The Dissertation Committee for Daniel You Liang certifies that this is the approved version of the following dissertation:

On Computationally Efficient Learning for Stabilizers and Beyond

Committee:

Scott Aaronson, Supervisor

Srinivasan Arunachalam

Adam Klivans

David Soloveichik

On Computationally Efficient Learning for Stabilizers and Beyond

by

Daniel You Liang

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2023

To my parents.

Acknowledgments

While I hope that it's not the peak, I can soundly say that my PhD has been one of the best times of my life so far. Through the highs, lows, and everything in-between, I have learned so much and grown so much both as a researcher as well as a person. It was also not possible without countless people, to whom I will now do my best (and fail) to properly thank.

I'd first like to thank my parents. While they didn't always understand the differences between my field and theirs, they understood what kind of commitment and undertaking a PhD is and always gave me love and support. They were also the ones who inspired me to pursue a PhD in the first place. Growing up and hearing about my parents having broadened the boundaries of science filled me with awe as a child, pushing me to go above and beyond academically to be just like them. So for that, I cannot thank them enough.

Next I'd like to thank my advisor, Dr. Scott Aaronson. As an undergraduate who only took up a CS major in his senior year and who had only done experimental physics research, I know that I was a gamble to admit as a student. I'd like to think, six years later, that gamble and the faith shown in me has now paid off. Dr. Aaronson has a notoriously hands-off advising approach and the freedom to pursue nearly whatever I wanted was a freedom I struggled with initially. Nevertheless, it has allowed me to become a stronger and more independent researcher than I could have imagined when I first started. The weekly group meetings and other less organized meetings also gave me so much insight into the various tools, results, and open questions in a wide variety of subfields in quantum computing/information and even some outside of it. These were a hugely important part of my early PhD experience when I had such little incoming experience. The QIS I and QIS II courses that I got to TA for Dr. Aaronson were also an incredible opportunity, allowing me to hone my fundamentals as well as learn some new material from a great lecturer. Needless to say, I get a little happy every time I get to reference the QIS II lecture notes I helped compile. It was an honor being a student in the Aaronson group and it is an honor to be inducted into a group of researchers whom I greatly admire.

While I can't say I enjoyed braving the Austin summer heat to get to my office, I'd like to thank all of the various students, postdocs, and administrators I've had the pleasure of interacting with while at UT-Austin. These were the people that made me look forward to go to the office, even when I was swamped with work and other duties. I'd like to thank Patrick Rall for graciously allowing me to piggyback on his Pauli propogation algorithm idea, leading to my first published work. While it did not directly make it into the dissertation, it was instrumental in allowing me to develop many of the tools now found in this dissertation. This was also in collaboration with William Kretschmer, whose desk was next to mine in the Gates Dell Complex and who has gone on to be my most frequent co-author and answer-er of my dumb questions on quantum computation/information. I also cannot thank Aravind Gollakota enough for co-authoring the first work (chronologically) that went into this dissertation and being a constant sounding board for all of my other dumb questions on learning theory. Thanks to Sabee Grewal and Vishnu Iyer as well for being wonderful and brilliant collaborators and group members. Without their efforts, much of the first part of this dissertation would likely be significantly worse off. Additionally, I'd like to thank (in no particular order) Justin Oh, Robert Shi, Justin Yirka, Jiahui Liu, Yuxuan Zhang, Matthew Jordan, Kevin Song, John Kallaugher, Akshay Kamath, Sushrut Karmalkar, Josh Vekhter, Ewin Tang, Andrea Rocchetto, Chunhao Wang, Han-Hsuan Lin, Nai-Hui Chia, Yosi Atia, Supartha Podder, Jason Pollack, Shih-Han Hung, Nick Hunter-Jones, Corey Ostrove, Ridwan Syed, Adrien Trejo Nuñes, Bryce Fuller, and countless others at UT-Austin for being great colleagues, classmates, people, and friends. I'd also like to extend my gratitude towards administrators Caitlin Young, Monica Aguilar, Kathryn Murphy, and Katie Traughber for making my life at UT a whole lot easier and letting me focus on research. Researchers outside of UT-Austin who helped me in my work include co-authors Stefan Leichenauer, Li Li, and Jeremy Cook, as well Srinivasan Arunachalam, Matthias Caro, Arkopal Dutt, John Wright, Robert Kueng, Marcel Hinsche, and Sepehr Nezami, all of whom provided helpful advice and conversations.

Outside of research, I had many other people supporting me. I'd like to thank all of my current/former roommates (William Ruys, Tyler Masthay, Graham Pash, Matthew Goldberg, Nereida Lewis-Ramirez, and Anna Yesypenko) for helping me deal with the Covid isolations as well as the power outages, the boil water notices, the freezes, and everything else that Austin threw at us. But also, thank you for your friendship and camaraderie. To my brother, Benjamin Liang and my childhood friend, Eric Wang, may raging barbarians always be on. To Andrew Kim, Michael Disare, and Samuel Newman-Stonebraker, I'm sorry I spoiled those Nintendo directs. To my friends who need auda, bring it in. To my group hardcore ironman squad, still five lives (as of the time of this writing). To Justin Lu and Matthew Wallingford, CHDL. Finally, I think I would be remiss to not also acknowledge the O's Campus Cafe Chicken Avocado Wrap that helped fuel me throughout most of my PhD.

Last but not least, I'd like to give special thanks to my girlfriend, Nisha Holay. For giving me a shoulder to lean on. For pushing me to have a good work-life balance when I swing too hard in one direction. For proofreading my slides even when you don't understand anything. For housing me during various natural disasters (such as having no A/C in the Austin summer). For trying all kinds of restaurants and food with me. And so much more.

On Computationally Efficient Learning for Stabilizers and Beyond

by

Daniel You Liang, Ph.D.
The University of Texas at Austin, 2023

Supervisor: Scott Aaronson

Artificial intelligence, big data, machine learning, neural networks – look up any recent research proposal and with good probability at least one of these phrases will appear. It's no secret that learning has taken this era of computer science by storm in our attempt to create software that perform extremely complicated tasks. As one of the most accurate models of our physical world currently known, it then makes sense to think about what kinds of quantum systems can or cannot be learned. As with many problems in quantum information and quantum computing, the simplest non-trivial versions of these problems start with the stabilizer formalism. In this dissertation, we examine learning problems centered around the stabilizer formalism in various different models from a theoretical standpoint using the tools of computer science and quantum information. Specifically, our focus will be on computational complexity, rather than sample complexity.

We begin by looking at learning in the tomographical sense. Here, one has black-box access to copies of an unknown quantum state $|\psi\rangle$ and want to learn properties of the state or outright given an approximation of $|\psi\rangle$. In this setting, [Mon17] gave an efficient learning algorithm for stabilizer states. The key algorithmic tool was *Bell difference sampling*, which allows one to sample from the stabilizer group of a stabilizer state. [GNW21] extended the analysis of Bell difference sampling beyond just stabilizer states. Throughout Part I we turn to Bell difference sampling to improve upon learning algorithms for states with only

a few (i.e., either $O(\log n)$ or strictly less than n depending on context) T gates. By using symplectic Fourier analysis, which is the generalization of Boolean Fourier analysis for a symplectic vector space over $\mathbb{F}_2^{2n}$, we derive powerful tools to understand the Bell difference sampling distribution.

With these tools we first give a tolerant property testing algorithm for stabilizer states. That is, we give an algorithm that distinguishes whether a state is ε_1 close to some stabilizer state or ε_2 far from all stabilizer states for certain parameter regimes of ε_1 and ε_2. We use our improved knowledge of Bell difference sampling to improve upon the completeness and soundness analysis of the property tester given by [GNW21], which is not tolerant.

A second application is stabilizer fidelity estimation and approximation. Given a state $|\psi\rangle$ that is $O(1)$ close to a stabilizer state, we output such a stabilizer state in time $2^{O(n)}$. This beats the previous $2^{O(n^2)}$ brute force search algorithm. Having such a stabilizer state also lets us figure out how close $|\psi\rangle$ is to being stabilizer.

A third application is extending Montanaro's learning algorithm to the output of Clifford + $O(\log n)$ non-Clifford gate circuits. More generally, our algorithm interpolates between Montanaro's algorithm and pure state tomography algorithms with runtime that is $\mathsf{poly}(n) * \exp(t)$ where t is the number of non-Clifford gates. This asymptotically matches the runtime of classical simulation algorithms for such circuits. A key algorithmic step in this work is the ability to "compress" the "stabilizer-ness" of a state onto a few qubits, allowing the "non-stabilizer-ness" to be brute-forced on the remaining qubits.

Our final application is pseudorandomness lower bounds. Introduced by [JLS18], a pseudorandom quantum state ensemble is a set of quantum states that are computationally indistinguishable from Haar random. By re-purposing algorithms from above, we produce a test that behaves differently when given a state produced by less than n T gates in a Clifford + T circuit versus being given a Haar random state. We note that this is tight assuming the existence of linear-time quantum-secure One-Way Functions.

Pivoting now, we also study the stabilizer formalism in the PAC learning framework proposed by [Val84]. Here one does not have control over the measurements, but must make do regardless (within information theoretic limits). We analyze the problem in two ways. First we show that, unlike stabilizer states, learning the associated Clifford unitaries in the proper PAC model is NP-hard. This is done by a reduction to the problem of finding a full rank matrix in an affine subspace of matrices over $\mathbb{F}_2$.

The second is studying stabilizer states in the presence of noise. We utilize the Statistical Query framework, a popular modification to the PAC learning framework that is inherently tolerant to noise. There, we also show hardness in this framework by a reduction to Learning Parities with Noise. This gives evidence that even in the PAC model stabilizer states are hard to learn with noise.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

As a description of our physical world, quantum mechanics is seemingly without equal. While Einstein and other great minds famously initially rejected many of its quirks [EPR35], its immense predictive abilities combined with the acclaimed Bell test [FC72, ADR82, WJS+98, BPA22] has cemented it as a fixture of physics. In some aspects, the ability to predict physical phenomena is then inherently tied to making predictions about quantum mechanics. One such sub-problem in that area is learning unknown quantum systems – given access to measurement data of some form, can one identify the quantum system at play? And can one do it efficiently? Such problems will be the focus of this dissertation.

