ρ under H_0 and H_1 respectively. In the rest of this section, let ξ be a slowly-growing function satisfying $\xi \gg \log^c(d/\varepsilon)$ for some absolute constant $c > 0$. We assume

$$n \lll \frac{1}{\xi m} \cdot \min_{\nu\in[m]} \frac{d_\nu^{1/2} d^2}{\varepsilon_\nu^2 2^{j_\nu}}. \tag{C.4}$$

and will prove $d_{\mathrm{TV}}(p_0, p_1) = o(1)$. By the hypothesis in Theorem C.1.2 that $N \geqslant 1/\varepsilon$, we may also assume

$$n \geqslant 1/\varepsilon \tag{C.5}$$

by adding superfluous measurements to the algorithm. We set α, β to be slowly-growing functions such that $m^{1/2}\xi^{1/2} \ll \alpha \ll \beta \ll \min_\nu \frac{d_\nu^{1/2} d^2}{\varepsilon^2 2^{j_\nu} n}$. Note that these choices are possible by (C.4) and (C.5).

We let $L^*(\cdot)$ denote the likelihood ratio between p_1 and p_0. That is, for a sequence of

vectors $\boldsymbol{x} = (x_1, \ldots, x_n)$, let $L^*(\boldsymbol{x}) \triangleq p_1(\boldsymbol{x})/p_0(\boldsymbol{x})$. For $\overline{M}_1 \sim \mathrm{GOE}^*_{U_1}(d_1), \ldots, \overline{M}_m \sim \mathrm{GOE}^*_{U_m}(d_m)$, note that because $\overline{M}_1, \ldots, \overline{M}_m$ are independent, $L^*(\boldsymbol{x}) = \prod_{\nu=1}^m L^*_\nu(\boldsymbol{x})$ where

$$L^*_\nu(\boldsymbol{x}) \triangleq \mathbb{E}[\overline{M}_\nu] * \prod_{i \in [n]: \nu(x_i) = \nu} \left(1 + \varepsilon_\nu \frac{x_i^\dagger \overline{M}_\nu x_i}{x_i^\dagger A_\nu x_i}\right). \tag{C.6}$$

For $M_1 \sim \mathrm{GOE}^*(d_1), \ldots, M_m \sim \mathrm{GOE}^*(d_m)$, define similarly $L(\boldsymbol{x}) = \prod_{\nu=1}^m L_\nu(\boldsymbol{x})$ for

$$L_\nu(\boldsymbol{x}) \triangleq \mathbb{E}[M_\nu] * \prod_{i \in [n]: \nu(x_i) = \nu} \left(1 + \varepsilon_\nu \frac{x_i^\dagger \overline{M}_\nu x_i}{x_i^\dagger A_\nu x_i}\right). \tag{C.7}$$

Note that $L(\boldsymbol{x})$ (resp. $L_\nu(\boldsymbol{x})$) is an estimate for the likelihood ratio $L^*(\boldsymbol{x})$ (resp. $L^*_\nu(\boldsymbol{x})$) where the conditioned Gaussian integral is replaced by a true Gaussian integral. Most of the computations in this section will be done in terms of L instead of L_ν; the proof of Theorem C.1.2 below quantifies that $L(\boldsymbol{x})$ is a close approximation of $L^*(\boldsymbol{x})$.

As before, we will somewhat abuse notation and write $L(\boldsymbol{z})$ for any sequence of unit vectors $\boldsymbol{z} = (z_1, \ldots, z_t)$ of length not necessarily n, that is, $L(\boldsymbol{z}) = \prod_{\nu=1}^m L_\nu(\boldsymbol{z})$ where $L_\nu(\boldsymbol{z})$ is defined the same way as in (C.7). We also write $L_\nu(\boldsymbol{x}, \boldsymbol{x})$ to denote the value of L_ν on input $(x_1, x_1, x_2, x_2, \ldots, x_n, x_n)$.

The main ingredient in the proof of Theorem C.1.2 is the following analogue of Proposition 5.3.4 giving a high-probability bound on each L_ν evaluated at the leaves of $\mathcal{T}$.

Proposition C.1.4. *There exists a subset S of the leaves of $\mathcal{T}$ such that $\mathbf{Pr}[p_0] * S = 1 - o(1)$ and for all $\boldsymbol{x} \in S$, $|L_\nu(\boldsymbol{x}) - 1| = o(1/m)$ and $L_\nu(\boldsymbol{x}, \boldsymbol{x}) \ll e^{\sqrt{d_\nu}}$ for all $\nu \in [m]$.*

Let us first prove Theorem C.1.2 assuming Proposition C.1.4.

Proof of Theorem C.1.2. Let U denote the event that $M_\nu \in U_\nu$ for all $\nu \in [m]$. Define

$$\overline{L}(\boldsymbol{x}) \triangleq \mathbb{E}[M_1 \sim \mathrm{GOE}^*(d_1), \ldots, M_m \sim \mathrm{GOE}^*(d_m)] * I\{U\} \prod_{i=1}^{n} \left(1 + \varepsilon_{\nu(x_i)} \frac{x_i^\dagger M_{\nu(x_i)} x_i}{x_i^\dagger A_{\nu(x_i)} x_i}\right) \tag{C.8}$$

$$= \prod_{\nu=1}^{m} \mathbb{E}[M_\nu \sim \mathrm{GOE}^*(d_\nu)] * \mathbf{1}U_\nu \prod_{i \in [n]: \nu(x_i) = \nu} \left(1 + \varepsilon_\nu \frac{x_i^\dagger M_\nu x_i}{x_i^\dagger A_\nu x_i}\right) \triangleq \prod_{\nu=1}^{m} \overline{L}_\nu(\boldsymbol{x}). \tag{C.9}$$

It is clear that $L_\nu^*(\boldsymbol{x}) = \mathbf{Pr} * {U_\nu}^{-1} \overline{L}_\nu(\boldsymbol{x})$. For all $\boldsymbol{x} \in S$ and $\nu \in [m]$, by Cauchy-Schwarz

$$\begin{aligned} |L_\nu(\boldsymbol{x}) - \overline{L}_\nu(\boldsymbol{x})| &= \left| \mathbb{E}[M_\nu] * I\{U_\nu^c\} \prod_{i \in [n]: \nu(x_i) = \nu} \left(1 + \varepsilon_\nu \frac{x_i^\dagger M_\nu x_i}{x_i^\dagger A_\nu x_i}\right) \right| \\ &\leqslant \sqrt{\mathbf{Pr} * U^c L_\nu(\boldsymbol{x}, \boldsymbol{x})} = o(1/m). \end{aligned}$$

Here we use that $\mathbf{Pr} * U_\nu^c \leqslant \mathrm{poly}(1/m) \cdot \exp(-\Omega(d_\nu))$ and $L_\nu(\boldsymbol{x}, \boldsymbol{x}) \ll e^{\sqrt{d_\nu}}$. Moreover, we have $|L_\nu(\boldsymbol{x}) - 1| = o(1/m)$. Thus, for all $\boldsymbol{x} \in S$ and $\nu \in [m]$, $\overline{L}_\nu(\boldsymbol{x}) = 1 + o(1/m)$ and

$$\begin{aligned} |L_\nu^*(\boldsymbol{x}) - 1| &\leqslant |L_\nu^*(\boldsymbol{x}) - \overline{L}_\nu(\boldsymbol{x})| + |\overline{L}_\nu(\boldsymbol{x}) - 1| \\ &= \frac{\mathbf{Pr} * U_\nu^c}{\mathbf{Pr} * U_\nu} \overline{L}_\nu(\boldsymbol{x}) + o(1/m) = o(1/m). \end{aligned}$$

Recalling that $L^*(\boldsymbol{x}) = \prod_{\nu=1}^m L_\nu^*(\boldsymbol{x})$, we conclude that $L^*(\boldsymbol{x}) = 1 + o(1)$. Finally,

$$\begin{aligned} d_{\mathrm{TV}}(p_0, p_1) &= 2\,\mathbb{E}[\boldsymbol{x} \sim p_0] * (L^*(\boldsymbol{x}) - 1)_- \\ &= 2\,\mathbb{E}[\boldsymbol{x} \sim p_0] * I\{\boldsymbol{x} \in S\}(L^*(\boldsymbol{x}) - 1)_- + 2\,\mathbb{E}[\boldsymbol{x} \sim p_0] * I\{\boldsymbol{x} \notin S\}(L^*(\boldsymbol{x}) - 1)_- \\ &\leqslant 2 \sup_{\boldsymbol{x} \in S} (L^*(\boldsymbol{x}) - 1)_- + 2\,\mathbf{Pr}[p_0] * S^c = o(1). \end{aligned}$$

□

C.1.1 Recursive Evaluation of Likelihood Ratio

Let $\boldsymbol{z} = (z_1, \dots, z_t)$ be a sequence of unit vectors. For $1 \leqslant i \leqslant t$, let $\boldsymbol{z}_{\sim i}$ be the sequence $\boldsymbol{z}$ with z_i omitted. Similarly, for $1 \leqslant i < j \leqslant t$, let $\boldsymbol{z}_{\sim i,j}$ be the sequence $\boldsymbol{z}$ with z_i, z_j omitted. The main result of this subsection is the following recursive formula for $L(\boldsymbol{z})$.

Lemma C.1.5. *The function L_ν satisfies*

$$L_\nu(\boldsymbol{z}) = L_\nu(\boldsymbol{z}_{\sim t}) + \frac{2\varepsilon_\nu^2}{d_\nu^2} \cdot \mathbf{1}\nu(z_t) = \nu \sum_{i<t:\nu(z_i)=\nu} \left[\frac{d_\nu \langle z_i, z_t\rangle^2 - 1}{(z_i^\dagger A_\nu z_i)(z_t^\dagger A_\nu z_t)} L_\nu(\boldsymbol{z}_{\sim i,t}) \right].$$

As with Lemma 5.3.5, the proof is based on Isserlis' theorem. For k even, recall that $\mathrm{PMat}(k)$ denotes the set of perfect matchings of $\{1, \dots, k\}$.

Proof of Lemma C.1.5. The case of $\nu(z_t) \neq \nu$ is clear. We now suppose $\nu(z_t) = \nu$. Let $J \subseteq [t]$ denote the indices s for which $\nu(x_s) = \nu$. For a set $S \subseteq J$ with $|S|$ even, let $\mathrm{PMat}(S)$ denote the set of perfect matchings of S. For even k, let $\mathrm{Mat}(J, k)$ denote the set of matchings of S consisting of $k/2$ pairs. We compute that

$$\begin{aligned}
L_\nu(\boldsymbol{z}) &= \sum_{S \subseteq J} \varepsilon_\nu^{|S|}\, \mathbb{E}[M_\nu \sim \mathrm{GOE}^*(d_\nu)] * \prod_{i \in S} \frac{z_i^\dagger M_\nu z_i}{z_i^\dagger A_\nu z_i} \qquad \text{(expanding definition of } L\text{)} \\
&= \sum_{\substack{S \subseteq J \\ |S| \text{ even}}} \varepsilon_\nu^{|S|} \sum_{\{\{a_1,b_1\},\dots,\{a_{|S|/2},b_{|S|/2}\}\} \in \mathrm{PMat}(S)} \prod_{i=1}^{|S|/2} \mathbb{E}[M_\nu \sim \mathrm{GOE}^*(d_\nu)] * \frac{z_{a_i}^\dagger M_\nu z_{a_i}}{z_{a_i}^\dagger A_\nu z_{a_i}} \cdot \frac{z_{b_i}^\dagger M_\nu z_{b_i}}{z_{b_i}^\dagger A_\nu z_{b_i}} \\
&= \sum_{k=0}^{\lfloor |J|/2 \rfloor} \varepsilon_\nu^{2k} \sum_{\{\{a_1,b_1\},\dots,\{a_k,b_k\}\} \in \mathrm{Mat}(J,2k)} \prod_{i=1}^{k} \mathbb{E}[M_\nu \sim \mathrm{GOE}^*(d_\nu)] * \frac{z_{a_i}^\dagger M_\nu z_{a_i}}{z_{a_i}^\dagger A_\nu z_{a_i}} \cdot \frac{z_{b_i}^\dagger M_\nu z_{b_i}}{z_{b_i}^\dagger A_\nu z_{b_i}} \\
&= \sum_{k=0}^{\lfloor |J|/2 \rfloor} \left(\frac{2\varepsilon_\nu^2}{d^2}\right)^k \sum_{\{\{a_1,b_1\},\dots,\{a_k,b_k\}\} \in \mathrm{Mat}(J,2k)} \prod_{i=1}^{k} \frac{d_\nu \langle z_{a_i}, z_{b_i}\rangle^2 - 1}{(z_{a_i}^\dagger A_\nu z_{a_i})(z_{b_i}^\dagger A_\nu z_{b_i})}. \qquad \text{(C.10)}
\end{aligned}$$

The lemma follows by partitioning the summands in (C.10) based on whether t appears in the matching, and if so which $i \in J$ it is paired with. □

C.1.2 High Probability Bound on Likelihood Ratio at Leaves

This subsection gives the main part of the proof of Proposition C.1.4. For any sequence of vectors $\boldsymbol{z} = (z_1, \dots, z_t)$ and $\nu \in [m]$, define

$$H_\nu(\boldsymbol{z}) = \sum_{i \leqslant t : \nu(z_i) = \nu} \frac{d_\nu z_i z_i^\dagger - I_{d_\nu}}{z_i^\dagger A_\nu z_i} \cdot \frac{L(\boldsymbol{z}_{\sim i})}{L(\boldsymbol{z})} \qquad \text{and} \qquad K_\nu(\boldsymbol{z}) = \sum_{i \leqslant t : \nu(z_i) = \nu} \frac{d_\nu z_i z_i^\dagger - I_{d_\nu}}{z_i^\dagger A_\nu z_i}.$$

H_ν enters our calculations by the following rewriting of Lemma C.1.5:

$$\frac{L_\nu(\boldsymbol{z})}{L_\nu(\boldsymbol{z}_{\sim t})} = 1 + \frac{2\varepsilon_\nu^2}{d_\nu^2} \cdot \mathbf{1}\nu(z_t) = \nu \cdot \frac{z_t^\dagger H_\nu(\boldsymbol{z}_{\sim t}) z_t}{z_t^\dagger A_\nu z_t}. \tag{C.11}$$

If $\boldsymbol{z} = \boldsymbol{x}_{\leqslant t} \triangleq (x_1, \dots, x_t)$ is a prefix of $\boldsymbol{x} \sim p_0$, then $\prod_{\nu=1}^m \frac{L_\nu(\boldsymbol{z})}{L_\nu(\boldsymbol{z}_{\sim t})} = \prod_{\nu=1}^m \frac{L_\nu(\boldsymbol{x}_{\leqslant t})}{L_\nu(\boldsymbol{x}_{\leqslant t-1})} = \frac{L(\boldsymbol{x}_{\leqslant t})}{L(\boldsymbol{x}_{\leqslant t-1})}$ is one step in the likelihood ratio martingale. We will control the contribution from each multiplicative martingale L_ν separately. As we will see (proof of Claim C.1.10) below, the multiplicative fluctuation of any such step is

$$\mathbb{E}[x_t] * \left(\frac{L_\nu(\boldsymbol{x}_{\leqslant t})}{L_\nu(\boldsymbol{x}_{\leqslant t-1})} \right)^2 = 1 + \frac{\varepsilon_\nu^4 2^{j_\nu}}{d^2 d_\nu^4} \cdot \|H_\nu(\boldsymbol{x}_{\leqslant t-1})\|_F^2.$$

Thus, an upper bound on $\|H_\nu(\boldsymbol{z})\|_F$ over all $\nu \in [m]$ and all prefixes $\boldsymbol{z}$ of $\boldsymbol{x}$ controls the fluctuations of the likelihood ratio martingale. Because the matrices output by H_ν are hard to control directly, we will use the function K_ν as a proxy for H_ν. The following analogue of Lemma 5.3.7 quantifies this relationship, showing that if $K_\nu(\boldsymbol{z})$ is bounded in Frobenius norm, $H_\nu(\boldsymbol{z})$ is bounded at the same scale.