At its core, quantum is a reformulation of probability, by moving from non-negative real numbers to complex amplitudes to describe the system. For countable distributions, we go from these non-negative numbers adding to 1, to the magnitude of the complex amplitudes adding to 1 instead. The evolution of such probabilistic systems are classically described by *stochastic* processes, but to accommodate for quantum we must now use unitary operations to describe evolution instead. While seemingly simple and benign, the consequences are systems that are complex and strange relative to the classical ones we are used to thinking about. Perhaps the simplest demonstration of this is the Mach-Zehnder interferometer experiment [Ved06]. While normally described using beam splitters, it will suffice to think just about coin flips. Imagine if one were to place a fair coin heads-side up, flip it, and then cover the coin when it lands before anyone could see it. We could classically describe the coin as being 1/2 probability heads and 1/2 probability tails. Regardless of whether or not we look at the coin, if we were to re-flip the coin we would again arrive at an even probability of heads and tails.

However, it is possible to define a 'quantum' coin flip, which also initially produces an even probability distribution upon the first flip. The Mach-Zehnder interferometer experiment shows that for such a quantum coin flip, whether or not one looks at the coin affects the distribution after the second flip. If one were to look at the coin, thus *measuring* the state, one would again get another even distribution upon re-flipping. This isn't terribly surprising because it matches the classical case. The surprise happens when we decide not to look at the coin before we re-flip, where we always get heads! It becomes as if we never flipped the coin in the first place, so somehow the second coin flip is "cancelling out" the first coin flip in a way that is not possible under purely stochastic processes.

On the flip side of things (no pun intended), given the massive success of machine learning, big data, etc on many of today's most complex problems like Natural Language Processing, Computer Vision, and robotics, it is only natural to wonder how these tools fare when we assume the systems are inherently quantum? Indeed, quantum systems are complex and often require an exponential number of bits (in n, the number of qubits) to even classically describe a system using only n qubits (the quantum equivalent of a bit). Without access to high-fidelity programmable quantum devices (i.e., error-corrected quantum computers), experimental tests of such tools are currently out of our reach. However, it is still entirely possible to study the question from a theoretical standpoint, specifically related to the branches of tomography and learning theory. Using ideas of complexity theory and cryptography, it is possible to conjecture that certain learning problems should be intractable lest our understanding of computer science be unraveled to some degree. It is also possible to give learning algorithms in certain frameworks with provable guarantees, giving evidence that general tools of machine learning (or even *quantum* machine learning) could one day prove useful for scientists.

In this relatively nascent area of learning quantum systems, a very natural simplification to consider is that of the stabilizer formalism, which has a rich algebraic structure. The stabiliser formalism is remarkably interesting because it seemingly allows for all of the complex interactions that arise from quantum, such as superposition, entanglement, phase, etc. This allows it to play a central role in quantum error-correction, quantum money schemes, quantum key distribution, quantum algorithms, and more [Sho95, CS96, RB00, KLR+08]! In spite of this, it is also known to not be able to account for all of quantum mechanics. It is even conjectured that Clifford circuits can't even solve all deterministically solvable problems [AG04]! Because of this, Clifford circuits can actually be efficiently simulated classically [Got98], making Clifford circuits some of the most well understood classes of quantum circuits today.

To illustrate the simplicity of Clifford circuits, we start by introducing the following matrices:

$$H = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \quad S = \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix} \quad \text{CNOT} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

In terms of realizing the 3 aspects of quantum mechanics – superposition, complex phase, and entanglement – each of these 3 gates accomplishes one of those aspects respectively. It turns out that every Clifford circuit can be generated via a combination of these 3 gates applied to the qubits in various ways, showing that they are indeed highly quantum in some sense. The set of states that can be produced via a Clifford circuit starting from the all zeros state $|0^n\rangle$ are called stabilizer states. They're called that because they are stabilized by some unique abelian subgroup of Pauli matrices. Here Pauli matrices are the n-fold tensor of the following four matrices:

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \; X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \; Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \; Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

These Pauli matrices under multiplication act like a symplectic vector space over $\mathbb{F}_2^{2n}$. As such, much of our algorithms for simulating/learning related to the stabilizer formalism ultimately depends on doing linear algebra over $\mathbb{F}_2$ (proceeded by dealing with the other add-ons, like phase, appropriately).

Given all that, a very fundamental question for learning in the quantum domain is: Can we learn a stabilizer state (computationally efficiently)? Naïvely, if you have an exponential number of copies and the ability to perform arbitrary measurements you can perform full quantum state tomography [OW16, OW17, HHJ+17] (see Section 7.1 for a *slightly* more detailed discussion). But since we know that the state is a stabilizer state, maybe we can do better? It turns out that using only a linear number of copies and $O(n^3)$ time [Mon17], with high probability one can learn the generators of the stabilizer group of the stabilizer state, thus uniquely identifying it. A wildly different unpublished algorithm by Aaronson and Gottesman [Aar22] also exists but requires $O(n^2)$ samples (and the same $O(n^3)$ time).

Naturally an interesting followup question is what happens when we take these learning problems *beyond* stabilizer states? We present 3 natural generalizations of this learning problem – Clifford + T states, stabilizer states with noise, and Clifford circuits – that each serve as important building blocks to learning problems in the stabilizer formalism. The first we study in the tomographical sense, and the other two are studied in the statistical learning theory framework (see Chapter 9).

Part I: Clifford + T States and Bell Difference Sampling

Since learning states and circuits within the stabilizer formalism have been examined, the areas just beyond the formalism are also of particular interest. Of note, if one adds just one non-Clifford gate to the Clifford gateset of H, S, and CNOT then one can in fact achieve quantum universality. The most common of such gates is known as the T gate:

$$T = \begin{pmatrix} 1 & 0 \\ 0 & e^{i\pi/4} \end{pmatrix}$$

But what is the effect on the stabilizer formalism? We'll take a look at simulation as an example, since efficient classical simulation of Clifford + T circuits would imply $\mathsf{BPP} = \mathsf{BQP}$.

Using current state-of-the-art simulation algorithms, simulating a quantum circuit comprised of t different T gates requires time exponential in t. This is because the algorithms work first by gadgetizing the Clifford + T circuit into a regular Clifford circuit with each T gate replaced by an ancillary state $|+T\rangle = TH|0\rangle$ as part of the input. $|+T\rangle$ is then decomposed into a linear combination of stabilizer states and by the linearity of quantum mechanics, the two circuits should have equivalent results. The number of stabilizer states in the decomposition grows exponentially with t to the best of our knowledge [1] and thus bottlenecks the efficiency for large t. As such, T gates are often considered a resource of sorts, and the area where t is small is considered to be just outside of the stabilizer formalism.

Extending the ideas of these Clifford + T simulation algorithms, it is possible to try and define how hard an input state is to simulate. The most obvious is known as the *stabilizer rank*, which is just the minimal number of states in any stabilizer decomposition. However, both this and the approximate stabilizer rank (i.e., minimal number of states to get somewhere close to the state) can be very hard to analyze. Some easier to work with measures include *stabilizer extent* and *stabilizer fidelity*. Given a state $|\psi\rangle$, its stabilizer extent $\xi(|\psi\rangle)$ is the minimal $\left(\sum_i c_i\right)^2$ such that $|\psi\rangle = \sum_i c_i |\phi_i\rangle$ where the $|\phi_i\rangle$ are all stabilizer states. A more intuitive definition is that it is the square of the minimal 1-norm, whereas stabilizer rank is the minimized 0-norm. The stabilizer fidelity $F_{\mathcal{S}}(|\psi\rangle$ is even simpler and is the maximal fidelity of $|\psi\rangle$ to any stabilizer state (i.e., $\arg\max_i |\langle\psi|\phi_i\rangle|^2$ where $|\phi_i\rangle$ are the stabilizer states). A very nice result by Bravyi et al. [BBC+19] is that for any state, the product of its stabilizer extent and stabilizer fidelity $\xi(|\psi\rangle) \cdot F_{\mathcal{S}}(|\psi\rangle) \leq 1$, casting them as somewhat of a dual to one another. In broad strokes, the runtime of classical simulation algorithms is polynomially related (or inversely related in the case of stabilizer fidelity) to stabilizer rank/extent/fidelity.

[1] And it is an important open problem in quantum information to show whether or not an exponential number is actually required.

Collectively, stabilizer rank, stabilizer extent, and stabilizer fidelity can be seen as a set of related *stabilizer complexity measures.* A final complexity measure used in this dissertation takes a different tact and tracks how much of the algebraic structure is preserved. Since stabilizer states are stabilized by an abelian subgroup of 2^n Pauli matrices, the *stabilizer dimension* of a state then corresponds to how many Pauli matrices stabilize an arbitrary quantum state.

Because much of the nice algebraic structure disappears once we start introducing T gates, learning in this area has largely been slow. The only notable example up until now was [LC22], which was able to give an algorithm for learning states produced by a very limited set of Clifford + T circuits[2]. Notably, their approach used *Bell difference sampling*, which was the same technique used in [Mon17]. Other results that used this (or similar techniques) include [GNW21], [HK23], and [HCP22].

To try and give results for general Clifford + T states, we greatly improve our understanding of Bell difference sampling. In broad strokes, Bell difference sampling is a distribution over $\mathbb{F}_2^{2n}$ produced by measuring 4 copies of a quantum state in the Bell basis, then taking a convolution (see Section 3.2 for details). Let q_ψ be the distribution produced by Bell difference sampling on copies of $|\psi\rangle$. At a high level, we establish our results by proving some structure on q_ψ for certain quantum states. For example, to prove our lower bound on the number of non-Clifford gates required to prepare pseudorandom states, we give an algorithm that distinguishes Haar-random states from quantum states prepared by circuits with fewer than $n/2$ non-Clifford single-qubit gates (Chapter 8). The key insight is that if $|\psi\rangle$ is the output of such a circuit, then q_ψ is concentrated on a proper subspace of $\mathbb{F}_2^{2n}$, whereas for Haar-random states, q_ψ is anticoncentrated on all such subspaces with overwhelming probability over the Haar measure. Proving these properties of q_ψ reveals a simple algorithm: draw a linear number of samples from q_ψ and compute the number of

[2] See Section 7.1.2 for *slightly* more details.

linearly independent vectors in the sample. Haar-random states will have $2n$ such vectors with high probability and, otherwise, there will be strictly less than $2n$ such vectors.

In this dissertation, we start by introducing Bell difference sampling in Chapter 3, capturing what was known prior to [GIKL23c][3]. The chapter ends with an explanation and proof of Montanaro's learning algorithm. We then present our new findings about q_ψ in Chapter 4, using techniques of symplectic Fourier analysis.

From there, we turn those findings into algorithmic results:

Chapter 5 Tolerant Testing of Stabilizer States

Gross et al. [GNW21] were able to come up with a Bell-difference-sampling-based test that distinguished stabilizer states from those far from stabilizer states. While only using 6 copies and linear time, the test has perfect completeness (i.e., always accepts for stabilizer states) and was shown to accept with probability no more than $\frac{1}{4}\left(F_{\mathcal{S}}(|\psi\rangle) + 3\right)$. We improve this result by giving a tighter characterization of the acceptance probability. The proof involves a generalization of Boolean Fourier analytic techniques similar to the BLR test The completeness is improved by showing that the acceptance probability is at least $\frac{1}{2}\left(F_{\mathcal{S}}(|\psi\rangle)^6 + 1\right)$ [GIKL23c] and the sixth power on stabilizer fidelity is a direct consequence of the six copies used in the test. We also give an improved soundness analysis, showing that the acceptance probability is at most $\frac{1}{8}\left(3F_{\mathcal{S}}(|\psi\rangle) + 5\right)$ [GIKL23b]. This is because our tighter character also allows us to use a tighter form of Markov's inequality.

Chapter 6 Stabilizer State Approximations and Stabilizer Fidelity Estimation

Another application of Bell difference sampling is to approximate an arbitrary quantum state with a stabilizer state (i.e., the one that realizes $F_{\mathcal{S}}(|\psi\rangle)$). The algorithm involves using Bell difference samples to (implicitly) build a list of possible stabilizer states and

[3] The (chronologically) first work that appears in Part I

sampling until the fidelity maximizing stabilizer state is in the list with high probability. Then by running a search algorithm to explicitly construct the states (via maximal clique enumeration [TTT06]) and using classical shadows [HKP20] we are able to find said maximizing stabilizer state. The number of samples is $O(n/F_{\mathcal{S}}(|\psi\rangle)^4)$ and the time complexity is $\exp(O(n/F_{\mathcal{S}}(|\psi\rangle)^4))$. For constant stabilizer fidelity, this is an super-polynomial improvement over the previous brute-force search of $\exp(O(n^2))$.

Chapter 7 Efficient Tomography of $O(\log n)$ T gate states

Our final unique algorithm returns to the task of tomography. Here, given a state $|\psi\rangle$ produced by no more than t T gates (or any t single-qubit non-Clifford gates for that matter), we are able to output a succinct classical approximation of $|\psi\rangle$ using $O(n) * \exp(t)$ samples and $O(n^3) * \exp(t)$ time. This significantly improves upon [LC22] as we have no restrictions on where the T gates are placed, what the state looks like before the T gates are applied, as well as can replace the T gates with arbitrary single-qubit non-Clifford. The algorithm starts by using Bell difference samples to learn $\text{Weyl}(|\psi\rangle)$. This allows us to map $|\psi\rangle$ to a state that looks like $|\psi'\rangle \otimes |0^{n-2t}\rangle$, thereby "compressing" the stabilizer-ness into the last $n - 2t$ qubits. From there, bruteforce pure state tomography is done on $|\psi'\rangle$ to complete the description. The algorithm itself can be seen as a major generalization of Montanaro's algorithm [Mon17], in that the first step is a more general form of it.

Chapter 8 T gate lower bounds for pseudorandom quantum states

Cryptography and learning are often seen as dual problems, in that things that are hard to learn can be used for cryptography and things that are cryptographically hard should be impossible to learn. In that sense, we examine computationally pseudorandom quantum states as introduced by [JLS18]. These are an ensemble of quantum states such that no polynomial time quantum algorithm can distinguish between a uniform distribution over said ensemble or a Haar random state. Since Clifford + T is a common universal gateset where the T gate is considerably more expensive than the other operations, we wish to characterize how many T gates are necessary to produce pseudorandom quantum states. We give two such lower bounds based on different measures of how far a state is from stabilizer: stabilizer fidelity and stabilizer dimension. These result in lower bounds of $\omega(\log n)$ and $\Omega(n)$ respectively. Furthermore, the two bounds are asymptotically tight (in terms of their respective stabilizer complexity measure) if linear-time quantum-secure one way functions exist. The distinguishing algorithms are actually just special cases of ones introduced in Chapters 5 and 7 respectively, with soundness analysis changed to be specifically for Haar random states.

Part II Statistical Learning Theory for Quantum Systems

Taking some of these ideas even further, one can try to learn in a scenario where one does not get to choose the measurements, also known as the PAC model [Val84, Aar07]. Here one does not try to learn in an absolute sense, like tomography, but in a statistical sense where the idea is to give good predictions with high probability relative to some distribution. More concretely, say there exists some unknown quantum state ρ that one is trying to learn. You are given access to an oracle that draws a measurement E from some (potentially unknown) distribution $\mathcal{D}$ and then returns to you the training datum $(E, \mathrm{Tr}[E\rho])$. The goal is then to output a hypothesis state σ such that the loss function $\Delta(\sigma, \rho)$ is small. An example of a simple and popular loss function is the squared loss, i.e., $\Delta(\sigma, \rho) = \mathbf{E}_{E\sim\mathcal{D}}\left[(\mathrm{Tr}[E\rho] - \mathrm{Tr}[E\sigma])^2\right]$. Rocchetto [Roc18] was able to develop an efficient distribution-agnostic PAC learning algorithm with Pauli measurements for stabilizer states. Here distribution-agnostic means that the learning algorithm does not need to know the input distribution in advance. The key idea was using the aforementioned relationship of Paulis to symplectic vector spaces over $\mathbb{F}_2$, creating an algorithm that achieved zero training loss and combined it with the generalization theorem given by Aaronson [Aar07].

Chapter 11: Statistical Queries for Quantum States

Continuing with the PAC learning model, it is possible to wonder what happens when noise is added. Now what does noise exactly mean here? In the standard PAC learning model of Boolean functions, noise simply means that the label is flipped with some probability. But since we're dealing with quantum states, another fair question is what happens when quantum noise is applied to the state first, such as depolarizing noise?

Classically, one of the main ways to PAC learn in the presence of noise is to use the Statistical Query (SQ) formalism [Kea98]. Here, rather than look at individual samples, the learner chooses a function $g : \mathcal{X} \times \{0,1\} \to [0,1]$ to give to the oracle. This function acts on the input-label pair $(x, f(x))$ and the oracle then returns the expectation $\mathbf{E}\left[g(x, f(x))\right]$ of such a function up to some error. While not all PAC learning algorithms that deal with noise lead to an SQ algorithm, all SQ algorithms imply a PAC learner that is able to resist noise via a simple Chernoff bound argument. As such, due to the error tolerance built-in to the SQ algorithm, it can be easily shown that very benign noises can be accommodated in the SQ algorithm. More interestingly, it is possible for SQ algorithms to systematically correct for larger errors. To give a small example of this, imagine that the error rate of bit-flips on the label is known. Then the expectation of the function we are expecting and the expectation of the function after it has gone through the noisy channel can be related to one another, thus allowing for correction.

In an attempt to partially answer this question of noisy learning of quantum states, we introduced the SQ model for learning quantum states [GL22]. Specifically, we cover the more difficult *probabilistic* concept (p-concept) model. By a p-concept, we mean that rather than receiving $\mathrm{Tr}[E\rho] \in [0,1]$ as a label, one receives a $\{0,1\}$ label with expectation $\mathrm{Tr}[E\rho]$. Thus, the effect of noise is altering the Bernoulli distribution associated with $\mathrm{Tr}[E\rho]$.

In this model, we were able to build on [AD98] and [Kea98] to show that SQ algorithms for general quantum states could systematically correct both classification and malicious classical noise, within information theoretic reasons. We additionally showed systematic correction for global depolarizing noise, a common form of quantum noise. Utilizing the work of Goel et al. [GGJ$^+$20] that introduced information theoretic lower bounds on learnability via the SQ-dimension, in [GL22] we give information theoretic lower bounds for learning stabilizer states under a uniform distribution of Pauli measurements. We also improve this lowerbound for a restricted subset of Pauli measurements via a reduction to learning parity functions, a problem that is provably hard to SQ learn [Kea98]. Finally, we give an efficient SQ algorithm for learning relative to uniform rank-1 product measurements, as well as give some connections to differential privacy [DR14].

Chapter 10: PAC Learning Clifford Circuits

Another interesting line of work is what happens when the object we are trying to learn is a circuit, rather than just a state? In the scenarios of absolute learning where one has the ability to apply the unitary C (and also $C^\dagger$), Low [Low09] was able to devise an algorithm in $O(n^2)$ time. The need for $C^\dagger$ was later removed by Lai and Cheng[LC22] with an algorithm that runs in $O(n^3)$.

The question of what happens in the PAC setting was then raised. One possible definition for PAC learning Clifford circuits is to get as input, a pair $(|\phi\rangle, E)$, where $|\phi\rangle$ is a stabilizer state and E is a Pauli matrix, resulting in a natural generalization of [Roc18]. The results of [CD20] can be modified[4] to give a generalization result for quantum circuits analogous to [Aar07]. This leads one to expect that a PAC learning algorithm for Clifford circuits should be possible just as with [Roc18] for stabilizer states. Contrary to popular intuition, it turns out that in the *proper* learning case, this possible if and only if $\mathsf{NP} \subseteq \mathsf{BPP}$

[4] For a proof of this see Appendix A

[Lia23][5]. By proper learning, the hypothesis is constrained to be a member of the concept class, which in this case is the set of all Clifford circuits. Note that the algorithm in [Roc18] was also a proper learner.