Lemma C.1.6. *Suppose $\gamma \gg m^{1/2}\xi^{1/2}$. If $\boldsymbol{z} = (z_1, \dots, z_t)$ is a sequence of unit vectors satisfying $t \leqslant n$ and $\|K_\nu(\boldsymbol{z})\|_F \leqslant \left(n^{1/2} \cdot 2^{j_\nu/2} d_\nu^{3/2}/d\right)\gamma$, and the number of $s \in [t]$ for which $\nu(z_s) = \nu$ is at most $n \cdot (d_\nu/2^{j_\nu}) \cdot m\xi$, then $\|H_\nu(\boldsymbol{z})\|_F \leqslant C\left(n^{1/2} \cdot 2^{j_\nu/2} d_\nu^{3/2}/d\right)\gamma$ for some absolute constant $C > 0$.*

This lemma is a "deterministic" statement about a sequence of vectors. We will prove this in Subsection C.1.3 using the same bootstrap argument from earlier.

Lemma C.1.6 requires that for every ν, the number of POVM elements supported on the coordinates $\mathcal{B}_\nu$ is not much greater than its expectation, which we show holds with high probability:

Lemma C.1.7. *With probability $1 - o(1)$ over $\boldsymbol{x} \sim p_0$, for all $\nu \in [m]$ there are at most $n \cdot (d_\nu / 2^{j_\nu}) \cdot m\xi$ indices $s \in [n]$ for which $\nu(x_s) = \nu$.*

Proof. Take any POVM $\{\omega_x d \cdot xx^\top\}$ where for every x there is some ν for which $x \in \mathbb{S}_\nu$. Now fix $\nu \in [m]$ and note that the probability of observing x for which $\nu(x) = \nu$ upon measuring a copy of ρ under the null hypothesis is

$$\sum_{x:\nu(x)=x} \omega_x x^\dagger A_\nu x \leqslant d \cdot 2^{-j_\nu+1} \cdot \sum_{x:\nu(x)=x} \omega_x \|x\|^2 = 2^{-j_\nu+1} d_\nu, \tag{C.12}$$

where in the last step we used (C.3). The lemma follows by Markov and a union bound over $\nu \in [m]$. □

Finally, Lemma C.1.6 also requires a bound on $K_\nu(\boldsymbol{z})$. The following analogue of Lemma 5.3.8 bounds $K_\nu(\boldsymbol{z})$ in Frobenius norm uniformly over all prefixes $\boldsymbol{z}$ of $\boldsymbol{x}$. We will prove this lemma in Subsection C.1.4.

Lemma C.1.8. *If $\boldsymbol{x} \sim p_0$, then $\mathbb{E} * \sup_{1 \leqslant t \leqslant n} \|K_\nu(\boldsymbol{x}_{\leqslant t})\|_F^2 \lesssim n \cdot 2^{j_\nu} d_\nu^3 / d^2$.*

We will now prove Proposition C.1.4 assuming Lemmas C.1.6 and C.1.8. Let $\boldsymbol{x} \sim p_0$. For $1 \leqslant t \leqslant n$, define the filtration $\mathcal{F}_t = \sigma(\boldsymbol{x}_{\leqslant t})$ and the sequences

$$H_{\nu,t} = H_\nu(\boldsymbol{x}_{\leqslant t}), \qquad K_{\nu,t} = K_\nu(\boldsymbol{x}_{\leqslant t}), \qquad \Phi_{\nu,t} = L_\nu(\boldsymbol{x}_{\leqslant t}), \qquad \Phi_t = L(\boldsymbol{x}_{\leqslant t}). \tag{C.13}$$

Consider the times

$$\tau_\nu = \{\infty\} \cup \inf \Bigg\{ t : \|K_{\nu,t}\|_F > (n^{1/2} \cdot 2^{j_\nu/2} d_\nu^{3/2}/d)\alpha \qquad \text{or}$$
$$|s \in [t] : \nu(x_s) = \nu| > n \cdot (d_\nu / 2^{j_\nu}) \cdot m\xi \qquad \text{or} \qquad |\Phi_{\nu,t} - 1| > n \cdot \frac{\varepsilon^2 2^{j_\nu}}{d_\nu^{1/2} d^2} \beta \Bigg\} \tag{C.14}$$

which are clearly stopping times with respect to $\mathcal{F}_t$. Also define the stopped sequences $\Psi_{\nu,t} = \Phi_{\nu,t\wedge\tau_\nu}$.

Claim C.1.9. With probability $1 - o(1)$, $\|K_{\nu,t}\|_F \leqslant n^{1/2} \cdot 2^{j_\nu/2} d_\nu^{3/2}/d$ for all $t \in [n]$ and all $\nu \in [m]$.

Proof. By Lemma C.1.8,

$$\mathbf{Pr} * \sup_{1\leqslant t\leqslant n} \|K_{\nu,t}\|_F > n^{1/2} \cdot 2^{j_\nu/2} d_\nu^{3/2}/d \leqslant \frac{\mathbb{E} * \sup_{1\leqslant t\leqslant n} \|K_{\nu,t}\|_F^2}{(n \cdot 2^{j_\nu} d_\nu^3/d^2)\alpha^2} \lesssim \alpha^{-2} = o(1/m). \quad \square$$

The claim follows by a union bound over ν.

Claim C.1.10. With probability $1 - o(1)$, $|\Psi_{n,\nu} - 1| \leqslant n \cdot \frac{\varepsilon_\nu^2 2^{j_\nu}}{d_\nu^{1/2} d^2} \beta$ for all $\nu \in [m]$.

Proof. Note that $\Psi_{\nu,t}$ is a multiplicative martingale: if $\tau \leqslant t-1$ then certainly $\mathbb{E} \frac{\Psi_{\nu,t}}{\Psi_{\nu,t-1}} | \mathcal{F}_{t-1} = 1$, and if $\tau > t-1$, (C.11) implies

$$\mathbb{E} * \frac{\Psi_{\nu,t}}{\Psi_{\nu,t-1}} | \mathcal{F}_{t-1} = 1 + \frac{2\varepsilon^2}{d_\nu^2} \mathbb{E} * \mathbf{1}\nu(x_t) = \nu \cdot \frac{x_t^\dagger H_{\nu,t-1} x_t}{x_t^\dagger A_\nu x_t} | \mathcal{F}_{t-1} = 1,$$

using that

$$\begin{aligned} \mathbb{E} * \mathbf{1}\nu(x_t) = \nu \cdot \frac{x_t^\dagger H_{\nu,t-1} x_t}{x_t^\dagger A_\nu x_t} | \mathcal{F}_{t-1} &= \sum_{x_t : \nu(x_t) = \nu} \omega_{x_t} (x_t^\dagger H_{\nu,t-1} x_t) \\ &= \left\langle H_{\nu,t-1}, \sum_{x_t : \nu(x_t) = \nu} \omega_{x_t} x_t x_t^\dagger \right\rangle = \langle H_{\nu,t-1}, I_{d_\nu}/d \rangle = 0. \end{aligned} \quad \text{(C.15)}$$

We next bound the quadratic increment $\mathbb{E} (\frac{\Psi_{\nu,t}}{\Psi_{\nu,t-1}})^2 | \mathcal{F}_{t-1}$. If $\tau \leqslant t-1$ this is 1, and otherwise it is given by

$$1 + \frac{4\varepsilon^2}{d_\nu^2} \mathbb{E} * \mathbf{1}\nu(x_t) = \nu \cdot \frac{x_t^\dagger H_{\nu,t-1} x_t}{x_t^\dagger A_\nu x_t} | \mathcal{F}_{t-1} + \frac{4\varepsilon^4}{d_\nu^4} \mathbb{E} * \mathbf{1}\nu(x_t) = \nu \cdot \frac{(x_t^\dagger H_{\nu,t-1} x_t)^2}{(x_t^\dagger A_\nu x_t)^2} | \mathcal{F}_{t-1}. \quad \text{(C.16)}$$

The first expectation is zero by (C.15). To bound the remaining expectation, note that for any unit vector $x \in \mathbb{S}_\nu$,

$$x^\dagger A_\nu x \geqslant d \cdot 2^{-j_\nu}. \tag{C.17}$$

So,

$$\begin{aligned}
\mathbb{E} * \mathbf{1}\nu(x_t) = \nu \cdot \frac{(x_t^\dagger H_{\nu,t-1} x_t)^2}{(x_t^\dagger A_\nu x_t)^2} | \mathcal{F}_{t-1} \leqslant (2^{j_\nu}/d)\, \mathbb{E} * \mathbf{1}\nu(x_t) &= \nu \cdot \frac{(x_t^\dagger H_{\nu,t-1} x_t)^2}{x_t^\dagger A_\nu x_t} | \mathcal{F}_{t-1} \\
&= (2^{j_\nu}/d) \sum_{x_t : \nu(x_t) = \nu} \omega_{x_t} x_t^\dagger H_{\nu,t-1}(x_t x_t^\dagger) H_{\nu,t-1} x_t \qquad \text{(C.18)} \\
&\leqslant (2^{j_\nu}/d) \sum_{x_t : \nu(x_t) = \nu} \omega_{x_t} x_t^\dagger H_{\nu,t-1}^2 x_t \qquad \text{(C.19)} \\
&= (2^{j_\nu}/d) \left\langle H_{\nu,t-1}^2, \sum_{x_t : \nu(x_t) = \nu} \omega_{x_t} x_t x_t^\dagger \right\rangle \qquad \text{(C.20)} \\
&= (2^{j_\nu}/d) \langle H_{\nu,t-1}^2, I_{d_\nu}/d \rangle = \frac{2^{j_\nu}}{d^2} \| H_{\nu,t-1} \|_F^2. \qquad \text{(C.21)}
\end{aligned}$$

Moreover, since $\tau > t-1$, $\|K_{\nu,t-1}\|_F \leqslant (n^{1/2} \cdot 2^{j_\nu/2} d_\nu^{3/2}/d)\alpha$ and Lemma C.1.6 implies $\|H_{\nu,t-1}\|_F \leqslant (Cn^{1/2} \cdot 2^{j_\nu/2} d_\nu^{3/2}/d)\alpha$. Thus,

$$\mathbb{E} * \left(\frac{\Psi_{\nu,t}}{\Psi_{\nu,t-1}} \right)^2 | \mathcal{F}_{t-1} \leqslant 1 + \frac{4\varepsilon_\nu^4 2^{j_\nu}}{d^2 d_\nu^4} \| H_{\nu,t-1} \|_F^2 \leqslant 1 + \frac{36\varepsilon_\nu^4 2^{2j_\nu} \cdot n}{d_\nu d^4} \alpha^2.$$

So, for all $1 \leqslant t \leqslant n$,

$$\mathbb{E}\, \Psi_{\nu,t}^2 = \mathbb{E} * \mathbb{E} * \left(\frac{\Psi_{\nu,t}}{\Psi_{\nu,t-1}} \right)^2 | \mathcal{F}_{t-1} \Psi_{\nu,t-1}^2 \leqslant \left(1 + \frac{36\varepsilon_\nu^4 2^{2j_\nu} \cdot n}{d_\nu d^4} \alpha^2 \right) \mathbb{E}\, \Psi_{\nu,t-1}^2,$$

and therefore

$$\mathbb{E}\, \Psi_{\nu,t}^2 \leqslant \left(1 + \frac{36\varepsilon_\nu^4 2^{2j_\nu} \cdot n}{d_\nu d^4} \alpha^2 \right)^n \leqslant \exp\left(\frac{36\varepsilon_\nu^4 2^{2j_\nu} \cdot n^2}{d_\nu d^4} \alpha^2 \right) \leqslant 2$$

since $\frac{\varepsilon_\nu^4 2^{2j_\nu} \cdot n^2}{d_\nu d^4} \alpha^2 \ll 1$ by assumption.

Moreover,

$$\begin{aligned}\mathbb{E}\,(\Psi_{\nu,t}-1)^2 &= \mathbb{E}*\mathbb{E}*\left(\frac{\Psi_{\nu,t}}{\Psi_{\nu,t-1}}\right)^2|\mathcal{F}_{t-1}\Psi_{\nu,t-1}^2 - 2\,\mathbb{E}*\frac{\Psi_{\nu,t}}{\Psi_{\nu,t-1}}|\mathcal{F}_{t-1}\Psi_{\nu,t-1}+1\\ &\leqslant \frac{36\varepsilon_\nu^4 2^{2j_\nu}\cdot n}{d_\nu d^4}\alpha^2\cdot\mathbb{E}\,\Psi_{\nu,t-1}^2+\mathbb{E}\,(\Psi_{\nu,t-1}-1)^2\\ &\leqslant \frac{72\varepsilon_\nu^4 2^{2j_\nu}\cdot n^2}{d_\nu d^4}\alpha^2+\mathbb{E}\,(\Psi_{\nu,t-1}-1)^2,\end{aligned}$$

so by induction

$$\mathbb{E}\,(\Psi_{\nu,n}-1)^2 \leqslant \frac{72\varepsilon_\nu^4 2^{2j_\nu}\cdot n^2}{d_\nu d^4}\alpha^2.$$

Thus

$$\mathbf{Pr}*|\Psi_{\nu,n}-1|>n\cdot\left(\frac{72\varepsilon_\nu^4 2^{2j_\nu}}{d_\nu d^4}\beta^2\right)^{1/2}\leqslant\frac{72\alpha^2}{\beta^2}=o(1).$$

Therefore, $|\Psi_{\nu,n}-1|\leqslant n\cdot\frac{\varepsilon_\nu^2 2^{j_\nu}}{d_\nu^{1/2}d^2}\beta$ with probability $1-o(1)$. □

Claim C.1.11. If $\|K_{\nu,n}\|_F\leqslant(n^{1/2}\cdot 2^{j_\nu/2}d_\nu^{3/2}/d)\alpha$, then $L_\nu(\boldsymbol{x},\boldsymbol{x})\ll e^{\sqrt{d_\nu}}$.

Proof. Using the elementary inequality $e^z\leqslant 1+z$, we can upper bound $L_\nu(\boldsymbol{x},\boldsymbol{x})$ by

$$L_\nu(\boldsymbol{x},\boldsymbol{x})\leqslant\mathop{\mathbb{E}}_{M\sim\mathrm{GOE}^*(d_\nu)}\left[\exp\left(\left\langle 2\varepsilon_\nu M,\sum_{i\in[n]:\nu(x_i)=\nu}\frac{x_ix_i^\dagger}{x_i^\dagger A_\nu x_i}\right\rangle\right)\right]=\mathbb{E}[M]*\exp\left(\frac{2\varepsilon_\nu}{d_\nu}\langle M,K_{\nu,n}\rangle\right),\tag{C.22}$$

where in the second step we used that $\mathrm{tr}(M)=0$. As $M=G-\frac{\mathrm{tr}(G)}{d}I_{d_\nu}$ for $G\sim\mathrm{GOE}(d_\nu)$, we have that $\langle M,K_{\nu,n}\rangle=\langle G,K_{\nu,n}\rangle$ is distributed as a Gaussian with variance $\frac{2}{d_\nu}\|K_{\nu,n}\|_F^2\leqslant(2n2^{j_\nu}d_\nu^2/d^2)\alpha^2$. So we can bound (C.22) by

$$\mathbb{E}[g\sim\mathcal{N}(0,8\varepsilon_\nu^2 n\alpha^2/d)]\exp(g)=e^{8\varepsilon_\nu^2 n2^{j_\nu}\alpha^2/d^2}\ll e^{\sqrt{d_\nu}}\tag{C.23}$$

where the last step follows by (C.4). □

Proof of Proposition C.1.4. Define the event

$$S=\left\{\sup_{1\leqslant t\leqslant n}\|K_{\nu,t}\|_F\leqslant\left(n^{1/2}\cdot 2^{j_\nu/2}d_\nu^{3/2}/d\right)\alpha\text{ and }|\Psi_{\nu,n}-1|\leqslant n\cdot\frac{\varepsilon_\nu^2 2^{j_\nu}}{d_\nu^{1/2}d^2}\beta\ \ \forall\,\nu\in[m]\right\}.$$

By Claims C.1.9 and C.1.10, $\mathbf{Pr}[p_0]S = 1 - o(1)$. We will show that if S holds, then $\tau = \infty$. Indeed, if $\tau = t < \infty$, then there exists ν such that either $\|K_{\nu,t}\|_F > (n^{1/2} \cdot 2^{j_\nu/2} d_\nu^{3/2}/d)\alpha$ or $|\Phi_{\nu,t} - 1| > n \cdot \frac{\varepsilon_\nu^2 2^{j_\nu}}{d_\nu^{1/2} d^2}\beta$ holds. Since $\Psi_{\nu,n} = \Phi_{\nu,t}$, this contradicts S.