The hardness result follows from showing that even deciding if there exists a hypothesis that produces zero error, also known as the *consistency* problem, is NP-Hard. This even holds for the subset of Clifford circuits that only utilize CNOT gates, creating a classical reversible circuit. Such circuits are called CNOT circuits. Here, each output bit is some parity function of the input bits, meaning that the whole circuit can be expressed as a linear transformation of the input bits over $\mathbb{F}_2$. Each sample that comes in constrains said matrix to some affine subspace, meaning that all consistent matrices lie in the intersection, which is just another affine subspace. The catch is that because the circuit is reversible, the matrix must also be full rank, and indeed for every full rank matrix there exists a corresponding CNOT circuit. This problem of finding a full rank matrix in some affine subspace is known as the NonSingularity problem over $\mathbb{F}_2$ and was shown to be NP-complete [BFS99]. Even moreso, we were able to show that the hardest instances of the NonSingularity problem over $\mathbb{F}_2$ (i.e., the ones that come from the original Buss et al. [BFS99] hardness reduction) can also formed as constraints arising from deciding consistency of CNOT circuits. Thus, the consistency problem for CNOT circuits is NP-hard, and simple reductions show that distribution-agnostic PAC learning of CNOT circuits and Clifford circuits must also be NP-hard, covering the more interesting 'only if' part of the statement.

To show the 'if' part, we note that deciding consistency is NP-complete. Since Clifford circuits can be encoded using only a polynomial number of classical bits, they can be used as a witness and the consistency can be classically verified by running Gottesman-Knill style simulation algorithms [AG04]. Search-to-decision reductions of NP-complete problems combined with the generalization result of [CD20] then give an efficient randomized/quantum

[5] This is an equivalent statement to $\mathsf{RP} = \mathsf{NP}$, which was the original hardness statement.

algorithm for PAC learning Clifford circuits should $\mathsf{NP} \subseteq \mathsf{BPP}$ or $\mathsf{NP} \subseteq \mathsf{BQP}$ respectively be true.

Finally, in the non-distribution agnostic case we give some efficient learning algorithms for Clifford and CNOT circuits for various distributions such as uniform, or where either the state of the measurement values are fixed.

Chapter 2

Quantum Information Preliminaries

This chapter is a review of quantum information and the stabilizer formalism. It borrows heavily from [Lia23], [GIKL23c], and [GIKL23b].

We briefly introduce general notation and convention for quantum information before diving into the stabilizer formalism, which is one of the central themes of this dissertation. For more background on the stabilizer formalism, see, e.g., [Got97, NC02].

2.1 Basic Preliminaries

We use log as the natural logarithm and $\log_2$ as the logarithm base 2. We state two concentration inequalities that we use in this work.

Fact 2.1 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be independent identically distributed random variables such that $a_i \leq X_i \leq b_i$. Let X denote their sum and let $\mu \coloneqq \mathsf{E}[X]$. Then for any $\delta > 0$,*

$$\mathbf{Pr}\left[X \leq \mu - \delta\right] \leq \exp\left(-\frac{2\delta^2\mu}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Fact 2.2 (Multiplicative Chernoff bound)**.** *Let $X_1, \ldots, X_n$ be independent identically distributed random variables taking values in $\{0, 1\}$. Let X denote their sum and let $\mu \coloneqq \mathsf{E}[X]$. Then for any $\delta > 0$,*

$$\mathbf{Pr}\left[X \leq (1-\delta)\mu\right] \leq e^{-\delta^2/2}.$$

2.2 Quantum States and Circuits

A quantum state ρ on n qubits is a $2^n \times 2^n$ PSD matrix with trace 1 If the matrix is rank 1 then we refer to ρ being a PURE STATE, since it can be decomposed as $\rho = |\psi\rangle\langle\psi|$ where $|\psi\rangle$ is a 2^n-dimensional column vector with norm 1 and $\langle\psi|$ is its complex conjugate. Oftentimes, we will simply refer to a quantum pure state as $|\psi\rangle$, noting that $e^{i\theta}|\psi\rangle$ refers to the same effective state $|\psi\rangle\langle\psi|$ (i.e., global phase does not matter).

An observable is simply a Hermitian matrix. By the spectral theorem, any Hermitian matrix H can be viewed as $H = \sum_i \lambda_i |\lambda_i\rangle\langle\lambda_i|$ where the $|\lambda_i\rangle$ are a complete set of orthonormal states i.e., a measurement basis. We can think of the $|\lambda_i\rangle\langle\lambda_i|$ as the projector describing the probability of measuring $|\lambda_i\rangle$. Measuring in this basis and returning λ_i for the corresponding eigenspace gives rise to a random variable whose expectation is $\mathrm{Tr}[H\rho]$. If $\rho = |\psi\rangle\langle\psi|$ is a pure state, we will alternatively write this as $\mathrm{Tr}[H\rho] = \langle\psi|H|\psi\rangle$ due to the cyclic property of the trace. A two-outcome measurement E is then a projector such that $E^2 = E$ such that the probability of a '1' outcome is $\mathrm{Tr}[E\rho]$ and the probability of a '0' outcome is $1 - \mathrm{Tr}[E\rho]$, leaving the expectation value as simply $\mathrm{Tr}[E\rho]$.

A quantum process is how one evolves a quantum state, and therefore it must preserve the trace 1 and the PSD condition. We will be primarily interested in quantum circuits, which are the subset of quantum processes that map pure states only to other pure states. These are constrained to be unitary operations, such that after acting on ρ with the circuit C, the state that we are left with is $C\rho C^\dagger$ where $C^\dagger$ is the complex conjugate of C.

Our most common measure of distance between two quantum states will be fidelity and trace distance.

Definition 2.3 (Fidelity). *Given quantum states ρ and σ, the fidelity between ρ and σ is*

$$F(\rho, \sigma) = \left(\mathrm{Tr}\left[\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}\right]\right)^2.$$

Alternatively, if $\rho = |\psi\rangle\langle\psi|$ is a pure state (as will be common for us) then the expression simplifies to

$$F(|\psi\rangle\langle\psi| , \sigma) \coloneqq \langle\psi|\sigma|\psi\rangle .$$

While the fidelity between two pure states is a very natural way to compare how close two states are, it is not a metric. For that, we will introduce the trace distance, which is in fact a metric.

To define the trace distance, we will first define the trace norm.

Definition 2.4 (Trace norm)**.** *For a Hermitian matrix A, let $\sum_i \lambda_i |\lambda_i\rangle\langle\lambda_i| = A$ be its spectral decomposition. The trace norm is defined to be*

$$\|A\|_{\mathrm{Tr}} \coloneqq \sum_i |\lambda_i|.$$

The trace distance simply arises from the trace norm as one would expect, but with a scale factor added to keep the value in $[0, 1]$.

Definition 2.5 (Trace distance)**.** *Given quantum states ρ and σ,*

$$\mathrm{TD}(\rho, \sigma) \coloneqq \frac{1}{2}\|\rho - \sigma\|_{\mathrm{Tr}}.$$

The trace distance can be seen as the quantum generalization of the total variation distance between distributions. As such, like total variation distance it exhibits a useful property of subadditivity with respect to tensor products.

Fact 2.6 (Subadditivity of trace distance)**.**

$$\mathrm{TD}(\rho \otimes \rho', \sigma \otimes \sigma') \leq \mathrm{TD}(\rho, \sigma) + \mathrm{TD}(\rho', \sigma')$$

Importantly, both fidelity and trace distance are unitarily invariant, in that for arbitrary unitary U, $F(U\rho U^\dagger, U\sigma U^\dagger) = F(\rho, \sigma)$ and $\mathrm{TD}(U\rho U^\dagger, U\sigma U^\dagger) = \mathrm{TD}(\rho, \sigma)$.

We can also relate trace distance to fidelity with the following well-known fact (see [NC02], noting the difference in square in the definition of Fidelity).

Fact 2.7 ([NC02] Eq 9.110)**.**

$$1 - \sqrt{F(\rho, \sigma)} \leq \mathrm{TD}(\rho, \sigma) \leq \sqrt{1 - F(\rho, \sigma)}$$

and the upper bound is an equality if both ρ and σ are pure states.

Since we are commonly dealing with pure states in this work, this tight upper bound will be very useful. For instance, we will sometimes want to convert from fidelity to trace distance, to make use of properties like the triangle inequality or subadditivity, then convert back to fidelity.

2.3 Symplectic Vector Spaces

We work extensively with $\mathbb{F}_2^{2n}$ as a symplectic vector space by equipping it with the symplectic product.

Definition 2.8 (Symplectic product)**.** *For $x, y \in \mathbb{F}_2^{2n}$, we define the* SYMPLECTIC PRODUCT *as $[x, y] = x_1 \cdot y_{n+1} + x_2 \cdot y_{n+2} + \ldots + x_n \cdot y_{2n} + x_{n+1} \cdot y_1 + x_{n+2} \cdot y_2 + \ldots + x_{2n} \cdot y_n$, where all operations are performed over $\mathbb{F}_2$.*

We can alternatively view x and y as vectors, such that $[x, y] = x^T \Lambda(n) y$ where

$$\Lambda(n) := \begin{bmatrix} 0 & I^{\otimes n} \\ I^{\otimes n} & 0 \end{bmatrix}$$

is a $2n \times 2n$ matrix.

The symplectic product gives rise to the notion of a *symplectic complement*, much like the orthogonal complement for the standard inner product modulo 2.

Definition 2.9 (Symplectic complement)**.** *Let $T \subseteq \mathbb{F}_2^{2n}$ be a subspace. The* SYMPLECTIC COMPLEMENT *of T, denoted by $T^\perp$, is defined by*

$$T^\perp := \{a \in \mathbb{F}_2^{2n} : \forall x \in T,\ [x, a] = 0\}.$$

We present the following useful facts about the symplectic complement, many of which are similar to that of the more familiar orthogonal complement.

Fact 2.10. *Let S and T be subspaces of $\mathbb{F}_2^{2n}$. Then:*

- *$T^\perp$ is a subspace.*
- *$(T^\perp)^\perp = T$.*
- *$|T| \cdot |T^\perp| = 4^n$, or equivalently $\dim T + \dim T^\perp = 2n$.*
- *$T \subseteq S \iff S^\perp \subseteq T^\perp$.*

A subspace $T \subset \mathbb{F}_2^{2n}$ is *isotropic* when for all $x, y \in T$, $[x, y] = 0$. Equivalently, T is isotropic if and only if $T \subseteq T^\perp$. A subspace $T \subset \mathbb{F}_2^{2n}$ is *Lagrangian* when $T^\perp = T$. Finally, a subspace $T \subseteq \mathbb{F}_2^{2n}$ is *coisotropic* if its symplectic complement is isotropic.

2.4 Pauli Group and Weyl Matrices

We will start by giving the following matrices, known as the Pauli matrices.

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \quad Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

Noting that these are all unitaries that act on a single qubit, we can generalize to n qubits.

Definition 2.11. *Let $\mathcal{P}_n = \{\pm 1, \pm i\} \times \{I, X, Y, Z\}^{\otimes n}$ be the matrix group consisting all n-qubit Paulis with phase ± 1 or $\pm i$.*

We refer $\mathcal{P}_n$ as the Pauli group, while the individual elements of $\mathcal{P}_n$ are interchangeably referred to as Pauli operators or n-qubit Pauli matrices throughout the dissertation.

We'll also introduce some shorthand notation:

Definition 2.12. *Let X_i and Z_i be the Pauli operator acting only on the i-th qubit with X or Z respectively and the identity matrix on all other qubits.*

Definition 2.13. *For $v \in \mathbb{F}_2^n$, let $X^v = \prod_{i=1}^n X_i^{v'_i}$ and $Z^v = \prod_{i=1}^n Z_i^{v'_i}$ respectively where v' is the representation of v as the n-bit string in $\{0,1\}^n$.*

While the phase part (i.e., $\{\pm 1, \pm i\}$) of the definition is necessary to make the Pauli operators a group, much of the interesting parts have nothing to do with it. Indeed, many times we will want to ignore the phase in front of the matrix and talk about the structure of the Pauli matrices. Note that $Z^v \cdot Z^w = Z^{v+w}$, assuming the dimensions of v and w match. It is easy to see that $v \neq w$ also implies that $Z^v \neq Z^w$. As such, we introduce the following matrices to formalize this idea.

Definition 2.14. *For $x = (a,b) \in \mathbb{F}_2^{2n}$, the* Weyl operator *$W_x$ is defined as*

$$W_x := i^{a' \cdot b'} X^a Z^b,$$

where $a', b' \in \mathbb{Z}_2^n$ are the embeddings of a, b into $\mathbb{Z}_2^n$ respectively.

Each Weyl operator is a Pauli operator, and every Pauli operator is a Weyl operator up to a phase. There is clearly a bijection between $\mathbb{F}_2^{2n}$ and the set of Weyl operators that we will use to freely go between the two. Importantly, commutation relations between Weyl operators (and the Pauli operators they represent up to phase) are determined by the symplectic product. In particular, for $x, y \in \mathbb{F}_2^{2n}$, the Weyl operators W_x, W_y commute when $[x,y] = 0$ and anticommute when $[x,y] = 1$. A useful identity that represents this idea is the following:

Fact 2.15.

$$W_x W_y W_x = (-1)^{[x,y]} W_y$$

So, if $T \subseteq \mathbb{F}_2^{2n}$ is a subspace, then T is isotropic if and only if $\{W_x : x \in T\}$ is a set of mutually commuting Weyl operators. Similarly, T is Lagrangian if and only if $\{W_x : x \in T\}$ is a set of 2^n mutually commuting Weyl operators.

An important fact is that all of the Weyl operators besides $W_{0^n} = I^{\otimes n}$ have trace zero. Another important fact is that for all $x \in \mathbb{F}_2^{2n}$, $W_x^2 = I^{\otimes n}$. In general, we find that $\mathrm{Tr}[W_x W_y] = 2^n 1_{x=y}$. As a result, the Weyl operators collectively form an orthogonal basis for $2^n \times 2^n$ matrices with respect to the Frobenius inner product $\langle A, B\rangle = \mathrm{Tr}[A^\dagger B]$. This gives rise to the so-called Weyl expansion of a matrix. We give the special case for Hermitian matrices.

Definition 2.16 (Weyl expansion)**.** *Let H be a Hermitian matrix. The Weyl expansion of H is*

$$H = \frac{1}{\sqrt{2^n}} \sum_{x \in F_2^{2n}} c_H(x) W_x,$$

where $c_H(x) = \frac{1}{\sqrt{2^n}} \mathrm{Tr}[W_x H] \in \mathbb{R}$.

Due to the structure of the Weyl operators and the fact that quantum states are represented as Hermitian matrices, we can view this as a form of Fourier expansion for quantum states. We now give a version of Plancherel's theorem for the Weyl expansion.

Lemma 2.17 (Weyl Operator Plancherel's theorem)**.** *For Hermitian matrices A and B,*

$$\mathrm{Tr}[AB] = \sum_{x \in \mathbb{F}_2^{2n}} c_A(x) c_B(x).$$

Proof.

$$\begin{aligned}\mathrm{Tr}[AB] &= \mathrm{Tr}\left[\left(\frac{1}{\sqrt{2^n}}\sum_{x\in F_2^{2n}} c_A(x)W_x\right)\left(\frac{1}{\sqrt{2^n}}\sum_{y\in F_2^{2n}} c_B(y)W_y\right)\right]\\ &= \frac{1}{2^n}\sum_{x,y\in\mathbb{F}_2^{2n}} c_A(x)c_B(y)\mathrm{Tr}[W_xW_y]\\ &= \frac{1}{2^n}\sum_{x,y\in\mathbb{F}_2^{2n}} c_A(x)c_B(y)2^n 1_{x=y}\\ &= \sum_{x\in\mathbb{F}_2^{2n}} c_A(x)c_B(x) \qquad \square\end{aligned}$$

Since quantum states are Hermitian matrices, every quantum state has a Weyl expansion. A useful application of Lemma 2.17 is that for a pure state $|\psi\rangle\langle\psi|$, the $c_\psi(x)^2$ must sum to 1. This can be seen as the analogue of Parseval's identity for quantum pure states.

Corollary 2.18 (Quantum Pure State Parseval's identity)**.** *For a pure state* $|\psi\rangle$*,*

$$\sum_{x\in\mathbb{F}_2^{2n}} c_\psi(x)^2 = 1.$$

Proof.

$$\begin{aligned}\sum_{x\in\mathbb{F}_2^{2n}} c_\psi(x)^2 &= \mathrm{Tr}[|\psi\rangle\langle\psi|^2] && \text{(Lemma 2.17)}\\ &= \mathrm{Tr}[|\psi\rangle\langle\psi|]\\ &= 1\end{aligned}$$

The second line follows from the fact that $|\psi\rangle\langle\psi|$ is a projector, so $|\psi\rangle\langle\psi|^2 = |\psi\rangle\langle\psi|$. □

Another way of seeing this is that the square of the $c_\psi(x)$'s gives rise to a distribution over $\mathbb{F}_2^{2n}$ and therefore over the Weyl operators. This will become extremely useful in Part I.

2.5 Stabilizer Groups and States

Let ρ be a state that can be written as $\frac{1}{2^n}\sum_{\sigma\in G}\sigma$, where $G \subset \mathcal{P}_n \setminus \{-I^{\otimes n}\}$ is an abelian subgroup without the negative identity. As it turns out, if G is of order 2^n then ρ will be a pure state. A STABILIZER STATE is then a pure state that can be written as $\frac{1}{2^n}\sum_{\sigma\in G}\sigma$, where $G \subset \mathcal{P}_n \setminus \{-I^{\otimes n}\}$ is an abelian subgroup of 2^n Pauli operators without the negative identity. G is known as the STABILIZER GROUP of ρ. We denote the set of n-qubit stabilizer pures states by $\mathcal{S}_n$. There is also the alternative (and more popular definition) where $|\psi\rangle$, is the unique state that is stabilized by G. That is, for all $g \in G$, $g|\psi\rangle = |\psi\rangle$. This definition shows why $-I^{\otimes n}$ isn't allowed to be in G, since $-I^{\otimes n}$ stabilizes nothing. It also shows why one must restrict the entries of G to only have real phase.

Proposition 2.19. *Any abelian subgroup of $G \subseteq \mathcal{P}_n \setminus \{-I^{\otimes n}\}$ cannot contain any Paulis with an imaginary phase.*

Proof. Given a Pauli with an imaginary phase, its square would be equal to $-I^{\otimes n}$, making the group not closed. This is a contradiction. □

As a consequence, we will often only want to consider the Paulis with real phase, $\mathcal{P}_n^{\pm} := \{\pm 1\} \times \{I, X, Y, Z\}^{\otimes n}$.

One of the reasons stabilizer states are so important is this bijection between the stabilizer group of a stabilizer state and the state itself; by simply knowing the generators of the group one can easily reconstruct the state. And since there are at most n generators, if one can efficiently write down the generators themselves then there is a polynomial size representation of a stabilizer state. We note that the Weyl operators allow us to efficiently write down generators of a stabilizer group, since we can write it as $\pm W_x$. Keeping track of the phase and storing $x \in \mathbb{F}_2^{2n}$ as a $2n$-bit string results in a $2n+1$ bit string for each generator. As a result, writing down a stabilizer state requires only $\mathcal{O}(n^2)$ bits to write down classically. This is generally known as the Aaronson-Gottesman tableau [AG04]. It immediately gives a $2^{O(n^2)}$ upper bound on the number of stabilizer states. This turns out to be tight.

Proposition 2.20 ([AG04] Proposition 2)**.** *The number of n-qubit stabilizer states grows as $2^{\Theta(n^2)}$.*

We now introduce a bit of notation that will be used extensively throughout the proofs of this dissertation.