So, $\tau = \infty$ on S. This implies that for all ν, $|L_\nu(\boldsymbol{x}) - 1| = |\Phi_{\nu,n} - 1| \leqslant n \cdot \frac{\varepsilon_\nu^2 2^{j_\nu}}{d_\nu^{1/2} d^2}\beta = o(1/m)$. Moreover $\|K_{\nu,n}\|_F \leqslant (n^{1/2} \cdot 2^{j_\nu/2} d_\nu^{3/2}/d)\alpha$ for all ν, so by Claim C.1.11 we have $L_\nu(\boldsymbol{x}, \boldsymbol{x}) \ll e^{\sqrt{d}_\nu}$ for all ν. □

C.1.3 Bounding H in Frobenius Norm by Bootstrapping

In this subsection, we prove Lemma C.1.6. Throughout this subsection, fix some $\nu \in [m]$. To ease notation, we will drop subscripts and refer to K_ν and H_ν simply as K and H. Let $\boldsymbol{z} = (z_1, \ldots, z_t)$ be a sequence of unit vectors satisfying $t \leqslant n$ and

$$\|K(\boldsymbol{z})\|_F \leqslant \left(n^{1/2} \cdot 2^{j_\nu/2} d_\nu^{3/2}/d\right)\gamma \tag{C.24}$$

for some $\gamma \gg m^{1/2}\xi^{1/2}$. Let $J \subseteq [t]$ denote the set of $s \in [t]$ for which $\nu(z_s) = \nu$. Suppose that

$$|J| \leqslant n \cdot (d_\nu/2^{j_\nu}) \cdot m\xi \tag{C.25}$$

as in Lemma C.1.7.

The following lemma bounds a variant of $K(\boldsymbol{z})$ where we multiply each summand by an adversarial $b_i \in [-1, 1]$. This will be used to control the discrepancy $H(\boldsymbol{z}) - K(\boldsymbol{z})$ in the bootstrapping argument.

Lemma C.1.12. *Uniformly over $b_1, \ldots, b_t \in [-1, 1]$, we have*

$$\|\sum_{i \in J} b_i \frac{d_\nu z_i z_i^\dagger - I_{d_\nu}}{z_i^\dagger A_\nu z_i}\|_F \lesssim n d_\nu^{1/2} \cdot m\xi + (n^{1/2} \cdot 2^{j_\nu/2} d_\nu^{3/2}/d)\gamma$$

Proof. For any choice of $b_1, \dots, b_t$,

$$\begin{aligned}\|\sum_{i\in J} b_i \frac{d_\nu z_i z_i^\dagger - I_{d_\nu}}{z_i^\dagger A_\nu z_i}\|_F &\leqslant \|\sum_{i\in J} b_i \frac{d_\nu z_i z_i^\dagger}{z_i^\dagger A_\nu z_i}\|_F + \|\sum_{i\in J} b_i \frac{I_{d_\nu}}{z_i^\dagger A z_i}\|_F \\ &\leqslant \|\sum_{i\in J} \frac{d_\nu z_i z_i^\dagger}{z_i^\dagger A_\nu z_i}\|_F + \|\sum_{i\in J} \frac{I_{d_\nu}}{z_i^\dagger A_\nu z_i}\|_F \\ &\leqslant \|K(\boldsymbol{z})\|_F + 2\|\sum_{i\in J} \frac{I_{d_\nu}}{z_i^\dagger A_\nu z_i}\|_F.\end{aligned}$$

The second inequality holds because the matrices $d_\nu z_i z_i^\dagger$ and I_{d_ν} are both psd. Using (C.17) and the assume bound on $|J|$ in (C.25), we have

$$\|\sum_{i\in J} \frac{I_{d_\nu}}{z_i^\dagger A_\nu z_i}\|_F \leqslant (2^{j_\nu}/d)|J| d_\nu^{1/2} \lesssim n d_\nu^{1/2} \cdot m\xi.$$

The result follows by our assumed bound on $\|K(\boldsymbol{z})\|_F$. □

For $S \subseteq J$, let $\boldsymbol{z}_S = (z_i)_{i\in S}$. Further, let

$$H_S = \sum_{i\in S} \frac{d_\nu z_i z_i^\dagger - I_{d_\nu}}{z_i^\dagger A_\nu z_i} \cdot \frac{L_\nu(\boldsymbol{z}_{S\setminus\{i\}})}{L_\nu(\boldsymbol{z}_S)} \qquad \text{and} \qquad K_S = \sum_{i\in S} \frac{d_\nu z_i z_i^\dagger - I_{d_\nu}}{z_i^\dagger A_\nu z_i}.$$

The following lemma gives a preliminary bound on $\|H_S\|_F$. In the proof of Lemma C.1.6, we will use this bound to control $\|H_S\|_F$ for $|S| = t - O(\log n)$, followed by the bootstrap argument over $O(\log n)$ recursive rounds to contract the bound to $O((2^{j_\nu}/d)^{1/2} d_\nu n^{1/2}\gamma)$.

Lemma C.1.13. *There exists an absolute constant C such that for all $S \subseteq [t]$, $\|H_S\|_F \leqslant C(n d_\nu^{1/2} \cdot m\xi + (n^{1/2} \cdot 2^{j_\nu/2} d_\nu^{3/2}/d)\gamma)$.*

Proof. We take C to be twice the constant hidden by the $\lesssim$ in Lemma C.1.12. Note that for any fixed $\overline{M}_\nu \in U_\nu$, for the U_ν given by Lemma C.1.1, and any unit vector $z \in \mathbb{S}_\nu$,

$$\varepsilon \left| \frac{z^\dagger \overline{M}_\nu z}{z^\dagger A_\nu z} \right| \leqslant \frac{1}{12} \cdot \frac{3}{1/2} = \frac{1}{2},$$

so $1 + \varepsilon \frac{z^\dagger \overline{M}_\nu z}{z^\dagger A_\nu z} \in [1/2, 3/2]$. Thus, for all i, $L_\nu(\boldsymbol{z}_S)/L_\nu(\boldsymbol{z}_{S\setminus\{i\}}) \in [1/2, 3/2]$, which implies

$$\frac{L(\boldsymbol{z}_{S\setminus\{i\}})}{L(\boldsymbol{z}_S)} \in [2/3, 2]. \tag{C.26}$$

Lemma C.1.12 gives

$$\frac{1}{2}\|H_S\|_F \leqslant \frac{1}{2}C(nd_\nu^{1/2} \cdot m\xi + (n^{1/2} \cdot 2^{j_\nu/2} d_\nu^{3/2}/d)\gamma).$$

as desired.

□

Proof of Lemma C.1.6. Let $\varepsilon^* \triangleq \left(n^{1/2} \cdot 2^{j_\nu/2} d_\nu^{3/2}/d\right)\gamma$ and $\varepsilon' \triangleq Cnd_\nu^{1/2} \cdot m\xi$. If $\varepsilon^* \geqslant \varepsilon'$, then we are already done by Lemma C.1.13. Otherwise, suppose $\varepsilon^* < \varepsilon'$. and let $D = \log(\varepsilon'/\varepsilon^*)$. If $t < D$, then by equations (C.17) and (C.26),

$$\|H(\boldsymbol{z})\|_F \leqslant \sum_{i=1}^{t} \frac{\|d_\nu z_i z_i^\dagger - I_{d_\nu}\|_F}{z_i^\dagger A z_i} \cdot \frac{L(\boldsymbol{z}_{\sim i})}{L(\boldsymbol{z})} \leqslant 2^{j_\nu+1} D d_\nu/d.$$

But note that $2^{j_\nu+1} D d_\nu/d \ll \varepsilon^*$ provided that $n \gg\gg 2^{j_\nu}/(d_\nu\gamma^2)$. By (C.2), $2^{j_\nu}/d_\nu \leqslant \log(d/\varepsilon)/(2\varepsilon)$, so this holds by (C.5) and our choice of $\gamma \gg m^{1/2}\xi^{1/2} \geqslant \mathrm{polylog}(d/\varepsilon)$. So $\|H(\boldsymbol{z})\|_F \ll \varepsilon^*$ when $t < D$.

Now suppose $t \geqslant D$. By Lemma C.1.13 and the assumption that $\varepsilon^* \leqslant \varepsilon'$, $\|H_S\|_F \leqslant 2\varepsilon'$. We will prove by induction on $a \geqslant 0$ that if $S \subseteq J$ satisfies $|S\setminus J| = D - a$, then

$$\|H_S\|_F \leqslant \xi_a \triangleq 2\varepsilon^* + e^{-a} \cdot 2\varepsilon'.$$

The base case $a = 0$ clearly holds. For the inductive step, assume $a \geqslant 1$. By the inductive hypothesis and equations (C.11) and (C.17), for all $i \in S$

$$\left|\frac{L_\nu(\boldsymbol{z}_S)}{L_\nu(\boldsymbol{z}_{S\setminus\{i\}})} - 1\right| \leqslant \frac{2\varepsilon_\nu^2}{d^2} \cdot \|\frac{H_{S\setminus i}}{z_i^\dagger A_\nu z_i}\|_{\mathrm{op}} \leqslant \frac{(2^{j_\nu}/d) \cdot 2\varepsilon_\nu^2}{d^2}\|H_{S\setminus i}\|_F \leqslant \frac{(2^{j_\nu}/d) \cdot 2\varepsilon_\nu^2}{d^2}\xi_{a-1}.$$

Since this upper bound is $o(1)$ by (C.4) and the second part of (C.2), we also have

$$\left|\frac{L(\boldsymbol{z}_{S\setminus\{i\}})}{L(\boldsymbol{z}_S)} - 1\right| \leqslant \frac{(2^{j_\nu}/d)\cdot 3\varepsilon_\nu^2}{d^2}\xi_{a-1}.$$

Write $\frac{L_\nu(\boldsymbol{z}_{S\setminus\{i\}})}{L_\nu(\boldsymbol{z}_S)} - 1 = \frac{(2^{j_\nu}/d)\cdot 3\varepsilon_\nu^2}{d^2}\xi_{a-1}b_i$ for $b_i \in [-1,1]$. By Lemma C.1.12, there is a constant c such that

$$\begin{aligned}\|\sum_{i\in S}\frac{d_\nu z_i z_i^\dagger - I_{d_\nu}}{z_i^\dagger A_\nu z_i}\cdot\left(\frac{L_\nu(\boldsymbol{z}_{S\setminus\{i\}})}{L_\nu(\boldsymbol{z}_S)}-1\right)\|_F &= \frac{(2^{j_\nu}/d)\cdot 3\varepsilon_\nu^2}{d^2}\xi_{a-1}\|\sum_{i\in S}\frac{d_\nu z_i z_i^\dagger - I_{d_\nu}}{z_i^\dagger A_\nu z_i}\cdot b_i\|_F \\ &\leqslant \frac{(2^{j_\nu}/d)\cdot 3\varepsilon_\nu^2 n d_\nu^{1/2}\cdot m\xi}{d^2}\xi_{a-1} \leqslant e^{-1}\xi_{a-1},\end{aligned}$$

where in the last step we used that $n \ll \frac{d^2 d_\nu^{1/2}}{\varepsilon_\nu^2 2^{j_\nu}\cdot m\xi}$. By the triangle inequality, equation (C.17), and our choice of D,

$$\|K_S\|_F \leqslant \|K(\boldsymbol{z})\|_F + \sum_{i\in J\setminus S}\frac{\|d_\nu z_i z_i^\dagger - I_{d_\nu}\|_F}{z_i^\dagger A_\nu z_i} \leqslant \varepsilon^* + 2^{j_\nu}Dd_\nu/d \leqslant \frac{101}{100}\varepsilon^*.$$

Hence

$$\begin{aligned}\|H_S\|_F &\leqslant \|K_S\|_F + \|\sum_{i\in S}\frac{dz_i z_i^\dagger - I_d}{z_i^\dagger A z_i}\cdot\left(\frac{L(\boldsymbol{z}_{S\setminus\{i\}})}{L(\boldsymbol{z}_S)}-1\right)\|_F \\ &\leqslant \frac{101}{100}\varepsilon^* + e^{-1}\xi_{a-1} \leqslant \xi_a,\end{aligned}$$

as $\frac{101}{100} + 2e^{-1} \leqslant 2$. This completes the induction. Finally,

$$\|H(\boldsymbol{z})\|_F = \|H_J\|_F \leqslant 2\varepsilon^* + e^{-D}2\varepsilon' = 4\varepsilon^*. \qquad \square$$

C.1.4 Uniform Frobenius Bound on the $K_\nu(\boldsymbol{x}_{\leqslant t})$ Matrix Martingale

In this subsection, we will prove Lemma C.1.8. Fix any $\nu \in [m]$, let $\boldsymbol{x} \sim p_0$, recall that $K_{\nu,t} = K_\nu(\boldsymbol{x}_{\leqslant t})$. To ease notation, we will drop the subscript ν and refer to this

as K_t. Also define $X = \sup_{1 \leqslant t \leqslant n} \|K_t\|_F$.

Lemma C.1.14. *We have that* $\mathbb{E}\, X^2 \leqslant 4\, \mathbb{E}\, \|K_{\nu,n}\|_F^2$

Proof. Analogous to Lemma 5.3.14. □

Lemma C.1.15. *We have that* $\mathbb{E} * \|K_n\|_F^2 \lesssim 2^{j_\nu} d_\nu^2 n/d$.

Proof. We can expand

$$\mathbb{E}\, \|K_n\|_F^2 = \sum_{i=1}^n \mathbb{E} * \mathbf{1}\nu(x_i) = \nu \cdot \| \frac{d_\nu x_i x_i^\dagger - I_{d_\nu}}{x_i^\dagger A_\nu x_i} \|_F^2 + 2 \sum_{1 \leqslant i < j \leqslant n} \mathbb{E} * \left\langle \mathbf{1}\nu(x_i) = \nu, \nu(x_j) = \nu \cdot \frac{d_\nu x_i x_i^\dagger - I_{d_\nu}}{x_i^\dagger A_\nu x_i}, \frac{d_\nu x_j x_j^\dagger - I_{d_\nu}}{x_j^\dagger A_\nu x_j} \right\rangle. \quad \text{(C.27)}$$

By (C.3),

$$\mathbb{E} * \mathbf{1}\nu(x_j) = \nu \cdot \frac{d_\nu x_j x_j^\dagger - I_{d_\nu}}{x_j^\dagger A_\nu x_j} | \mathcal{F}_{j-1} = \sum_{x_j : \nu(x_j) = \nu} \omega_{x_j} (d_\nu x_j x_j^\dagger - I_{d_\nu}) = 0,$$

so for any $i < j$ we have

$$\begin{aligned} &\mathbb{E} * \mathbf{1}\nu(x_i) = \nu, \nu(x_j) = \nu \cdot \left\langle \frac{d_\nu x_i x_i^\dagger - I_{d_\nu}}{x_i^\dagger A_\nu x_i}, \frac{d_\nu x_j x_j^\dagger - I_{d_\nu}}{x_j^\dagger A_\nu x_j} \right\rangle \\ &= \mathbb{E} * \mathbf{1}\nu(x_i) = \nu \left\langle \frac{d_\nu x_i x_i^\dagger - I_{d_\nu}}{x_i^\dagger A_\nu x_i}, \mathbb{E} * \mathbf{1}\nu(x_j) = \nu \cdot \frac{d_\nu x_j x_j^\dagger - I_{d_\nu}}{x_j^\dagger A_\nu x_j} | \mathcal{F}_{j-1} \right\rangle = 0. \quad \text{(C.28)} \end{aligned}$$

The other expectation in (C.27) can be bounded by (recalling (C.17))

$$\mathbb{E} * \mathbf{1}\nu(x_i) = \nu \cdot \| \frac{d_\nu x_i x_i^\dagger - I_{d_\nu}}{x_i^\dagger A_\nu x_i} \|_F^2 \leqslant (2^{j_\nu}/d)\, \mathbb{E} * \mathbf{1}\nu(x_i) = \nu \cdot \frac{\langle d_\nu x_i x_i^\dagger - I_{d_\nu}, d_\nu x_i x_i^\dagger - I_{d_\nu} \rangle}{x_i^\dagger A_\nu x_i} \quad \text{(C.29)}$$

$$= (2^{j_\nu}/d) \cdot d_\nu (d_\nu - 1)\, \mathbb{E} * \frac{\mathbf{1}\nu(x_i) = \nu}{x_i^\dagger A_\nu x_i} \quad \text{(C.30)}$$

$$\leqslant (2^{j_\nu}/d) \cdot d_\nu(d_\nu - 1) \cdot \sum_{x:\nu(x)=x} \omega_x \leqslant 2^{j_\nu} d_\nu^3/d^2. \tag{C.31}$$

Therefore $\mathbb{E} * \|K_n\|_F^2 \leqslant n \cdot 2^{j_\nu} d_\nu^3/d^2$. □

Proof of Lemma C.1.8. Follows immediately from Lemmas C.1.14 and C.1.15. □

C.2 Refined Bounds for State Certification

In this section we use the lower bound instance from Appendix C.1 to give a refined version of the analysis in Section 5.5 and prove Theorem 5.5.1. The steps in this section are essentially already present in [65] (see Sections 5.1, 5.2.2, and 5.5 therein), but we include them for the sake of completeness.