Definition 2.21 (Stabilizer group)**.** *Let $\mathit{Stab}(|\psi\rangle) \coloneqq \{\sigma \in \mathcal{P}_n^{\pm} : \sigma |\psi\rangle = |\psi\rangle\}$ denote the stabilizer group of $|\psi\rangle$.*

Definition 2.22 (Unsigned stabilizer group)**.** *Let $\mathrm{Weyl}(|\psi\rangle) \coloneqq \{x \in \mathbb{F}_2^{2n} : W_x |\psi\rangle = \pm |\psi\rangle\}$ denote the unsigned stabilizer group of $|\psi\rangle$.*

By standard linear algebra, we can show that $\mathrm{Weyl}(|\psi\rangle)$ is closed under addition.

Fact 2.23. *$\mathrm{Weyl}(|\psi\rangle)$ is a subspace of $\mathbb{F}_2^{2n}$.*

Proof. We note that if $x, y \in \mathrm{Weyl}(|\psi\rangle)$ then $x + y \in \mathrm{Weyl}(|\psi\rangle)$ since

$$|\langle\psi|W_{x+y}|\psi\rangle|^2 = |\langle\psi|W_x W_y|\psi\rangle|^2 = 1,$$

where the last equality comes from the definition of $x, y \in \mathrm{Weyl}(|\psi\rangle)$. This shows that $\mathrm{Weyl}(|\psi\rangle)$ is a subspace. □

We now show that, as a consequence of the uncertainty principle, $\mathrm{Weyl}(|\psi\rangle)$ is isotropic. [1]

[1] In 1927, Heisenberg observed a tradeoff between knowing a particle's position and momentum, which has since been generalized in several ways. This particular uncertainty relation was derived by Schrödinger in 1930.

Fact 2.24 (Schrödinger uncertainty relation [Sch30, AB08])**.** *For Hermitian observables A and B,*

$$\left(\langle\psi|A^2|\psi\rangle - \langle\psi|A|\psi\rangle^2\right)\left(\langle\psi|B^2|\psi\rangle - \langle\psi|B|\psi\rangle^2\right) \geq \left|\frac{1}{2}\langle\psi|(AB+BA)|\psi\rangle - \langle\psi|A|\psi\rangle\langle\psi|B|\psi\rangle\right|^2$$

Fact 2.25. *Let $M = \{x \in \mathbb{F}_2^{2n} : \langle\psi|W_x|\psi\rangle^2 > \frac{1}{2}\}$. Then for all $x, y \in M$, $[x,y] = 0$.*

Proof. Since the Weyl operators are Hermitian, they can be seen as observables. Let W_x and W_y be two Weyl operators in M. Simple calculations tell us that

$$\left(\langle\psi|W_x^2|\psi\rangle - \langle\psi|W_x|\psi\rangle^2\right)\left(\langle\psi|W_y^2|\psi\rangle - \langle\psi|W_y|\psi\rangle^2\right) = \left(1 - \langle\psi|W_x|\psi\rangle^2\right)\left(1 - \langle\psi|W_y|\psi\rangle^2\right) < \frac{1}{4}$$

where we utilize the fact that for all $x \in \mathbb{F}_2^{2n}$, $W_x^2 = I^{\otimes n}$. Now, assume that $[x,y] = 1$ for the sake of contradiction. We would find that $W_xW_y + W_yW_x = 0$ such that

$$\left|\frac{1}{2}\langle\psi|\left(W_xW_y + W_yW_x\right)|\psi\rangle - \langle\psi|W_x|\psi\rangle\langle\psi|W_y|\psi\rangle\right|^2 = |\langle\psi|W_x|\psi\rangle\langle\psi|W_y|\psi\rangle|^2 > \frac{1}{4}.$$

By Fact 2.24, this is a contradiction so $[x,y]$ must be zero instead. □

Corollary 2.26. Weyl$(|\psi\rangle)$ *is an isotropic subspace of* $\mathbb{F}_2^{2n}$*.*

Proof. The subspace part is dealt with by Fact 2.23. We note that Weyl$(|\psi\rangle) \subseteq \{x \in \mathbb{F}_2^{2n} : \langle\psi|W_x|\psi\rangle^2 > \frac{1}{2}\}$. By Fact 2.25, all pairs of $x, y \in$ Weyl$(|\psi\rangle)$ must have have $[x,y] = 0$, making the subspace isotropic. □

Additionally, for all $x \in$ Weyl$(|\psi\rangle)$ we find that $\mathsf{sgn}(\mathrm{Tr}[W_x |\psi\rangle\langle\psi|]) \cdot W_x$ is contained in Stab$(|\psi\rangle)$. As a result Weyl$(|\psi\rangle)$ is Lagrangian if and only if $|\psi\rangle$ is a stabilizer state. Furthermore, if $T \subset \mathbb{F}_2^{2n}$ is a Lagrangian subspace, then the set of states $\{|\varphi\rangle : \text{Weyl}(|\varphi\rangle) = T\}$ forms an orthonormal basis of the n-qubit Hilbert space. Moreover, since each basis state $|\varphi\rangle$ is stabilized by 2^n Weyl operators (up to phase), every basis state is a stabilizer state.

One final detail we need is that given $x, y \in$ Weyl$(|\phi\rangle)$ for stabilizer state $|\phi\rangle$, we need to be able to efficiently determine which of $\pm W_{x+y}$ is in Stab$(|\phi\rangle)$. Luckily, this was handled by [AG04]:

Fact 2.27 ([AG04]). *Given $(-1)^a W_x, (-1)^b W_y \in \mathit{Stab}(|\phi\rangle)$ for stabilizer state $|\phi\rangle$ and $a, b \in \mathbb{F}_2$, then $i^{[x',y']}(-1)^{a+b} W_{x+y} \in \mathit{Stab}(|\phi\rangle)$ where x' and y' are the embedding of x and y in $\mathbb{Z}$ (or $\mathbb{Z}_4$) respectively.*

Proof. Since $(-1)^a W_x, (-1)^b W_y \in \mathrm{Stab}(|\phi\rangle)$ then $x, y \in \mathrm{Weyl}(|\phi\rangle)$. By Corollary 2.26, W_x and W_y must commute.

The major question is, what is the phase of $W_x \cdot W_y$ with respect to W_{x+y}? It is well known that $XY = iZ$, $YZ = iX$, and $ZX = iY$ so we pick up an imaginary phase every time the Pauli matrix on a particular qubit doesn't commute between W_x and W_y. Luckily, since W_x and W_y commute, we know that the total phase will end up being real. The key idea is that, looking closely, what the symplectic product (without the modular arithmetic) actually computes is exactly how many times W_x and W_y don't commute over all n qubits. Due to the periodicity of i, we can also take the symplectic product modulo 4 without affecting the phase. □

We will implicitly apply this fact throughout the remainder of this work, since it allows us to focus on finding generators of $\mathrm{Stab}(|\phi\rangle)$ without worrying about finding phases of the items in the span.

2.6 Clifford Circuits

Informally, a Clifford circuit maps stabilizer states to other stabilizer states. Due to the close relationship between stabilizer states and Pauli operators, they must also map Pauli operators to Pauli operators as well.

Definition 2.28. *A* Clifford circuit *is a unitary C on n qubits such that $C\mathcal{P}_n C^\dagger = \mathcal{P}_n$, while ignoring global phase on the unitary. More formally, consider the normalizer $\mathcal{N}(\mathcal{P}_n) = \{C \in U(2^n) \mid C\mathcal{P}_n C^\dagger = \mathcal{P}_n\}$, and let $\mathcal{C}_n = \mathcal{N}(\mathcal{P}_n)/U(1)$ be the* Clifford group *on n qubits.*

Like stabilizer states, generators are an important part of how we deal with Clifford circuits. How a given Clifford circuit C acts on the generators of the Pauli matrices completely characterizes the unitary C [Low09]. To borrow the notation of Koenig [KS14], this relationship can be efficiently described via:

$$CX_jC^\dagger = (-1)^{p_j} W_{(\alpha_j,\beta_j)} \qquad CZ_jC^\dagger = (-1)^{q_j} \prod_{i=1}^{n} W_{(\gamma_j,\theta_j)} \tag{2.1}$$

where $j \in [n]$, $p_j, q_j \in \{0,1\}$, $\alpha_j, \beta_j, \gamma_j, \theta_j \in \{0,1\}^n$, and $W_{(v,w)}$ are Weyl operators (see Definition 2.14). It will sometimes be useful to view the sets $\{\alpha_j\}$, $\{\beta_j\}$, $\{\gamma_j\}$, and $\{\theta_j\}$ as matrices $A, B, \Gamma, \Theta \in \mathbb{F}_2^{n\times n}$ by stacking the bitstrings as columns then converting to $\mathbb{F}_2$. This gives us a simple upper-bound on the number of Clifford circuits.

Proposition 2.29. *There are at most $2^{\mathcal{O}(n^2)}$ Clifford circuits.*

Proof. The total number of bits we use to represent p, q, A, B, Γ, and Θ is $4n^2+2n = \mathcal{O}(n^2)$. There can then be at most $2^{\mathcal{O}(n^2)}$ Clifford circuits. □

However, because commutation relations are preserved, not all possible values of $\alpha, \beta, \gamma, \theta$ are allowed (the p and q values can be arbitrary). This leads us to the idea of symplectic matrices. We note that a Clifford circuit can be encoded as a $(2n+1) \times 2n$ boolean matrix S defined as

$$\begin{bmatrix} A & \Gamma \\ B & \Theta \\ p^T & q^T \end{bmatrix}.$$

We will call this the *full encoding* of the Clifford circuit.

Definition 2.30. *A symplectic matrix over $\mathbb{F}_2^{2n}$ is a $2n \times 2n$ matrix S with entries in $\mathbb{F}_2$ such that*

$$S^T \Lambda(n) S = \Lambda(n) = \begin{bmatrix} 0 & I^{\otimes n} \\ I^{\otimes n} & 0 \end{bmatrix} \tag{2.2}$$

These matrices form the symplectic group $Sp(2n, \mathbb{F}_2)$.