C.2.1 Bucketing and Mass Removal

We will use the following bucketing scheme from [65, Definition 5.2].

For $j \in \mathbb{Z}_{\geqslant 0}$, let S_j denote the set of indices $i \in [d]$ for which $2^{-j-1} < \sigma_i \leqslant 2^{-j}$, and define $d_j \triangleq |S_j|$. Let $\mathcal{J}$ denote the set of j for which $S_j \neq \emptyset$. We will refer to $j \in \mathcal{J}$ as *buckets*. Given $i \in [d]$, let $j(i)$ denote the index of the bucket for which $i \in S_j$.

Our bounds are based on the following modification of σ given by removing a small fraction of its entries:

Definition C.2.1. If b is the largest number for which the b smallest entries of σ sum to at most ε, define $S^1_{\mathsf{light}} \subseteq [d]$ to be the indices of these b smallest entries. Let S^2_{light} denote the set of $i \in [d]$ for which $\sum_{i' \in S_{j(i)}} \sigma_{i'} \leqslant 2\varepsilon/\log(d/\varepsilon)$. Define $S_{\mathsf{light}} \triangleq S^1_{\mathsf{light}} \cup S^2_{\mathsf{light}}$.

Without loss of generality, assume that all σ_i are sorted in increasing order based on $\sigma_i/d^2_{j(i)}$. Recall the constant a^* from Lemma C.1.1 and Theorem C.1.2. Let $d' \leqslant d$

denote the largest index for which $\sum_{i \notin S_{\mathsf{light}}: i \leqslant d'} \sigma_i \leqslant C_{a^*}\varepsilon$ for a constant C_{a^*} sufficiently lage depending on a^*. Let $S_{\mathsf{tail}} \triangleq \{i : i \notin S_{\mathsf{light}}, i \leqslant d'\}$.

Let m denote the number of buckets $j \in \mathcal{J}$ for which S_j and S_{light} are disjoint. Let $S_{\mathsf{few}} \subseteq [d]$ denote the set of i belonging to a bucket of size less than a^*, and let $S_{\mathsf{many}} \subseteq [d]$ denote the set of i belonging to a bucket of size at least a^*.

Let σ' denote the matrix given by zeroing out the entries indexed by $S_{\mathsf{tail}} \cup S_{\mathsf{light}}$. Let σ'' denote the matrix by further zeroing out the largest entry of σ'. Let σ^* denote the density matrix $\sigma'/\operatorname{tr}(\sigma^*)$.

Lastly, define $\mathcal{J}^*$ to be the set of $j \in \mathcal{J}$ for which S_j has nonempty intersection with $S_{\mathsf{many}} \backslash S_{\mathsf{light}}$.

Fact C.2.2. *We have $m \leqslant O(\log(d/\varepsilon))$, that is, there are at most $O(\log(d/\varepsilon))$ indices $j \in \mathcal{J}$ for which S_j and S_{light} are disjoint. Furthermore, the total mass of σ in $S_{\mathsf{light}} \cup S_{\mathsf{tail}}$ is $O(\varepsilon)$.*

Proof. This is a slight modification of [65, Fact 5.3]. By definition of S^1_{light}, the $(b+1)$-st smallest entry of σ is at least ε/d. There are thus at most $\log_2(d/\varepsilon)$ buckets containing $[d]\backslash S^1_{\mathsf{light}}$, which concludes the proof of the first part. The second part follows by construction. □

C.2.2 Tuning the Perturbations

The goal of this section will be to tune the perturbations $\{\varepsilon_j\}$ from the lower bound instance in Theorem C.1.2 in order to show the following:

Lemma C.2.3. *For $0 < \varepsilon < \widetilde{O}(1/\log\log(d))$, for any mixed state $\sigma \in \mathbb{C}^{d\times d}$, the copy complexity of state certification with respect to σ to error ε is at least $\Omega(1/\varepsilon) \vee \widetilde{\Omega}(\|\sigma''\|_{2/5}/(\varepsilon^2 \operatorname{polylog}(d/\varepsilon)))$.*

First we handle a minor corner case. Note that Theorem C.1.2 can only be applied to the buckets of σ which are of size at least a^*. We now verify that if the Schatten

2/5-norm of σ' is dominated by such buckets, then the $\widetilde{\Omega}(\|\sigma''\|_{2/5}/\varepsilon^2)$ lower bound follows from *classical* lower bounds.

Lemma C.2.4. *If* $\sum_{i\in S_{\mathsf{few}}\setminus(S_{\mathsf{tail}}\cup S_{\mathsf{light}})}\sigma_i^{2/5} \geqslant \frac{1}{2}\|\sigma''\|_{2/5}^{2/5}$, *then state certification with respect to* σ *using incoherent measurements has copy complexity at least* $\Omega(\|\sigma''\|_{2/5}/\varepsilon^2)$.

For this, we use the following instance-optimal lower bound for classical identity testing:

Theorem C.2.5 (Theorem 1.1 from [204])**.** *Given a known distribution* p *and samples from an unknown distribution* q, *any tester that can distinguish between* $q = p$ *and* $\|p-q\|_1 \geqslant \varepsilon$ *with probability* $2/3$ *must draw at least* $\Omega(1/\varepsilon)\vee\Omega(\|p_{-\varepsilon}^{-\max}\|_{2/3}/\varepsilon^2)$ *samples, where* $p_{-\varepsilon}^{-\max}$ *denotes the vector given by removing from* p *the largest element and the smallest elements summing up to at most* ε.

Note that this immediately implies a lower bound for state certification by considering only diagonal mixed states:

Corollary C.2.6. *State certification with respect to any known mixed state* σ *to error* ε *using incoherent measurements requires at least* $\Omega(1/\varepsilon)\vee\Omega(\|\sigma_{-\varepsilon}^{-\max}\|_{2/3}/\varepsilon^2)$ *samples, where* $\sigma_{-\varepsilon}^{-\max}$ *denotes the matrix given by projecting out from* σ *the largest eigenvalue and the smallest eigenvalues summing up to at most* ε.

Proof of Lemma C.2.4. This is a slight modification of [65, Lemma 5.12]. The idea is that if the hypothesis of the lemma holds, then the spectrum of σ is essentially dominated by eigenvalues in geometric progression, in which case there is no distinction between the 2/5- and 2/3-quasinorms and we can simply apply Corollary C.2.6.

Formally, Corollary C.2.6 implies a lower bound of $\Omega(\|\sigma_{\varepsilon}^{-\max}\|_{2/3}/\varepsilon^2)$. We would like to relate $\|\sigma_{\varepsilon}^{-\max}\|_{2/3}$ to

$$\left(\sum_{i\in S_{\mathsf{few}}\setminus(S_{\mathsf{tail}}\cup S_{\mathsf{light}})}\sigma_i^{2/3}\right)^{3/2} \geqslant a^{*-5/2}\cdot(1-2^{-2/5})^{5/2}\cdot\left(\sum_{i\in S_{\mathsf{few}}\setminus S_{\mathsf{tail}}}\sigma_i^{2/5}\right)^{5/2} \geqslant \Omega(\|\sigma''\|_{2/5}), \tag{C.32}$$

where the last step follows by the hypothesis of the lemma and Fact C.2.2.

Suppose that there is some i for which $d_{j(i)} \leqslant a^*$ and i is not among the indices removed in the definition of $\sigma_{-\varepsilon}^{-\max}$. Then we can lower bound $\|\sigma_{-\varepsilon}^{-\max}\|_{2/3}$ by σ_i, which is at least $a^{*-3/2}(1-2^{-2/3})^{3/2} = \Omega(1)$ times the left-hand side of (C.32).

On the other hand, suppose that all i for which $d_{j(i)} \leqslant a^*$ are removed in the definition of $\sigma_{-\varepsilon}^{-\max}$. As long as $\sigma_{-\varepsilon}^{-\max}$ has some nonzero entry, call it σ_{i^*}, then $\sigma_{i^*} \geqslant \max_{i \in S_{\mathsf{few}} \setminus (S_{\mathsf{tail}} \cup S_{\mathsf{light}})} \sigma_i$, so we can similarly guarantee that $\|\sigma_{-\varepsilon}^{-\max}\|_{2/3} \geqslant \sigma_{i^*}$ is at least $a^{*-3/2}(1-2^{-2/3})^{3/2} = \Omega(1)$ times the left-hand side of (C.32). Otherwise, we note that σ'' is zero as well, in which case we are also done. □

It remains to consider the primary case where the hypothesis of Lemma C.2.4 does not hold, which we can express as

$$\sum_{i \in S_{\mathsf{many}} \setminus (S_{\mathsf{light}} \cup S_{\mathsf{tail}})} \sigma_i^{2/5} > \frac{1}{2} \|\sigma''\|_{2/5}^{2/5}, \tag{C.33}$$

and this is the case where we will use Theorem C.1.2. Because Corollary C.2.6 already shows that the copy complexity is at least $\Omega(1/\varepsilon)$, we will assume henceforth that the lower bound in Theorem C.1.2 is at least $\Omega(1/\varepsilon)$.

First for every $i \in S_{\mathsf{many}} \setminus S_{\mathsf{light}}$, define the perturbations

$$\varepsilon_{j(i)} \triangleq d \cdot \left\{ 2^{-j(i)-1} \Big/ \left(12 + \Theta\left(\sqrt{\log(m)/d_{j(i)}} \right) \right) \right\} \wedge \left\{ \zeta 2^{-2/3(j(i)+1)} d_{j(i)}^{2/3} \right\} \tag{C.34}$$

for normalizing quantity ζ satisfying

$$\sum_{j \in \mathcal{J}^*} d_j \cdot \left\{ 2^{-j-1} \wedge \zeta 2^{-2/3(j+1)} d_j^{2/3} \right\} = \varepsilon. \tag{C.35}$$

Note that this choice ζ ensures that the trace distance between the two states under H_0 and H_1 in Theorem C.1.2 is $\Omega(\varepsilon)$.

The rest of the proof is devoted to analyzing what Theorem C.1.2 gives for this choice

of $\{\varepsilon_j\}$. The main step is to upper bound the normalizing quantity ζ.

Lemma C.2.7. $\zeta \leqslant O(\varepsilon) \cdot \left(\sum_{j \in \mathcal{J}^*} 2^{-2j/3} d_j^{5/3}\right)^{-1}$.

To prove this, we will need the following elementary fact.

Fact C.2.8. *Let $u_1 \leqslant \cdots \leqslant u_m$ be numbers for which there are at most ℓ elements in any interval $[2^{-j-1}, 2^{-j}]$. Let $v_1 \leqslant \cdots v_n$ and let $d_1, \ldots, d_n > 1$ be arbitrary integers. Let $w_1 \leqslant \cdots \leqslant w_{m+n}$ be these numbers $u_1, \ldots, u_m, v_1, \ldots, v_n$ in sorted order. For $i \in [m+n]$, define d_1^* to be 1 if w_i corresponds to some u_j, and d_j if w_i corresponds to some v_j.*

There is an absolute constant C_ℓ depending on ℓ such that the following holds. Let s be the largest index for which $\sum_{i=1}^{s} w_i d_i^* \leqslant C_\ell \varepsilon$. Let a, b be the largest indices for which u_a, v_b are present among $w_1, \ldots, w_s$ (if none exists, take it to be 0). Then either $b = n$ or $\sum_{i=1}^{b+1} v_i d_i > 2\varepsilon$.

Proof. This is Fact 5.16 from [65] with minor modifications. We may assume $s < m+n$ (otherwise obviously $b = n$). Assume to the contrary that $\sum_{i=1}^{b+1} v_i d_i \leqslant \varepsilon$. We proceed by casework based on whether $w_{s'+1} = u_{a+1}$ or $w_{s'+1} = v_{b+1}$.

If $w_{s'+1} = u_{a+1}$, then

$$C_\ell \varepsilon < \sum_{i=1}^{s+1} w_i d_i^* = \sum_{i=1}^{a+1} u_i + \sum_{i=1}^{b} v_i d_i \leqslant \sum_{i=1}^{a+1} v_{b+1} \cdot 2^{\lceil (1-i)/\ell \rceil} + \sum_{i=1}^{b} v_i d_i \leqslant O_\ell(1)\varepsilon + \sum_{i=1}^{b} v_i d_i, \tag{C.36}$$

where in the first step we used maximality of s, in the third step we used that $u_{a+1} \leqslant v_{b+1}$ and the assumption on $\{u_i\}$, and in the last step we used that $v_{b+1} \leqslant \sum_{i=1}^{b+1} v_i d_i \leqslant \varepsilon$. From this, if C_ℓ is sufficiently large, then we conclude that $\sum_{i=1}^{b} v_i d_i > 2\varepsilon$, a contradiction. The argument for $w_{s'+1} = v_{b+1}$ is analogous. □

Corollary C.2.9. *If* (C.33) *holds, then $S_{\mathsf{many}} \backslash (S_{\mathsf{light}} \cup S_{\mathsf{tail}})$ is nonempty, and there exists an absolute constant $c > 0$ such that for any $i \in S_{\mathsf{many}} \backslash (S_{\mathsf{light}} \cup S_{\mathsf{tail}})$ in some bucket j, $\zeta \cdot 2^{-2/3(j+1)} d_j^{2/3} \leqslant 2^{-j-1}$.*

Proof. The first part immediately follows from (C.33). For the second part, take some constant $c \geqslant 1$ to be optimized later and suppose to the contrary that for some $i^* \in S_{\mathsf{many}} \backslash (S_{\mathsf{light}} \cup S_{\mathsf{tail}})$, lying in some bucket j^*, we have $2^{-j^*-1} < \zeta \cdot 2^{-2/3(j^*+1)} d_{j^*}^{2/3}$, or equivalently $2^{-j^*-1}/d_{j^*}^2 < \zeta^3$. Because in the definition of S_{tail}, we sorted by $\sigma_i/d_{j(i)}^2$, we then also have that $2^{-j(i)-1}/d_{j(i)}^2 < \zeta^3$ for all $i \in S_{\mathsf{tail}}$, or equivalently, $2^{-j(i)-1} < \zeta \cdot 2^{-2/3(j+1)} d_{j(i)}^{2/3}$.