As a result, the symplectic matrices preserve the symplectic product $[x, y] = x^T\Lambda(n)y$ on $\mathbb{F}_2^{2n}$. It turns out that if we consider the submatrix

$$S := \begin{bmatrix} A & \Gamma \\ B & \Theta \end{bmatrix},$$

a necessary and sufficient condition to preserve the commutation relations of the generators is for S to be symplectic, as $\{X_i\} \cup \{Z_i\}$ forms a symplectic basis that generates $\mathbb{F}_2^{2n}$. Formally, $\mathcal{C}_n/\mathcal{P}_n \cong \mathrm{Sp}(2n, \mathbb{F}_2)$. In a slight abuse of notation, we will define the action of Clifford circuit C on $x \in \mathbb{F}_2^{2n}$, to be $C(x) := S \cdot x$. In this way, $CW_xC^\dagger = \pm W_{C(x)}$.

2.6.1 CNOT Circuits and $\oplus$L

It is a well known fact that every Clifford circuit can be generated using only H, P, and CNOT gates as defined below:

$$H = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \quad P = \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix} \quad \mathrm{CNOT} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

We note that $X = HP^2H$. If we restrict to the subset of circuits that are generated by only X and CNOT, we get what are known as CNOT circuits [AG04], which are a clear subset of Clifford circuits. As it turns out, they will contain much of the same features of Clifford circuits while greatly simplifying calculations.

Let us now consider the set of all Clifford circuit that map computational basis states to other computational basis states, thereby stabilizing the subgroup $\{\pm 1\} \times \{I, Z\}^{\otimes n}$. Very briefly, we will call these *classical Clifford circuits* as we will now prove that they are largely equivalent to CNOT circuits. The following lemmas will be useful.

Lemma 2.31. *Let Θ be the matrix form of the θ_j from Eq. (2.1). Any CNOT circuit C must have Θ be full rank.*

Proof. Let us first consider what happens to a computational basis state when acted upon by C. Referencing Eq. (2.1), Γ must be the zero matrix, since we always map computational basis states to computational basis states (i.e., never introduces X as a stabilizer). Since every member of $\mathrm{Sp}(2n, \mathbb{F}_2)$ is full rank, Θ must be full rank if Γ is rank zero. □

Lemma 2.32. *Let Θ be the matrix form of the θ_j from Eq. (2.1) for some Clifford circuit C. If Θ is full rank then there exists a CNOT circuit with the same Θ.*

Proof. One can verify that the Θ matrix of the circuit that does nothing, which is a valid CNOT circuit as well, is the identity matrix. We note that a CNOT from qubit i to qubit j performs the rowsum operation of adding row j to row i of Θ. Thus it is possible to efficiently construct a circuit with matching Θ using rowsum operations via CNOT gates. □

We can now prove our desired goal leveraging these two lemmas.

Proposition 2.33. *Let C be an arbitrary classical Clifford circuit. It can be efficiently generated using solely X, Z, and CNOT gates. Moreover, its effect on the computational basis states can be entirely simulated using only X and CNOT.*

Proof. Let us first consider what happens to a computational basis state when acted upon by C. Referencing Eq. (2.1), Γ must be 0 so that we map computational basis states to computational basis states. Let us focus on Θ and q. By Lemma 2.31, Θ must have full rank. By Lemma 2.32, there exists a CNOT matrix that achieves the same Θ as well. To get a matching q, one can simply apply an X gate at the beginning of each qubit that has $q_j = 1$, since $XZX = -Z$, and the following CNOT gates will not introduce any negative phases. From here, we have already proved the *moreover* statement.

To prove the full result, we return to A and B. We will show that there exists a single unique solution. Based on Eq. (2.2), to form a symplectic basis we find that $A^T\Theta = I$ and $A^TB = 0$, since $\Gamma = 0$. Clearly $A^T = \Theta^{-1}$, which is guaranteed to exist, and $B = 0$ since A

will also be full rank. To match the p_j values we simply place Z gates in front of the qubits where $p_j = 1$, similar to the X gates for q_j. □

This means we do not lose any kinds of interactions by only considering CNOT circuits, since the only differentiating factors (i.e., the Z gates) do not actually affect the outcome when fed with a computational basis state. As such, all given results will be given in terms of simply CNOT circuits.

2.7 Stabilizer Complexity

Throughout most of Part I, we will be concerned with states that are "close" to a stabilizer state, and showing that enough of the structure that we care about with stabilizer states remains to perform useful algorithms. But what exactly does "close" mean? We first define some complexity measures that characterize how far a general quantum state is from being stabilizer in terms of stabilizer state decompositions.

Definition 2.34 (stabilizer extent [BBC+19])**.** *Suppose $|\psi\rangle$ is a pure n-qubit state. The* STABILIZER EXTENT *of $|\psi\rangle$, denoted $\xi(|\psi\rangle)$, is the minimum of $\|c\|_1^2$ over all decompositions $|\psi\rangle = \sum_i c_i |\phi_i\rangle$, where $|\phi_i\rangle \in \mathcal{S}_n$ and c is some vector in $\mathbb{C}^{|\mathcal{S}_n|}$.*

Definition 2.35 (stabilizer fidelity [BBC+19])**.** *Suppose $|\psi\rangle$ is a pure n-qubit state. The* STABILIZER FIDELITY *of $|\psi\rangle$, denoted $F_{\mathcal{S}}$, is*

$$F_{\mathcal{S}}(|\psi\rangle) \coloneqq \max_{|\phi\rangle \in \mathcal{S}_n} F(|\psi\rangle\langle\psi|, |\phi\rangle\langle\phi|) = \max_{|\phi\rangle \in \mathcal{S}_n} |\langle\phi|\psi\rangle|^2$$

the maximum fidelity of $|\psi\rangle$ with any stabilizer state.

Below we give a useful relation between the complexity measures defined above.

Claim 2.36. *Let $|\psi\rangle$ be an n-qubit pure state. Then,*

$$\xi(|\psi\rangle) \geq \frac{1}{F_{\mathcal{S}}(|\psi\rangle)}.$$

Proof. Let $|\psi\rangle = \sum_{|\phi\rangle \in \mathcal{S}_n} c_\phi |\phi\rangle$ be such that $\left(\sum_\phi |c_\phi|\right)^2 = \xi(|\psi\rangle)$. Suppose towards a contradiction that $F_{\mathcal{S}}(|\psi\rangle) < \frac{1}{\xi(|\psi\rangle)}$ and therefore $|\langle\phi|\psi\rangle| < \frac{1}{\xi(|\psi\rangle)}$ for all $|\phi\rangle \in \mathcal{S}_n$. Then,

$$\begin{aligned} 1 = |\langle\psi|\psi\rangle| = \left| \sum_{|\phi\rangle \in \mathcal{S}_n} c_\phi^* \langle\phi|\psi\rangle \right| &\leq \sum_{|\phi\rangle \in \mathcal{S}_n} |c_\phi| \, |\langle\phi|\psi\rangle| \\ &\leq \max_i |\langle\phi_i|\psi\rangle| \sum_{|\phi\rangle \in \mathcal{S}_n} |c_\phi| \\ &\leq \sqrt{F_{\mathcal{S}}(|\psi\rangle)\xi(|\psi\rangle)} \\ &< 1, \end{aligned}$$

which is a clear contradiction. □

The claim above also follows as a special case of [BBC+19, Theorem 4], though its proof is more complicated.

We also define a new stabilizer complexity measure based on the unsigned stabilizer group. Rather than characterize states in terms of stabilizer decompositions, which is useful for classical simulation, it asks how much of its algebraic structure remains.

Definition 2.37 (Stabilizer dimension)**.** *Let $|\psi\rangle$ be an n-qubit pure state. The* STABILIZER DIMENSION OF $|\psi\rangle$ *is the dimension of* Weyl$(|\psi\rangle)$ *as a subspace of* $\mathbb{F}_2^{2n}$.

The stabilizer dimension of a stabilizer state is n, which is maximal, and, for most states, the stabilizer dimension is 0. It is closely related to the stabilizer nullity [BCHK20]. In fact, for n-qubit states, the stabilizer dimension is simply n minus the stabilizer nullity.

Part I

Bell Difference Sampling: Analysis and Algorithms

Chapter 3

Introduction to Bell Difference Sampling

This chapter introduces Bell difference sampling, a major algorithmic tool used throughout Part I. It borrows heavily from [GIKL23c], [GIKL23b] and [GIKL23a], which were all joint work with Sabee Grewal, Vishnu Iyer, and William Kretschmer.

The unifying tool in Part I is *Bell difference sampling*, a measurement primitive that has recently found applications in a variety of algorithms related to stabilizer states [Mon17, GNW21, GIKL23c, GIKL23b, GIKL23a]. We defer a full definition of Bell difference sampling to Section 3.2, but note some of its important properties here. Recall that Corollary 2.18 tells us that for pure states, the squared coefficients in this expansion sum to 1, and therefore form a distribution over $\mathbb{F}_2^{2n}$. We denote this distribution by $p_\psi(x) \coloneqq 2^{-n} \langle \psi | W_x | \psi \rangle^2$. Bell difference sampling involves measuring pairs of qubits of $|\psi\rangle^{\otimes 4}$ in the Bell basis and combining the measurements to interpret the result as corresponding to an n-qubit Weyl operator. Gross, Nezami, and Walter [GNW21] showed that Bell difference sampling a quantum pure state $|\psi\rangle$ is equivalent to sampling from the following distribution:

$$q_\psi(x) \coloneqq \sum_{a \in \mathbb{F}_2^{2n}} p_\psi(a) p_\psi(a + x),$$

i.e., the convolution of p_ψ with itself.

In this chapter, we will show why Bell difference sampling works. This will be our first glimpse into the power of symplectic Fourier analysis. We then show that when Bell difference sampling is performed on a stabilizer state, the result is the uniform distribution over the unsigned stabilizer group (see Definition 2.14). This will allow us to explain the learning algorithm from [Mon17] in Section 3.4. Along the way, we will also prove an algorithmic preliminary about finding important subspaces of $\mathbb{F}_2^{2n}$ using Bell difference sampling.