To induce a contradiction, we lower bound the sum on the left-hand side of (C.35) by the contribution from $j \in \mathcal{J}^*$ for which S_j contains an index i satisfying $i \leqslant i^*$. The above discussion implies that for such j, the corresponding summand in (C.35) is given by $d_j \cdot 2^{-j(i)-1}$. So the left-hand side of (C.35) is at least

$$\sum_{i \in S_{\mathsf{many}} \backslash S_{\mathsf{light}} : i \leqslant i^*} \sigma_i > \varepsilon, \tag{C.37}$$

where in the latter inequality we used Fact C.2.8 applied to the numbers $\ell \triangleq a^*$, $\{u_i\} \triangleq \{\sigma_i\}_{i \in S_{\mathsf{few}} \backslash S_{\mathsf{light}}}$, $\{v_i\} \triangleq \{\sigma_i/d_{j(i)}^2\}_{i \in S_{\mathsf{many}} \backslash S_{\mathsf{light}}}$, and $\{d_i\} \triangleq \{d_{j(i)}^2\}_{i \in S_{\mathsf{many}} \backslash S_{\mathsf{light}}}$, in light of our definition for S_{tail}. This contradicts (C.35). □

We are finally ready to upper bound the normalizing constant ζ.

Proof of Lemma C.2.7. By Corollary C.2.9 and (C.35),

$$\varepsilon \geqslant \Omega(\zeta) \cdot \sum_{j \in \mathcal{J}^*} d_j \cdot 2^{-2/3(j+1)} d_j^{2/3} \geqslant \Omega(\zeta) \sum_{j \in \mathcal{J}^*} 2^{-2j/3} d_j^{5/3}. \tag{C.38}$$

The claimed bound follows. □

We are now ready to complete the proof of Lemma C.2.3.

Proof of Lemma C.2.3. As discussed above, because of Lemma C.2.4 it suffices to consider the case where (C.33) holds. We will apply Theorem C.1.2 to the principal submatrix of σ indexed by the indices from buckets in $\mathcal{J}^*$. It suffices to show that

the copy complexity in that theorem, when specialized to ε_j from (C.34), is at least $\widetilde{\Omega}(\|\sigma''\|_{2/5}/(\varepsilon^2 \operatorname{polylog}(d/\varepsilon)))$. Note that we can apply Theorem C.1.2 to this submatrix because by our definition of S_{light}, the second part of (C.2) holds, by the first argument of each minimum in (C.34), the first part of (C.2) holds, and by the definition of S_{many}, d_j is sufficiently large for every j that appears in this submatrix. Note that our definition of m in Definition C.2.1 is the same as the parameter m in Theorem C.1.2. Recall from Fact C.2.2 that $m \leqslant O(\log(d/\varepsilon))$.

$$\varepsilon_j \triangleq d \cdot \left\{ 2^{-j-1} \Big/ \left(12 + \Theta\left(\sqrt{\log(m)/d_j} \right) \right) \right\} \wedge \left\{ \zeta 2^{-2/3(j+1)} d_j^{2/3} \right\} \tag{C.39}$$

First, let us rewrite the lower bound from that theorem as

$$\frac{1}{m} \min_{j \in \mathcal{J}^*} \frac{d_j^{1/2} d^2}{\varepsilon_j^2 2^j} \geqslant \frac{1}{m} \left(\sum_{j \in \mathcal{J}^*} \frac{\varepsilon_j^4 2^{2j}}{d_j d^4} \right)^{-1/2} \tag{C.40}$$

Substituting our choice of $\{\varepsilon_j\}$ from (C.34) and denoting $\alpha_j \triangleq 12 + \Theta(\sqrt{\log(m)/d_j})$, we get

$$\gtrsim \frac{1}{m} \left(\sum_{j \in \mathcal{J}^*} \frac{2^{-2j}}{\alpha_j^4 d_j} \wedge \zeta^4 2^{-2j/3} d_j^{5/3} \right)^{-1/2} \tag{C.41}$$

$$\geqslant \frac{1}{m} \left(\sum_{j \in \mathcal{J}^*} \alpha_j^{-1} \zeta^3 2^{-j} d_j \wedge \zeta^4 2^{-2j/3} d_j^{5/3} \right)^{-1/2} \tag{C.42}$$

$$\gtrsim \frac{\zeta^{-3/2}}{m} \left(\sum_{j \in \mathcal{J}^*} d_j \{ \alpha_j^{-1} 2^{-j-1} \wedge \zeta 2^{-2/3(j+1)} d_j^{2/3} \} \right)^{-1/2} \tag{C.43}$$

$$\gtrsim \zeta^{-3/2} \cdot \varepsilon^{-1/2}/m \gtrsim (\varepsilon^{-2}/m) \cdot \left(\sum_{j \in \mathcal{J}^*} 2^{-2j/3} d_j^{5/3} \right)^{3/2} \tag{C.44}$$

$$\geqslant \max_{j \in \mathcal{J}^*, i \in S_j} \sigma_i d_j^{5/2} / (\varepsilon^2 \log(d/\varepsilon)) \gtrsim \|\sigma''\|_{2/5} / (\varepsilon^2 \operatorname{polylog}(d/\varepsilon)), \tag{C.45}$$

where in the second step we used that the minimum of two nonnegative numbers increases if we replace one of them by a weighted geometric mean of the two, in the

fourth step we used (C.35), in the fifth step we used Lemma C.2.7, in the penultimate step we used Fact C.2.2, and in the last step we used (C.33). □

C.2.3 Putting Everything Together

Proof of Theorem 5.5.1. The proof will be given by modifying a few places in the proof in Section 5.5. We proceed by the same casework of whether or not $d_j = 1$ for all $j \in \mathcal{J}^*$ (note that our definition of $\mathcal{J}^*$ is slightly different from the one used in Section 5.5).

First by Fact C.2.2 we have that $\mathrm{tr}(\sigma') \geqslant 1 - O(\varepsilon) \geqslant \Omega(1)$, so by Fact 5.5.3 it suffices to lower bound the copy complexity by

$$\Omega\left(d_{\mathsf{eff}}\|\sigma'\|_{1/2}/(\varepsilon^2 \log^{\Theta(1)}(d/\varepsilon))\right). \tag{C.46}$$

Case 1. $d_j = 1$ for all $j \in \mathcal{J}^*$. Note that in this case,

$$\|\sigma'\|_{1/2}^{1/2} = \sum_{j \in \mathcal{J}^*} 2^{-j/2} = O(1) \tag{C.47}$$

and $\|\sigma^*\|_{1/2} = \Theta(\|\sigma'\|_{1/2})$. As $d_{\mathsf{eff}} = 1$, it thus suffices to show a lower bound of $\Omega(1/\varepsilon^2)$ in this case.

If additionally we have $|\mathcal{J}^*| = 1$, then for ε at most a sufficiently small constant, the maximum entry of σ is at least $3/4$, so we can apply Lemma 5.5.7 to obtain a lower bound of $\Omega(1/\varepsilon^2)$ as desired.

Otherwise, let j, j' be the two smallest bucket indices in $\mathcal{J}^*$, and let i, i' be the elements of the singleton sets $S_j, S_{j'}$. If $\varepsilon \leqslant c2^{-j/2-j'/2-1}$ for sufficiently small absolute constant $c > 0$, we can invoke [65, Lemma A.4] to conclude a lower bound of $\Omega(1/\varepsilon^2)$.

Otherwise, suppose $\varepsilon > c2^{-j/2-j'/2-1}$. Because $2^{-j} > 2^{-j'}$, we know that $2^{-j} \leqslant O(\varepsilon)$. In particular, consider the state σ^{**} given by zeroing out $\sigma_{i'}$ from σ' and normalizing. For this matrix, $d_{\mathsf{eff}} = 1$ and $\|\sigma^{**}\|_{1/2} = O(1)$. Furthermore, because ε is smaller than

some absolute constant, we conclude that the nonzero entry of σ^{**} is at least $3/4$, so we can again apply Lemma 5.5.7 to conclude a lower bound $\Omega(1/\varepsilon^2)$.

Case 2. $d_j > 1$ for some $j \in \mathcal{J}^*$. In this case, let $j_1 \triangleq \arg\max_{j\in\mathcal{J}^*} d_j$ and $j_2 \triangleq \arg\max_{j\in\mathcal{J}^*} d_j^2 2^{-j}$. Note that $d2^{-j_1} \geqslant d_{j_1}2^{-j_1} \gtrsim \varepsilon/\log(d/\varepsilon)$, so

$$j_1 \leqslant O(\log(d/\varepsilon)). \tag{C.48}$$

If $\varepsilon \leqslant cd_{j_2}2^{-j_1/2-j_2/2-1}/j_1$ for sufficiently small constant $c > 0$, then we can invoke the lower bound instance from Section 5.4. The proof in this case is identical to the corresponding part of the proof in Section 5.5.

It remains to consider the case of

$$\varepsilon > cd_{j_2}2^{-j_1/2-j_2/2-1}/j_1. \tag{C.49}$$

We would like to use the lower bound from Lemma C.2.3. We would first like to relate $\|\sigma'\|_{2/5}$ to $\|\sigma''\|_{2/5}$ (recall that the difference is that σ'' is defined by removing the largest entry from σ'.

Lemma C.2.10. *Either $\|\sigma''\|_{2/5} \geqslant \Omega(\|\sigma'\|_{2/5})$, or the following holds. Let j° be the index maximizing $d_j^{5/2}2^{-j}$. Then 1) $j^\circ = \min_{j\in\mathcal{J}^*} j$, 2) $d_{j^\circ} = 1$, and 3) $j^\circ = 0$.*

Proof. This is essentially Lemma 5.26 from [65]. We will assume that $\|\sigma''\|_{2/5} = o(\|\sigma'\|_{2/5})$ and show that 1)-3) must hold. Let $i_{\max}$ be the index of the top entry of σ'. Suppose 1) does not hold. Then

$$\frac{\|\sigma'\|_{2/5}^{2/5}}{\|\sigma''\|_{2/5}^{2/5}} \leqslant \frac{\sigma_{i_{\max}}^{2/5} + \sum_{i\in S_{j^\circ}}\sigma_i^{2/5}}{\sum_{i\in S_{j^\circ}}\sigma_i^{2/5}} \leqslant 2, \tag{C.50}$$

where the first inequality follows by the elementary fact that for $a \geqslant b \geqslant 0$ and $c \geqslant 0$, $\frac{a+c}{b+c} \leqslant \frac{a}{b}$, and the second inequality follows by the definition of j°. This contradicts the assumption that $\|\sigma''\|_{2/5} = o(\|\sigma'\|_{2/5})$.

Next, suppose 1) holds but 2) does not. Then

$$\frac{\|\sigma'\|_{2/5}^{2/5}}{\|\sigma''\|_{2/5}^{2/5}} \leqslant \frac{\sum_{i\in S_{j^\circ}} \sigma_i^{2/5}}{\sum_{i\in S_{j^\circ}\setminus\{i_{\max}\}} \sigma_i^{2/5}} \leqslant O(1), \tag{C.51}$$

where in the first step we again used the above elementary fact and in the second step we used that 2) does not hold. We again get a contradiction.

Finally, suppose 1) and 2) hold, but 3) does not. Because 1) holds and $j^\circ > 0$, this implies that $\|\sigma'\|_{\mathrm{op}} \leqslant 1/2$. On the other hand, $\|\sigma''\|_{2/5} \geqslant \|\sigma''\|_1 \geqslant (1 - O(\varepsilon)) - 1/2 \geqslant 1/2 - O(\varepsilon)$. So for ε smaller than a sufficiently large constant, we get that $\|\sigma''\|_{2/5} \geqslant \Omega(\|\sigma'\|_{\mathrm{op}})$, so $\|\sigma''\|_{2/5} \geqslant \Omega(\|\sigma'\|_{2/5})$, a contradiction. □

Suppose the latter scenario in Lemma C.2.10 happens, but the former does not. In this case, because $d_{j^\circ} = 1$, we also have that $j^\circ = j_2$, so $1 \geqslant d_{j^\circ} 2^{-j^\circ} = d_{j^\circ}^2 2^{-j^\circ} = d_{j_c}^2 2^{-j_2}$. Note that this implies that $\|\sigma'\|_{1/2} \leqslant \log(d/\varepsilon)$. Furthermore, it implies that

$$1 \geqslant d_{j_2}^2 2^{-j_2} \geqslant d_{j_1}^2 2^{-j_1} \geqslant \Omega(d_{j_1}^{3/2}\varepsilon/\log(d/\varepsilon)), \tag{C.52}$$

where in the second step we used that $j_2 \arg\max_{j\in\mathcal{J}^*} d_j^2 2^{-j}$, and in the last step we used that $d_j 2^{-j} \geqslant \Omega(\varepsilon/\log(d/\varepsilon))$ by definition of S_{light} and $\mathcal{J}^*$. We conclude that

$$\varepsilon \leqslant O(d_{j_1}^{-3/2}\log(d/\varepsilon)). \tag{C.53}$$

But recall that we are assuming that (C.49) holds, i.e.

$$\varepsilon \gtrsim d_{j_2} \cdot 2^{-j_1/2 - j_2/2}/j_1 = 2^{-j_1/2-j_2/2}/j_1 \geqslant \Omega(\varepsilon/(d_{j_1}\log(d/\varepsilon)))^{1/2}/j_1, \tag{C.54}$$

where the second step is by $d_{j_2} = d_{j^\circ} = 1$ and the last step is by 3) in Lemma C.2.10 and the fact that $d_j 2^{-j} \geqslant \Omega(\varepsilon/\log(d/\varepsilon))$ for all $j \in \mathcal{J}^*$. Combining (C.53) and (C.54), we conclude that $d_{j_1} \leqslant \mathrm{polylog}(d/\varepsilon) \cdot j_1 \leqslant \mathrm{polylog}(d/\varepsilon)$, where in the last step we used (C.48). But if $d_{j_1} \leqslant \mathrm{polylog}(d/\varepsilon)$, then $d_{\mathsf{eff}} \leqslant \mathrm{polylog}(d/\varepsilon)$. Then because we

also have $\|\sigma'\|_{1/2} \leqslant O(\log(d/\varepsilon))$, the claimed lower bound in the theorem would follow from a lower bound of $\Omega(1/\varepsilon^2)$. This then follows in a similar fashion to the analysis from Case 1 above.

Finally, suppose instead that the former scenario in Lemma C.2.10 happens, in which case Lemma C.2.3 gives a lower bound of $\Omega(\|\sigma'\|_{2/5}/(\varepsilon^2 \log(d/\varepsilon)))$. Let j° be as defined in Lemma C.2.10. As $\|\sigma'\|_{2/5} \geqslant d_{j^\circ}^{5/2} 2^{-j^\circ}$, to complete the proof, it suffices to show that

$$d_{j^\circ}^{5/2} 2^{-j^\circ} \operatorname{polylog}(d/\varepsilon) \geqslant \Omega\left(\sqrt{d_{j_1}} d_{j_2}^2 2^{-j_2}\right). \tag{C.55}$$

Suppose to the contrary. Then because $d_{j_1}^{5/2} 2^{-j_1} \leqslant d_{j^\circ}^{5/2} 2^{-j^\circ}$, we would get from the negation of (C.55) that

$$d_{j_1}^2 2^{-j_1} \operatorname{polylog}(d/\varepsilon) = o(d_{j_2}^2 2^{-j_2}). \tag{C.56}$$

But by (C.49) and (C.48),

$$c d_{j_2} 2^{-j_1/2 - j_2/2 - 1} / O(\log(d/\varepsilon)) \leqslant \varepsilon \leqslant O(d_{j_1} 2^{-j_1} \log(d/\varepsilon)), \tag{C.57}$$

where in the last step we used that $d_{j_1} 2^{-j_1} \geqslant \Omega(\varepsilon \log(d/\varepsilon))$ by definition of S_{light}. Squaring and rearranging, we find that $d_{j_2}^2 2^{-j_2} \leqslant O(d_{j_1}^2 2^{-j_1} \log^2(d/\varepsilon))$, contradicting (C.56). □

C.3 Regularity Bounds on GOE and Ginibre Ensembles

Here we provide the proofs of Lemmas 5.3.2 and 5.4.1, restated for convenience.

Lemma 5.3.2. *There exists $U \subseteq \boldsymbol{R}^{d \times d}$ such that if $M \sim \mathrm{GOE}^*(d)$, then $\mathbf{Pr}\, M \notin U \leqslant \exp(-\Omega(d))$ and on the event $M \in U$, we have $\|M\|_{\mathrm{op}} \leqslant 3$ and $\|M\|_1 \geqslant d/12$.*

Proof of Lemma 5.3.2. Let U denote the event that $\|M\|_{\mathrm{op}} \leqslant 3$ and $\|M\|_F^2 \geqslant d/4$.

Let $\lambda_1, \ldots, \lambda_d$ denote the eigenvalues of U. On the event U, we have

$$\|M\|_F^2 = \sum_{i=1}^d \lambda_d^2 \leqslant \left(\max_{1 \leqslant i \leqslant d} |\lambda_i|\right) \sum_{i=1}^d |\lambda_i| = \|M\|_{\mathrm{op}} \|M\|_1,$$

so $\|M\|_1 \geqslant \|M\|_F^2 / \|M\|_{\mathrm{op}} \geqslant d/12$. We will show $\mathbf{Pr}\, U^c \leqslant \exp(-\Omega(d))$. We generate $M = G - \frac{\mathrm{tr}(G)}{d} I_d$, where $G \sim \mathrm{GOE}(d)$. Note that

$$\begin{aligned}\mathbf{Pr}\, \|M\|_{\mathrm{op}} > 3 &\leqslant \mathbf{Pr}\, \|G\|_{\mathrm{op}} > 5/2 + \mathbf{Pr}\, |\mathrm{tr}(G)| > d/2 \\ &\leqslant \exp(-\Omega(d)) + \exp(-\Omega(d^2)),\end{aligned}$$

where the first term is bounded by [44, Theorem 6.2] (because $5/2 > 2$) and the second term is bounded by $\mathrm{tr}(G) \sim \mathcal{N}(0, 2)$. Moreover, since

$$\|M\|_F^2 = \|G\|_F^2 - \mathrm{tr}(G)^2/d,$$

we have

$$\mathbf{Pr}\, \|M\|_F^2 < d/4 \leqslant \mathbf{Pr}\, \|G\|_F^2 < d/2 + \mathbf{Pr}\, |\mathrm{tr}(G)| > d/2$$

and the second probability is $\exp(-\Omega(d^2))$ as explained above. To bound the first probability, write $G_{i,i} = \sqrt{\frac{2}{d}} Z_{i,i}$ and for $i < j$, $G_{i,j} = G_{j,i} = \frac{1}{\sqrt{d}} Z_{i,j}$ for i.i.d. $Z_{i,i}, Z_{i,j} \sim \mathcal{N}(0, 1)$. Then

$$\|G\|_F^2 = \frac{2}{d} \sum_{1 \leqslant i \leqslant j \leqslant d} Z_{i,j}^2.$$

By a standard Chernoff bound, if $X \sim \chi^2(n)$, then $\mathbf{Pr}\, X \leqslant (1-\varepsilon)n \leqslant ((1-\varepsilon)e^{\varepsilon})^{n/2}$. Thus

$$\mathbf{Pr}\, *\|G\|_F^2 \leqslant d/4 = \mathbf{Pr}\, *\chi^2(d(d+1)/2) \leqslant d^2/8 \leqslant \exp(-\Omega(d^2)). \qquad \square$$

Lemma 5.4.1. *For $d_1 \geqslant d_2$, there exists $U \subseteq \mathbb{R}^{d_1 \times d_2}$ such that if $G \sim \mathrm{Gin}(d_1, d_2)$, then $\mathbf{Pr}\, G \notin U \leqslant \exp(-0.1 d_1)$ and on the event $G \in U$, we have $\|G\|_{\mathrm{op}} \leqslant 3$ and*

$\|M\|_1 \geqslant d_2/3$ *for*

$$M = \begin{pmatrix} 0 & G \\ G^\dagger & 0 \end{pmatrix}.$$

Proof of Lemma 5.4.1. Let U be the event $s_{\max}(G) \leqslant 3$ and $\|G\|_F^2 \geqslant d_2/2$, where $s_{\max}$ denotes the largest singular value. On this event, certainly $\|M\|_{\mathrm{op}} \leqslant 3$ and $\|M\|_F^2 = 2\|G\|_F^2 \geqslant d_2$. Similarly to the proof of Lemma 5.3.2, we have $\|M\|_1 \geqslant \|M\|_F^2/\|M\|_{\mathrm{op}} \geqslant d_2/3$. It remains to show $\mathbf{Pr}\, U^c \leqslant \exp(-0.1d_1)$ By [206, Corollary 5.35], $\mathbf{Pr}\, s_{\max}(G) > 3 \leqslant \exp(-0.11d_1)$. Moreover, $\|G\|_F^2 =_d \frac{1}{d_1}\chi^2(d_1 d_2)$, so similarly to the proof of Lemma 5.3.2 we have $\mathbf{Pr}\, \|G\|_F^2 < d_2/2 \leqslant \exp(-\Omega(d_1^2))$. □

C.4 Separating K and κ

In this short section we construct an example of a transcript $(\boldsymbol{z}, \boldsymbol{w})$ for which $K((\boldsymbol{z}, \boldsymbol{w})) = 0$, but for which $\kappa((\boldsymbol{z}, \boldsymbol{w})) \gg d_1 d_2^2/\varepsilon^2$. For simplicity, consider $A = a \cdot I_{d_1}$ and $B = b \cdot I_{d_2}$.

Consider a unit vector $(z, w) \in \mathbb{S}^{d_1+d_2-1}$ for which $\|z\|^2 = b/(a+b)$ and $\|w\|^2 = a/(a+b)$. Now note that if $(\boldsymbol{z}, \boldsymbol{w}) = ((z,w),(z,-w),(z,w),(z,-w),\ldots)$, then clearly $K((\boldsymbol{z}, \boldsymbol{w})) = 0$. On the other hand, if we take $b_i = (-1)^{i+1}$, we find that

$$\kappa((\boldsymbol{z}, \boldsymbol{w})) \geqslant \|\sum_{i=1}^{t} \frac{z w^\dagger \|z\|\|w\|}{(a\|z\|^2 + b\|w\|^2)^2}\|_F = t \cdot \frac{\|z\|^2\|w\|^2}{(a\|z\|^2 + b\|w\|^2)^2} = \frac{t}{4ab}. \tag{C.58}$$

Note that for $\varepsilon \asymp d_2\sqrt{ab}$, $d_1 d_2^2/\varepsilon^2 \asymp d_1/(ab)$, so for $t \gg d_1$, $\kappa((\boldsymbol{z}, \boldsymbol{w})) \gg d_1 d_2^2/\varepsilon^2$.

Bibliography

[1] Scott Aaronson. Shadow tomography of quantum states. In *STOC*, pages 325–338, 2018.

[2] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, and Shengjun Pan. Competitive closeness testing. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 47–68. JMLR Workshop and Conference Proceedings, 2011.

[3] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, and Ananda Suresh. Competitive classification and closeness testing. In *Conference on Learning Theory*, pages 22–1. JMLR Workshop and Conference Proceedings, 2012.

[4] David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985.

[5] Dorit Aharonov, Jordan Cotler, and Xiao-Liang Qi. Quantum algorithmic measurement. *Nature communications*, 13(1):1–9, 2022.

[6] Asma Al-Qasimi and Daniel F. V. James. Sudden death of entanglement at finite temperature. 77(1):012117. ISSN 1094-1622. doi: 10.1103/physreva.77.012117.

[7] Álvaro M. Alhambra. Quantum many-body systems in thermal equilibrium. 4 (4):040201. ISSN 2691-3399. doi: 10.1103/prxquantum.4.040201.

[8] M. P. Almeida, F. de Melo, M. Hor-Meyll, A. Salles, S. P. Walborn, P. H. Souto

Ribeiro, and L. Davidovich. Environment-induced sudden death of entanglement. 316(5824):579–582. ISSN 1095-9203. doi: 10.1126/science.1139892.

[9] Luigi Amico, Rosario Fazio, Andreas Osterloh, and Vlatko Vedral. Entanglement in many-body systems. 80(2):517–576. doi: 10.1103/revmodphys.80.517.

[10] Luigi Amico, Rosario Fazio, Andreas Osterloh, and Vlatko Vedral. Entanglement in many-body systems. *Reviews of modern physics*, 80(2):517–576, 2008.

[11] Nima Anari, Kuikui Liu, and Shayan Oveis Gharan. Spectral independence in high-dimensional expanders and applications to the hardcore model. ISSN 1095-7111. doi: 10.1137/20m1367696.

[12] Greg Anderson, Alice Guionnet, and Ofer Zeitouni. *An Introduction to Random Matrices*. Cambridge University Press, 2005.

[13] Anurag Anshu and Srinivasan Arunachalam. A survey on the complexity of learning quantum states. 2023.

[14] Anurag Anshu, Srinivasan Arunachalam, Tomotaka Kuwahara, and Mehdi Soleimanifar. Efficient learning of commuting hamiltonians on lattices. URL `https://anuraganshu.seas.harvard.edu/links`.

[15] Anurag Anshu, Srinivasan Arunachalam, Tomotaka Kuwahara, and Mehdi Soleimanifar. Sample-efficient learning of quantum many-body systems. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 685–691. IEEE, 2020. doi: 10.1109/focs46700.2020.00069.

[16] Anurag Anshu, Srinivasan Arunachalam, Tomotaka Kuwahara, and Mehdi Soleimanifar. Sample-efficient learning of interacting quantum systems. *Nature Physics*, May 2021. doi: 10.1038/s41567-021-01232-0. Preliminary version in FOCS 2020.

[17] Anurag Anshu, Zeph Landau, and Yunchao Liu. Distributed quantum inner product estimation. *arXiv preprint arXiv:2111.03273*, 2021.

[18] Leandro Aolita, Fernando de Melo, and Luiz Davidovich. Open-system dynamics of entanglement: A key issues review. *Reports on Progress in Physics*, 78(4): 042001, March 2015. ISSN 1361-6633. doi: 10.1088/0034-4885/78/4/042001.

[19] Leandro Aolita, Christian Gogolin, Martin Kliesch, and Jens Eisert. Reliable quantum certification of photonic state preparations. *Nature communications*, 6 (1):1–8, 2015.

[20] Itai Arad, Tomotaka Kuwahara, and Zeph Landau. Connecting global and local energy distributions in quantum spin models on a lattice. 2016(3):033301. doi: 10.1088/1742-5468/2016/03/033301.

[21] M. C. Arnesen, S. Bose, and V. Vedral. Natural thermal and magnetic entanglement in the 1D Heisenberg model. *Physical Review Letters*, 87(1):017901, June 2001. ISSN 1079-7114. doi: 10.1103/physrevlett.87.017901.

[22] MC Arnesen, S Bose, and V Vedral. Natural thermal and magnetic entanglement in the 1d heisenberg model. *Physical Review Letters*, 87(1):017901, 2001.

[23] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando GSL Brandao, David A Buell, et al. Quantum supremacy using a programmable superconducting processor. *Nature*, 574(7779):505–510, 2019.

[24] Guillaume Aubrun and Stanisław J. Szarek. Tensor products of convex sets and the volume of separable states on N qudits. 73(2):022109. ISSN 1094-1622. doi: 10.1103/physreva.73.022109.

[25] Koenraad MR Audenaert, Michael Nussbaum, Arleta Szkoła, and Frank Verstraete. Asymptotic error rates in quantum hypothesis testing. *Communications in Mathematical Physics*, 279(1):251–283, 2008.

[26] Costin Bădescu and Ryan O'Donnell. Improved quantum data analysis. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1398–1411, 2021.

[27] Costin Bădescu, Ryan O'Donnell, and John Wright. Quantum state certification. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 503–514, 2019.

[28] Emilio Bagan, MA Ballester, Richard D Gill, Alex Monras, and Ramon Munoz-Tapia. Optimal full estimation of qubit mixed states. *Physical Review A*, 73(3): 032301, 2006.

[29] Eyal Bairey, Itai Arad, and Netanel H. Lindner. Learning a local Hamiltonian from local measurements. *Physical Review Letters*, 122:020504, January 2019. doi: 10.1103/PhysRevLett.122.020504.

[30] Ainesh Bakshi, Ilias Diakonikolas, He Jia, Daniel M Kane, Pravesh K Kothari, and Santosh S Vempala. Robustly learning mixtures of k arbitrary gaussians. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1234–1247, 2022.

[31] Ainesh Bakshi, Allen Liu, Ankur Moitra, and Ewin Tang. High-temperature gibbs states are unentangled and efficiently preparable. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1027–1036. IEEE, 2024.

[32] Ainesh Bakshi, Allen Liu, Ankur Moitra, and Ewin Tang. Learning quantum hamiltonians at any temperature in polynomial time. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 1470–1477, 2024.

[33] Ainesh Bakshi, Allen Liu, Ankur Moitra, and Ewin Tang. Structure learning of hamiltonians from real-time evolution. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1037–1050. IEEE, 2024.

[34] Konrad Banaszek, Marcus Cramer, and David Gross. Focus on quantum tomography. *New Journal of Physics*, 15(12):125020, 2013.

[35] Boaz Barak. Proofs, beliefs, and algorithms through the lens of sum-of-squares.

[36] Boaz Barak and David Steurer. Sum-of-squares proofs and the quest toward optimal algorithms. *arXiv preprint arXiv:1404.5236*, 2014.

[37] Boaz Barak, Fernando GSL Brandao, Aram W Harrow, Jonathan Kelner, David Steurer, and Yuan Zhou. Hypercontractivity, sum-of-squares proofs, and their applications. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 307–326, 2012.

[38] Boaz Barak, Jonathan A Kelner, and David Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 143–151, 2015.

[39] Boaz Barak, Pravesh K Kothari, and David Steurer. Quantum entanglement, sum of squares, and the log rank conjecture. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 975–988, 2017.

[40] Ivan Bardet, Ángela Capel, Li Gao, Angelo Lucia, David Pérez-García, and Cambyse Rouzé. Rapid thermalization of spin chain commuting Hamiltonians. *Physical Review Letters*, 130(6):060401, February 2023. ISSN 1079-7114. doi: 10.1103/physrevlett.130.060401.

[41] Stephen M Barnett and Sarah Croke. Quantum state discrimination. *Advances in Optics and Photonics*, 1(2):238–278, 2009.

[42] Alexander Barvinok. *Combinatorics and complexity of partition functions*. Springer International Publishing. ISBN 9783319518299. doi: 10.1007/978-3-319-51829-9.

[43] J. S. Bell. On the einstein podolsky rosen paradox. *Physics Physique Fizika*, 1:195–200, Nov 1964. doi: 10.1103/PhysicsPhysiqueFizika.1.195. URL `https://link.aps.org/doi/10.1103/PhysicsPhysiqueFizika.1.195`.

[44] Gérard Ben Arous, Amir Dembo, and Alice Guionnet. Aging of spherical spin glasses. *Probab. Theory Related Fields*, 120(1):1–67, 2001.

[45] Ingemar Bengtsson and Karol Zyczkowski. *Geometry of quantum states: an introduction to quantum entanglement*. Cambridge University Press. ISBN 9780511535048. doi: 10.1017/cbo9780511535048.

[46] Charles H Bennett and David P DiVincenzo. Quantum information and computation. *nature*, 404(6775):247–255, 2000.

[47] Charles H Bennett, David P DiVincenzo, John A Smolin, and William K Wootters. Mixed-state entanglement and quantum error correction. *Physical Review A*, 54(5):3824, 1996.

[48] Eric Blais, Clément L Canonne, and Tom Gur. Distribution testing lower bounds via reductions from communication complexity. *ACM Transactions on Computation Theory (TOCT)*, 11(2):1–37, 2019.

[49] Andrew Blake, Pushmeet Kohli, and Carsten Rother. *Markov random fields for vision and image processing*. MIT press, 2011.

[50] Fernando G. S. L. Brandão and Michael J. Kastoryano. Finite correlation length implies efficient preparation of quantum thermal states. *Communications in Mathematical Physics*, 365(1):1–16, May 2018. ISSN 1432-0916. doi: 10.1007/s00220-018-3150-8.

[51] Fernando GSL Brandão, Matthias Christandl, and Jon Yard. A quasipolynomial-time algorithm for the quantum separability problem. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 343–352, 2011.

[52] Guy Bresler. Efficiently learning Ising models on arbitrary graphs. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing - STOC '15*. ACM Press, 2015. doi: 10.1145/2746539.2746631.

[53] Guy Bresler, Elchanan Mossel, and Allan Sly. Reconstruction of Markov random fields from samples: some observations and algorithms. *SIAM Journal on Computing*, 42(2):563–578, January 2013. doi: 10.1137/100796029.

[54] Sebastien Bubeck, Sitan Chen, and Jerry Li. Entanglement is necessary for optimal quantum property testing. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 692–703. IEEE, 2020.

[55] R. Bubley and M. Dyer. Path coupling: A technique for proving rapid mixing in Markov chains. In *Proceedings 38th Annual Symposium on Foundations of Computer Science*, SFCS-97. IEEE Comput. Soc, 1997. doi: 10.1109/sfcs.1997.646111.

[56] N. Canosa and R. Rossignoli. Global entanglement in XXZ chains. 73(2):022347. ISSN 1094-1622. doi: 10.1103/physreva.73.022347.

[57] Jose Carrasco, Andreas Elben, Christian Kokail, Barbara Kraus, and Peter Zoller. Theoretical and experimental perspectives of quantum verification. 2(1): 010102. doi: 10.1103/prxquantum.2.010102.

[58] Anthony Chefles. Quantum state discrimination. *Contemporary Physics*, 41(6): 401–424, 2000.

[59] Chi-Fang Chen and Fernando G. S. L. Brandão. Fast thermalization from the eigenstate thermalization hypothesis. 2021.

[60] Chi-Fang Chen, Michael J. Kastoryano, Fernando G. S. L. Brandão, and András Gilyén. Quantum thermal state preparation. 2023.

[61] Chi-Fang Chen, Michael J. Kastoryano, and András Gilyén. An efficient and exact noncommutative quantum Gibbs sampler. 2023.

[62] S. Chen, B. Huang, J. Li, A. Liu, and M. Sellke. When does adaptivity help for quantum state learning? In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 391–404, Los Alamitos, CA, USA, nov 2023. IEEE Computer Society. doi: 10.1109/FOCS57990.2023.00029. URL https://doi.ieeecomputersociety.org/10.1109/FOCS57990.2023.00029.

[63] Senrui Chen, Sisi Zhou, Alireza Seif, and Liang Jiang. Quantum advantages for pauli channel estimation. *Physical Review A*, 105(3):032435, 2022.

[64] Sitan Chen, Jordan Cotler, Hsin-Yuan Huang, and Jerry Li. A hierarchy for replica quantum advantage. *arXiv preprint arXiv:2111.05874*, 2021.

[65] Sitan Chen, Jerry Li, and Ryan O'Donnell. Toward instance-optimal state certification with incoherent measurements. *arXiv preprint arXiv:2102.13098*, 2021.

[66] Sitan Chen, Jordan Cotler, Hsin-Yuan Huang, and Jerry Li. Exponential separations between learning with and without quantum memory. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 574–585. IEEE, 2022.

[67] Sitan Chen, Jerry Li, Brice Huang, and Allen Liu. Tight bounds for quantum state certification with incoherent measurements. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1205–1213, 2022. doi: 10.1109/FOCS54457.2022.00118.

[68] Sitan Chen, Jerry Li, and Allen Liu. An optimal tradeoff between entanglement and copy complexity for state tomography. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 1331–1342, 2024.

[69] Yuansi Chen and Ronen Eldan. Localization schemes: A framework for proving mixing bounds for Markov chains. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. doi: 10.1109/focs54457.2022.00018.

[70] Anirban Narayan Chowdhury and Rolando D Somma. Quantum algorithms for gibbs sampling and hitting-time estimation. *arXiv preprint arXiv:1603.02940*, 2016.

[71] Peter Clifford. Markov random fields in statistics. *Disorder in physical systems: A volume in honour of John M. Hammersley*, pages 19–32, 1990.

[72] Marcus Cramer, Martin B. Plenio, Steven T. Flammia, Rolando Somma, David Gross, Stephen D. Bartlett, Olivier Landon-Cardinal, David Poulin, and Yi-Kai Liu. Efficient quantum state tomography. 1(1). doi: 10.1038/ncomms1147.

[73] Marcus P. da Silva, Olivier Landon-Cardinal, and David Poulin. Practical characterization of quantum devices without tomography. 107(21):210404. doi: 10.1103/physrevlett.107.210404.

[74] Marcus P da Silva, Olivier Landon-Cardinal, and David Poulin. Practical characterization of quantum devices without tomography. *Physical Review Letters*, 107(21):210404, 2011.

[75] Ilias Diakonikolas and Daniel M Kane. A new approach for testing properties of discrete distributions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 685–694. IEEE, 2016.

[76] Ilias Diakonikolas and Daniel M Kane. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*, 2019.

[77] Andrew C Doherty, Pablo A Parrilo, and Federico M Spedalieri. Distinguishing separable and entangled states. *Physical Review Letters*, 88(18):187904, 2002.

[78] M. Endres, M. Cheneau, T. Fukuhara, C. Weitenberg, P. Schauß, C. Gross, L. Mazza, M. C. Bañuls, L. Pollet, I. Bloch, and S. Kuhr. Observation of correlated particle-hole pairs and string order in low-dimensional Mott insulators. 334(6053):200–203. doi: 10.1126/science.1209284.

[79] Omar Fawzi, Nicolas Flammarion, Aurélien Garivier, and Aadil Oufkir. Quantum channel certification with incoherent measurements. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 1822–1884. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/fawzi23a.html.

[80] Omar Fawzi, Nicolas Flammarion, Aurélien Garivier, and Aadil Oufkir. On Adaptivity in Quantum Testing. working paper or preprint, May 2023. URL https://hal.science/hal-04107265.

[81] Christopher Ferrie and Robin Blume-Kohout. Minimax quantum tomography: the ultimate bounds on accuracy. *arXiv preprint arXiv:1503.03100*, 2015.

[82] Richard P Feynman. Simulating physics with computers. In *Feynman and computation*, pages 133–153. cRc Press, 2018.

[83] B. V. Fine, F. Mintert, and A. Buchleitner. Equilibrium entanglement vanishes at finite temperature. 71(15):153105. ISSN 1550-235X. doi: 10.1103/physrevb.71.153105.

[84] Steve Flammia and Ryan O'Donnell. Quantum chi-squared tomography and mutual information testing. *preprint*, 2023.

[85] Steven T Flammia and Yi-Kai Liu. Direct fidelity estimation from few pauli measurements. *Physical review letters*, 106(23):230501, 2011.

[86] Steven T Flammia, David Gross, Yi-Kai Liu, and Jens Eisert. Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators. *New Journal of Physics*, 14(9):095022, 2012.

[87] Noah Fleming, Pravesh Kothari, Toniann Pitassi, et al. *Semialgebraic Proofs and Efficient Algorithm Design.* now the essence of knowledge, 2019.

[88] Sacha Friedli and Yvan Velenik. *Statistical mechanics of lattice systems: a concrete mathematical introduction.* Cambridge University Press, 2017. ISBN 978-1-107-18482-4. doi: 10.1017/9781316882603.

[89] Christopher A Fuchs and Jeroen Van De Graaf. Cryptographic distinguishability measures for quantum-mechanical states. *IEEE Transactions on Information Theory*, 45(4):1216–1227, 1999.

[90] Andreas Galanis, Daniel Stefankovic, and Eric Vigoda. Inapproximability of the partition function for the antiferromagnetic Ising and hard-core models. 25(4): 500–559. ISSN 1469-2163. doi: 10.1017/s0963548315000401.

[91] Sevag Gharibian, Yichen Huang, Zeph Landau, and Seung Woo Shin. Quantum Hamiltonian complexity. 10(3):159–282. doi: 10.1561/0400000066.

[92] András Gilyén, Yuan Su, Guang Hao Low, and Nathan Wiebe. Quantum singular value transformation and beyond: Exponential improvements for quantum matrix arithmetics. In *Proceedings of the 51st ACM Symposium on the Theory of Computing (STOC)*, pages 193–204. ACM, June 2019. doi: 10.1145/3313276.3316366.

[93] Roe Goodman, Nolan R Wallach, et al. *Symmetry, representations, and invariants*, volume 255. Springer, 2009.

[94] Dima Grigoriev. Complexity of positivstellensatz proofs for the knapsack. *computational complexity*, 10(2):139–154, 2001.

[95] Christian Gross and Immanuel Bloch. Quantum simulations with ultracold atoms in optical lattices. 357(6355):995–1001. doi: 10.1126/science.aal3837.

[96] Martin Grötschel, László Lovász, and Alexander Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1:169–197, 1981.

[97] Andi Gu, Lukasz Cincio, and Patrick J. Coles. Practical black box hamiltonian learning. 2022.

[98] Otfried Gühne and Géza Tóth. Entanglement detection. *Physics Reports*, 474 (1-6):1–75, 2009.

[99] D. Gunlycke, V. M. Kendon, V. Vedral, and S. Bose. Thermal concurrence mixing in a one-dimensional ising model. *Physical Review A*, 64(4):042302, September 2001. ISSN 1094-1622. doi: 10.1103/physreva.64.042302.

[100] Leonid Gurvits and Howard Barnum. Separable balls around the maximally mixed multipartite quantum states. *Physical Review A*, 68(4):042312, October 2003. ISSN 1094-1622. doi: 10.1103/physreva.68.042312.

[101] Madalin Guţă, Jonas Kahn, Richard Kueng, and Joel A Tropp. Fast state tomography with optimal error bounds. *Journal of Physics A: Mathematical and Theoretical*, 53(20):204001, 2020.

[102] Jeongwan Haah, Aram W Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu. Sample-optimal tomography of quantum states. *IEEE Trans. Inf. Theory*, 63 (9):5628–5641, 2017.

[103] Jeongwan Haah, Robin Kothari, and Ewin Tang. Optimal learning of quantum Hamiltonians from high-temperature Gibbs states. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, October 2022. doi: 10.1109/focs54457.2022.00020.

[104] Hartmut Häffner, Wolfgang Hänsel, CF Roos, Jan Benhelm, D Chek-al Kar, M Chwalla, T Körber, UD Rapol, M Riebe, PO Schmidt, et al. Scalable multiparticle entanglement of trapped ions. *Nature*, 438(7068):643–646, 2005.

[105] Linus Hamilton, Frederic Koehler, and Ankur Moitra. Information theoretic properties of markov random fields, and their algorithmic applications. *Advances in Neural Information Processing Systems*, 30, 2017.

[106] J.M. Hammersley and P. Clifford. Markov field on finite graphs and lattices. Available at `http://www.statslab.cam.ac.uk/~grg/books/hammfest/hamm-cliff.pdf`, 1971.

[107] Aram W Harrow. The church of the symmetric subspace. *arXiv preprint arXiv:1308.6595*, 2013.

[108] Aram W. Harrow, Saeed Mehraban, and Mehdi Soleimanifar. Classical algorithms, correlation decay, and complex zeros of partition functions of quantum many-body systems. In ., editor, *Proceedings of the 52nd Annual ACM*

SIGACT Symposium on Theory of Computing, STOC '20. ACM, June 2020. doi: 10.1145/3357713.3384322.

[109] O. Hart and C. Castelnovo. Entanglement negativity and sudden death in the toric code at finite temperature. 97(14):144410. ISSN 2469-9969. doi: 10.1103/physrevb.97.144410.

[110] Matthew B Hastings. Locality in quantum systems. In Manfred Salmhofer, Vieri Mastropietro, Wojciech De Roeck, and Leticia F. Cugliandolo, editors, *Quantum Theory from Small to Large Scales*, volume 95 of *Lecture Notes of the les Houches Summer School Ser.*, pages 171–212. Oxford University Press, Incorporated. ISBN 9780191623769.

[111] Matthew B Hastings. Field theory and the sum-of-squares for quantum systems. *arXiv preprint arXiv:2302.14006*, 2023.

[112] Matthew B Hastings and Ryan O'Donnell. Optimizing strongly interacting fermionic hamiltonians. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 776–789, 2022.

[113] Masahito Hayashi and Keiji Matsumoto. Asymptotic performance of optimal state estimation in qubit system. *Journal of Mathematical Physics*, 49(10): 102101, 2008.

[114] Roland Hildebrand. Entangled states close to the maximally mixed state. 75(6): 062330. ISSN 1094-1622. doi: 10.1103/physreva.75.062330.

[115] Sarah Hirthe, Thomas Chalopin, Dominik Bourgund, Petar Bojović, Annabelle Bohrdt, Eugene Demler, Fabian Grusdt, Immanuel Bloch, and Timon A. Hilker. Magnetically mediated hole pairing in fermionic ladders of ultracold atoms. 613 (7944):463–467. doi: 10.1038/s41586-022-05437-y.

[116] Samuel B Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034, 2018.

[117] Michał Horodecki, Paweł Horodecki, and Ryszard Horodecki. Mixed-state entanglement and distillation: is there a "bound" entanglement in nature? 80 (24):5239–5242, . ISSN 1079-7114. doi: 10.1103/physrevlett.80.5239.

[118] Paweł Horodecki and Artur Ekert. Method for direct detection of quantum entanglement. *Physical review letters*, 89(12):127902, 2002.

[119] Ryszard Horodecki, Paweł Horodecki, Michał Horodecki, and Karol Horodecki. Quantum entanglement. 81(2):865–942, . ISSN 1539-0756. doi: 10.1103/revmodphys.81.865.

[120] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. 16(10):1050–1057, . doi: 10.1038/s41567-020-0932-7.

[121] Hsin-Yuan Huang, Yu Tong, Di Fang, and Yuan Su. Learning many-body Hamiltonians with Heisenberg-limited scaling. 130(20):200403, . doi: 10.1103/physrevlett.130.200403.

[122] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics*, 16(10): 1050–1057, 2020.

[123] Hsin-Yuan Huang, Michael Broughton, Jordan Cotler, Sitan Chen, Jerry Li, Masoud Mohseni, Hartmut Neven, Ryan Babbush, Richard Kueng, John Preskill, et al. Quantum advantage in learning from experiments. *arXiv preprint arXiv:2112.00778*, 2021.

[124] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Information-theoretic bounds on quantum advantage in machine learning. *Physical Review Letters*, 126(19):190505, 2021.

[125] Hsin-Yuan Huang, Michael Broughton, Jordan Cotler, Sitan Chen, Jerry Li, Masoud Mohseni, Hartmut Neven, Ryan Babbush, Richard Kueng, John Preskill,

et al. Quantum advantage in learning from experiments. *Science*, 376(6598): 1182–1186, 2022.

[126] Rajibul Islam, Ruichao Ma, Philipp M Preiss, M Eric Tai, Alexander Lukin, Matthew Rispoli, and Markus Greiner. Measuring entanglement entropy in a quantum many-body system. *Nature*, 528(7580):77–83, 2015.

[127] Leon Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1–2): 134–139, 1918.

[128] Mark Jerrum, Alistair Sinclair, and Eric Vigoda. A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. 51(4): 671–697. ISSN 1557-735X. doi: 10.1145/1008731.1008738.

[129] Mark R Jerrum, Leslie G Valiant, and Vijay V Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical computer science*, 43:169–188, 1986.

[130] Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Minimax estimation of the l_1 distance. *IEEE Transactions on Information Theory*, 64(10):6672–6706, 2018.

[131] Anatoli Juditsky and Arkadii S Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. *arXiv preprint arXiv:0809.0813*, 2008.

[132] Michael J. Kastoryano and Fernando G. S. L. Brandão. Quantum Gibbs samplers: The commuting case. 344(3):915–957. ISSN 1432-0916. doi: 10.1007/s00220-016-2641-8.

[133] Kohtaro Kato and Fernando G. S. L. Brandão. Quantum approximate Markov chains are thermal. *Communications in Mathematical Physics*, 370(1):117–149, June 2019. doi: 10.1007/s00220-019-03485-6.

[134] M. Keyl and R. F. Werner. Estimating the spectrum of a density operator. *Physical Review A*, 64(5), October 2001. ISSN 1094-1622. doi: 10.1103/physreva.64.052311. URL http://dx.doi.org/10.1103/PhysRevA.64.052311.

[135] Michael Keyl. Quantum state estimation and large deviations. *Reviews in Mathematical Physics*, 18(01):19–60, 2006.

[136] Alexei Yu Kitaev, Alexander Shen, and Mikhail N Vyalyi. *Classical and quantum computation*. Number 47. American Mathematical Soc., 2002.

[137] M. Kliesch, C. Gogolin, M.J. Kastoryano, A. Riera, and J. Eisert. Locality of temperature. *Physical Review X*, 4(3):031019, July 2014. ISSN 2160-3308. doi: 10.1103/physrevx.4.031019.

[138] Adam Klivans and Raghu Meka. Learning graphical models using multiplicative weights. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, October 2017. doi: 10.1109/focs.2017.39.

[139] Adam Klivans, Pravesh K Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory*, pages 1420–1430. PMLR, 2018.

[140] Christian Kokail, Rick van Bijnen, Andreas Elben, Benoit Vermersch, and Peter Zoller. Entanglement Hamiltonian tomography in quantum simulation. 17(8): 936–942. doi: 10.1038/s41567-021-01260-w.

[141] R. Kotecký and D. Preiss. Cluster expansion for abstract polymer models. 103 (3):491–498. ISSN 1432-0916. doi: 10.1007/bf01211762.

[142] Pravesh K Kothari, Jacob Steinhardt, and David Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1035–1046, 2018.

[143] H. A. Kramers and G. H. Wannier. Statistics of the two-dimensional ferromagnet. part i. *Phys. Rev.*, 60:252–262, Aug 1941. doi: 10.1103/PhysRev.60.252. URL `https://link.aps.org/doi/10.1103/PhysRev.60.252`.

[144] Richard Kueng, Holger Rauhut, and Ulrich Terstiege. Low rank matrix recovery from rank one measurements. *Applied and Computational Harmonic Analysis*, 42(1):88–116, 2017.

[145] Tomotaka Kuwahara, Kohtaro Kato, and Fernando G. S. L. Brandão. Clustering of conditional mutual information for quantum Gibbs states above a threshold temperature. *Physical Review Letters*, 124(22), June 2020. doi: 10.1103/physrevlett.124.220601.

[146] Tomotaka Kuwahara, Álvaro M. Alhambra, and Anurag Anshu. Improved thermal area law and quasilinear time algorithm for quantum Gibbs states. *Physical Review X*, 11(1):011047, March 2021. doi: 10.1103/physrevx.11.011047.

[147] Jean B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001. doi: 10.1137/S1052623400366802. URL https://doi.org/10.1137/S1052623400366802.

[148] Monique Laurent. Sums of squares, moment matrices and optimization over polynomials. In *Emerging applications of algebraic geometry*, pages 157–270. Springer, 2009.

[149] J. P. F. LeBlanc, Andrey E. Antipov, Federico Becca, Ireneusz W. Bulik, Garnet Kin-Lic Chan, Chia-Min Chung, Youjin Deng, Michel Ferrero, Thomas M. Henderson, Carlos A. Jiménez-Hoyos, E. Kozik, Xuan-Wen Liu, Andrew J. Millis, N. V. Prokof'ev, Mingpu Qin, Gustavo E. Scuseria, Hao Shi, B. V. Svistunov, Luca F. Tocchio, I. S. Tupitsyn, Steven R. White, Shiwei Zhang, Bo-Xiao Zheng, Zhenyue Zhu, and Emanuel Gull. Solutions of the two-dimensional hubbard model: Benchmarks and results from a wide range of numerical algorithms. *Phys. Rev. X*, 5:041041, Dec 2015. doi: 10.1103/PhysRevX.5.041041. URL https://link.aps.org/doi/10.1103/PhysRevX.5.041041.

[150] David A. Levin and Yuval Peres. *Markov chains and mixing times*. American Mathematical Society, Providence, RI, second edition, 2017. ISBN 978-1-4704-2962-1. doi: 10.1090/mbk/107.

[151] Yahui Li, Frank Pollmann, Nicholas Read, and Pablo Sala. Highly-entangled stationary states from strong symmetries. 2024.

[152] Elliott H. Lieb and Derek W. Robinson. The finite group velocity of quantum spin systems. 28(3):251–257. doi: 10.1007/bf01645779.

[153] Norbert M Linke, Sonika Johri, Caroline Figgatt, Kevin A Landsman, Anne Y Matsuura, and Christopher Monroe. Measuring the rényi entropy of a two-site fermi-hubbard model on a trapped ion quantum computer. *Physical Review A*, 98(5):052334, 2018.

[154] Allen Liu and Ankur Moitra. Settling the robust learnability of mixtures of gaussians. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 518–531, 2021.

[155] Seth Lloyd. Universal quantum simulators. *Science*, 273(5278):1073–1078, 1996.

[156] Angus Lowe. Learning quantum states without entangled measurements. Master's thesis, University of Waterloo, 2021.

[157] Muzhou Ma, Steven T Flammia, John Preskill, and Yu Tong. Learning k-body hamiltonians via compressed sensing. *arXiv preprint arXiv:2410.18928*, 2024.

[158] Xiao-Song Ma, Thomas Herbst, Thomas Scheidl, Daqing Wang, Sebastian Kropatschek, William Naylor, Bernhard Wittmann, Alexandra Mech, Johannes Kofler, Elena Anisimova, et al. Quantum teleportation over 143 kilometres using active feed-forward. *Nature*, 489(7415):269–273, 2012.

[159] Dylan H Mahler, Lee A Rozema, Ardavan Darabi, Christopher Ferrie, Robin Blume-Kohout, and AM Steinberg. Adaptive quantum state tomography improves accuracy quadratically. *Physical review letters*, 111(18):183601, 2013.

[160] Ryan L. Mann and Tyler Helmuth. Efficient algorithms for approximating quantum partition functions. 62(2). ISSN 1089-7658. doi: 10.1063/5.0013689.

[161] Marc Mézard and Andrea Montanari. *Information, Physics, and Computation*. Oxford Graduate Texts. Oxford University Press, 2009. ISBN 9780198507540.

[162] Amin Moharramipour, Leonardo A. Lessa, Chong Wang, Timothy H. Hsieh, and Subhayan Sahu. Symmetry-enforced entanglement in maximally mixed states. 5 (4):040336. ISSN 2691-3399. doi: 10.1103/prxquantum.5.040336.

[163] Andrea Montanari. Computational implications of reducing data to sufficient statistics. 2015.

[164] Ashley Montanaro. Learning stabilizer states by bell sampling, 2017.

[165] Ashley Montanaro and Ronald de Wolf. A survey of quantum property testing. *Theory of Computing*, pages 1–81, 2016.

[166] Yurii Nesterov. Squared functional systems and optimization problems. In *High performance optimization*, pages 405–440. Springer, 2000.

[167] K. Netočný and F. Redig. Large deviations for quantum spin systems. 117(3–4): 521–547. ISSN 1572-9613. doi: 10.1007/s10955-004-3452-4.

[168] Michael A Nielsen and Isaac Chuang. Quantum computation and quantum information, 2002.

[169] Ryan O'Donnell and John Wright. Quantum spectrum testing. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 529–538, 2015.

[170] Ryan O'Donnell and John Wright. Efficient quantum tomography. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 899–912, 2016.

[171] Ryan O'Donnell and John Wright. Efficient quantum tomography II. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 962–974, 2017.

[172] Lars Onsager. Crystal statistics. i. a two-dimensional model with an order–disorder transition. *Physical Review*, 65:117–149, 1944.

[173] Tobias J Osborne and Michael A Nielsen. Entanglement in a simple quantum phase transition. *Physical Review A*, 66(3):032110, 2002.

[174] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10): 4750–4755, 2008.

[175] Ojas Parekh and Kevin Thompson. Application of the level-2 quantum lasserre hierarchy in quantum approximation algorithms. *arXiv preprint arXiv:2105.05698*, 2021.

[176] Gilles Parez and William Witczak-Krempa. The fate of entanglement. 2024.

[177] Pablo A Parrilo. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*. PhD thesis, California Institute of Technology, 2000.

[178] R. K. Pathria and Paul D. Beale. *Statistical Mechanics*. Elsevier, Amsterdam, 3rd edition, 2011. ISBN 978-0123821881. See Ch. 13 for a discussion of liquid–gas (boiling) phase transitions.

[179] Martin B Plenio and Vlatko Vedral. Teleportation, entanglement and thermodynamics in the quantum world. *Contemporary physics*, 39(6):431–446, 1998.

[180] Martin B. Plenio and S. Virmani. An introduction to entanglement measures, 2006. URL `https://arxiv.org/abs/quant-ph/0504163`.

[181] John Preskill. Quantum computing in the NISQ era and beyond. *Quantum*, 2: 79, August 2018. doi: 10.22331/q-2018-08-06-79.

[182] Xiao-Liang Qi and Daniel Ranard. Determining a local Hamiltonian from a single eigenstate. *Quantum*, 3:159, July 2019. ISSN 2521-327X. doi: 10.22331/q-2019-07-08-159.

[183] Prasad Raghavendra. Optimal algorithms and inapproximability results for every csp? In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pages 245–254, 2008. doi: 10.1145/1374376.1374414. URL https://doi.org/10.1145/1374376.1374414.

[184] Dan Romik. *The surprising mathematics of longest increasing subsequences*. Number 4. Cambridge University Press, 2015.

[185] Cambyse Rouzé, Daniel Stilck França, and Álvaro M. Alhambra. Efficient thermalization and universal quantum computing with quantum Gibbs samplers. 2024.

[186] Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC, 2005.

[187] Bruce E Sagan. *The symmetric group: representations, combinatorial algorithms, and symmetric functions*, volume 203. Springer Science & Business Media, 2013.

[188] Elad Schneidman, Michael J. II Berry, Ronen Segev, and William Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–1012, 2006.

[189] Tselil Schramm and Alexander S Wein. Computational barriers to estimation from low-degree polynomials. *arXiv preprint arXiv:2008.02269*, 2020.

[190] E. Schrödinger. Discussion of probability relations between separated systems. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(4):555–563, 1935. doi: 10.1017/S0305004100013554.

[191] Nicholas E. Sherman, Trithep Devakul, Matthew B. Hastings, and Rajiv R. P. Singh. Nonzero-temperature entanglement negativity of quantum spin models: area law, linked cluster expansions, and sudden death. 93(2):022128. ISSN 2470-0053. doi: 10.1103/physreve.93.022128.

[192] Naum Zuselevich Shor. An approach to obtaining global extremums in polynomial mathematical programming problems. *Cybernetics*, 23(5):695–700, 1987.

[193] M. D. Shulman, S. P. Harvey, J. M. Nichol, S. D. Bartlett, A. C. Doherty, V. Umansky, and A. Yacoby. Suppressing qubit dephasing using real-time Hamiltonian estimation. 5(1). doi: 10.1038/ncomms6156.

[194] Alistair Sinclair and Mark Jerrum. Approximate counting, uniform generation and rapidly mixing Markov chains. 82(1):93–133. ISSN 0890-5401. doi: 10.1016/0890-5401(89)90067-9.

[195] Alistair Sinclair, Piyush Srivastava, and Marc Thurley. Approximation algorithms for two-state anti-ferromagnetic spin systems on bounded degree graphs. 155(4):666–686. ISSN 1572-9613. doi: 10.1007/s10955-014-0947-5.

[196] Allan Sly. Computational transition at the uniqueness threshold. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE. doi: 10.1109/focs.2010.34.

[197] Allan Sly and Nike Sun. Counting in two-spin models on d-regular graphs. 42 (6). ISSN 0091-1798. doi: 10.1214/13-aop888.

[198] David Steurer and Stefan Tiegel. SoS degree reduction with applications to clustering and robust moment estimation. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 374–393. Society for Industrial and Applied Mathematics. doi: 10.1137/1.9781611976465.24.

[199] Szilárd Szalay. Multipartite entanglement measures. *Physical Review A*, 92(4): 042329, 2015.

[200] Ewin Tang and Kevin Tian. A CS guide to the quantum singular value transformation, 2023.

[201] K. Temme, T. J. Osborne, K. G. Vollbrecht, D. Poulin, and F. Verstraete. Quantum Metropolis sampling. 471(7336):87–90. ISSN 1476-4687. doi: 10.1038/nature09770.

[202] The Nobel Prize Committee. The nobel prize in physics 2021 – press release. https://www.nobelprize.org/prizes/physics/2021/press-release/, 2021. Accessed: 2025-05-09.

[203] Daniel Ueltschi. Cluster expansions and correlation functions. *Mosc. Math. J.*, 4(2):511–522, 536, 2004. ISSN 1609-3321. doi: 10.17323/1609-4514-2004-4-2-511-522.

[204] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017.

[205] Joran van Apeldoorn, Arjan Cornelissen, András Gilyén, and Giacomo Nannicini. Quantum tomography using state-preparation unitaries. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1265–1318. SIAM, 2023.

[206] Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices*. Cambridge University Press, 2012.

[207] G. Vidal and R. F. Werner. Computable measure of entanglement. 65(3):032314. ISSN 1094-1622. doi: 10.1103/physreva.65.032314.

[208] Guifré Vidal. Entanglement monotones. *Journal of Modern Optics*, 47(2-3): 355–376, 2000.

[209] Adam Bene Watts, Anirban Chowdhury, Aidan Epperly, J William Helton, and Igor Klep. Relaxations and exact solutions to quantum max cut via the algebraic structure of swap operators. *arXiv preprint arXiv:2307.15661*, 2023.

[210] Dror Weitz. Counting independent sets up to the tree threshold. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of Computing*, STOC06. ACM. doi: 10.1145/1132516.1132538.

[211] Robin Yunfei Wen and Achim Kempf. Separable ball around any full-rank multipartite product state. 56(33):335302. ISSN 1751-8121. doi: 10.1088/1751-8121/ace810.

[212] Nathan Wiebe, Cassandra Granade, Christopher Ferrie, and D. G. Cory. Hamiltonian learning and certification using quantum resources. 112(19):190501. doi: 10.1103/physrevlett.112.190501.

[213] Michael M. Wolf, Frank Verstraete, Matthew B. Hastings, and J. Ignacio Cirac. Area laws in quantum systems: mutual information and correlations. *Physical Review Letters*, 100(7):070502, February 2008. ISSN 1079-7114. doi: 10.1103/physrevlett.100.070502.

[214] William K Wootters. Entanglement of formation of an arbitrary state of two qubits. *Physical Review Letters*, 80(10):2245, 1998.

[215] John Wright. How to learn a quantum state. *Ph. D. dissertation*, 2016.

[216] Yihong Wu. Lecture notes on information-theoretic methods for high-dimensional statistics. `http://www.stat.yale.edu/~yw562/teaching/it-stats.pdf`, 2017.

[217] C. N. Yang. The spontaneous magnetization of a two-dimensional ising model. *Physical Review*, 85(5):808–816, 1952. doi: 10.1103/PhysRev.85.808.

[218] Chao Yin and Andrew Lucas. Polynomial-time classical sampling of high-temperature quantum Gibbs states. 2023.

[219] Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

[220] Nengkun Yu. Quantum closeness testing: A streaming algorithm and applications, 2019.

[221] Ting Yu and J. H. Eberly. Finite-time disentanglement via spontaneous emission. 93(14):140404, . ISSN 1079-7114. doi: 10.1103/physrevlett.93.140404.

[222] Ting Yu and J. H. Eberly. Sudden death of entanglement. 323(5914):598–601, . ISSN 1095-9203. doi: 10.1126/science.1167343.

[223] Andrew Zhao. Learning the structure of any hamiltonian from minimal assumptions. *arXiv preprint arXiv:2410.21635*, 2024.

[224] Assaf Zubida, Elad Yitzhaki, Netanel H. Lindner, and Eyal Bairey. Optimal short-time measurements for Hamiltonian learning. 2021.