3.1 Symplectic Fourier Analysis

We now define and cover the basics of *symplectic* Fourier analysis. It is similar to Boolean Fourier analysis (see e.g., [O'D14]), except the Fourier characters are defined with respect to the symplectic product (Definition 2.8).

Definition 3.1 (Symplectic Fourier transform)**.** *Let $f : \mathbb{F}_2^{2n} \to \mathbb{R}$. We define the* SYMPLECTIC FOURIER TRANSFORM *of f, which is given by a function $\widehat{f} : \mathbb{F}_2^{2n} \to \mathbb{R}$, by*

$$\widehat{f}(a) = \frac{1}{4^n} \sum_{x \in \mathbb{F}_2^{2n}} (-1)^{[a,x]} f(x).$$

Hence, the SYMPLECTIC FOURIER EXPANSION *of f is*

$$f(x) = \sum_{a \in \mathbb{F}_2^{2n}} (-1)^{[a,x]} \widehat{f}(a).$$

We prove a fact that will be critical in our symplectic Fourier analysis.

Lemma 3.2. *For any subspace $T \subseteq \mathbb{F}_2^{2n}$ and a fixed $x \in \mathbb{F}_2^{2n}$,*

$$\sum_{a \in T} (-1)^{[a,x]} = |T| \cdot 1_{x \in T^\perp}.$$

Proof. If $x \in T^\perp$ then this is easy to see. Suppose $x \notin T^\perp$. Then we claim $[a,x] = 0$ for exactly half of the elements $a \in T$. To see this, we observe that there exists a $y \in T$ such that $[y,x] = 1$. Let T/y denote T modulo addition by y. Given a pair $\{a, a+y\} \in T/y$, observe that exactly one of $[a,x]$ and $[a+y,x]$ is 0 and the other is 1. As such we have that for half of all $a \in T$, $[a,x] = 0$ and for the other half, $[a,x] = 1$, giving us $\sum_{a \in T} (-1)^{[a,x]} = 0$. □

Lemma 3.2 immediately implies orthogonality of the new Fourier characters.

Corollary 3.3.

$$\sum_{x\in\mathbb{F}_2^{2n}}(-1)^{[a,x]}(-1)^{[b,x]} = 4^n 1_{a=b}$$

Proof. By linearity, $(-1)^{[a,x]}(-1)^{[b,x]} = (-1)^{[a+b,x]}$. Set $T = \mathbb{F}_2^{2n}$ (such that $T^\perp$ is the trivial subspace) and apply Lemma 3.2, then note that $a+b=0 \iff a=b$ over $\mathbb{F}_2$. □

As a basis change, we can then re-think inner products to be over the symplectic Fourier coefficients as well.

Fact 3.4 (Plancherel's theorem)**.**

$$\frac{1}{4^n}\sum_{x\in\mathbb{F}_2^{2n}}(-1)^{[a,x]}f(x)g(x) = \sum_{x\in\mathbb{F}_2^{2n}}\widehat{f}(x)\widehat{g}(x+a).$$

Proof.

$$\begin{aligned}\frac{1}{4^n}\sum_{x\in\mathbb{F}_2^{2n}}(-1)^{[a,x]}f(x)g(x) &= \frac{1}{4^n}\sum_{b,c\in\mathbb{F}_2^{2n}}\widehat{f}(b)\widehat{g}(c)\sum_{x\in\mathbb{F}_2^{2n}}(-1)^{[a+b+c,x]}\\ &= \frac{1}{4^n}\sum_{b,c\in\mathbb{F}_2^{2n}}\widehat{f}(b)\widehat{g}(c)(4^n 1_{a+b=c}) && \text{(Corollary 3.3)}\\ &= \sum_{x\in\mathbb{F}_2^{2n}}\widehat{f}(x)\widehat{g}(x+a)\end{aligned}$$

□

We will mostly call Fact 3.4 with $a=0$, recovering the usual Plancherel's theorem of $\frac{1}{4^n}\sum_{x\in\mathbb{F}_2^{2n}}f(x)g(x) = \sum_{x\in\mathbb{F}_2^{2n}}\widehat{f}(x)\widehat{g}(x)$. The one notable exception is the proof of Lemma 3.8.

Finally, the convolution will play a central role in our work as the Bell difference sampling distribution, q_ψ, will later be defined using it.

Definition 3.5 (Convolution)**.** *Let $f,g:\mathbb{F}_2^{2n}\to\mathbb{R}$. Their convolution is the function $f*g:\mathbb{F}_2^{2n}\to\mathbb{R}$ defined by*

$$(f*g)(x) = \frac{1}{4^n}\sum_{t\in\mathbb{F}_2^{2n}}f(t)g(t+x).$$

Convolution corresponds to the multiplication of Fourier coefficients, even under the symplectic Fourier transform.

Proposition 3.6. *Let $f, g : \mathbb{F}_2^{2n} \to \mathbb{R}$. Then for all $a \in \mathbb{F}_2^{2n}$,*

$$\widehat{f * g}(a) = \widehat{f}(x)\widehat{g}(a).$$

Proof. We again use the fact the symplectic product is bilinear, such that $[a, x] = [a, t] + [a, x + t]$. Using this, we can expand and simplify:

$$\begin{aligned}
\widehat{f * g}(a) &= \frac{1}{4^n} \sum_{x \in \mathbb{F}_2^{2n}} (-1)^{[a,x]} (f * g)(x) \\
&= \frac{1}{16^n} \sum_{x,t \in \mathbb{F}_2^{2n}} (-1)^{[a,x]} f(t) g(x + t) \\
&= \frac{1}{16^n} \sum_{t \in \mathbb{F}_2^{2n}} (-1)^{[a,t]} f(t) \sum_{x \in \mathbb{F}_2^{2n}} (-1)^{[a,x+t]} g(x + t) \\
&= \widehat{f}(a)\widehat{g}(a).
\end{aligned}$$

□

3.2 Bell Difference Sampling

Recall the Weyl expansion of a quantum state (see Definition 2.16). For compactness of notation, let $\psi = |\psi\rangle\langle\psi|$ be a pure quantum state. By Corollary 2.18, squaring the $c_\psi(x)$'s gives rise to a distribution over $\mathbb{F}_2^{2n}$ and therefore over the Weyl operators. We denote this distribution by $p_\psi(x) \coloneqq c_\psi(x)^2$ and refer to it as the CHARACTERISTIC DISTRIBUTION. Note that, for all x, $p_\psi(x) \in [0, 2^{-n}]$. While p_ψ is *not* the distribution that comes from Bell difference sampling, it will be very closely related.

We first show an extremely important fact about p_ψ in its invariance (up to scaling) under the symplectic Fourier transform.

Fact 3.7. *For any n-qubit pure state $|\psi\rangle$ and any $x \in \mathbb{F}_2^{2n}$, $p_\psi(x) = 2^n \widehat{p}_\psi(x)$.*

Proof.

$$\begin{aligned}
\widehat{p}_\psi(x) &= \frac{1}{4^n}\sum_{a\in\mathbb{F}_2^{2n}}(-1)^{[a,x]}p_\psi(a) \\
&= \frac{1}{4^n}\sum_{a\in\mathbb{F}_2^{2n}}(-1)^{[a,x]}c_\psi(a)c_\psi(a) \\
&= \frac{1}{4^n}\sum_{a\in\mathbb{F}_2^{2n}}(-1)^{[a,x]}c_\psi(a)\left(\frac{\langle\psi|W_a|\psi\rangle}{\sqrt{2^n}}\right) \\
&= \frac{1}{4^n}\sum_{a\in\mathbb{F}_2^{2n}}c_\psi(a)\left(\frac{\langle\psi|W_xW_aW_x|\psi\rangle}{\sqrt{2^n}}\right) && \text{(Fact 2.15)} \\
&= \frac{1}{4^n}\sum_{a\in\mathbb{F}_2^{2n}}c_\psi(a)c_{W_x\psi W_x}(a) \\
&= \frac{1}{4^n}\mathrm{Tr}[\psi(W_x\psi W_x)] && \text{(Lemma 2.17)} \\
&= \frac{1}{2^n}p_\psi(x)
\end{aligned}$$

□

Note our slight difference in normalization from [GNW21]. One can also refer to [GIKL23c, Proposition 3.3], where the normalization is consistent with this work, but uses the standard Boolean Fourier transform rather than the symplectic one. Despite this difference, the proof goes through in a similar way.

We now introduce the titular algorithmic primitive in Part I, *Bell difference sampling* [Mon17, GNW21]. Let $|\Phi^+\rangle := \frac{|00\rangle+|11\rangle}{\sqrt{2}}$ be the Bell state. Then, the set of quantum states

$$\{|W_x\rangle := (I\otimes W_x)\,|\Phi^+\rangle : x\in\mathbb{F}_2^2\} \tag{3.1}$$

forms an orthonormal basis of $\mathbb{C}^2\otimes\mathbb{C}^2$, which we call the *Bell basis.* Bell difference sampling an n-qubit state $|\psi\rangle$ just means the following. First, take two copies of a pure state $|\psi\rangle$. Take the first qubit in each copy and measure them in the Bell basis. Repeat this for each remaining pair of qubits. Let (a_i, b_i) denote the two-bit measurement outcome from measuring the ith pair of qubits. Then, we denote the measurement outcome on the two copies by $x = (a_1, \dots, a_n, b_1, \dots, b_n) \in \mathbb{F}_2^{2n}$. Repeat this once more with two fresh copies of $|\psi\rangle$ to obtain a string $y \in \mathbb{F}_2^{2n}$. Finally, output $x + y$.[1] Historically, Bell difference sampling has found use in algorithms for stabilizer states. However, Gross, Nezami, and Walter proved that Bell difference sampling is meaningful for all quantum states. This will be the first showcase of the power of symplectic Fourier analysis

Lemma 3.8 (Bell difference sampling, [GNW21, Theorem 3.2])**.** *Let $|\psi\rangle$ be an arbitrary n-qubit pure state. Bell difference sampling corresponds to drawing a sample from the following distribution:*

$$q_\psi(x) \coloneqq 4^n (p_\psi * p_\psi)(x) = \sum_{y \in \mathbb{F}_2^{2n}} p_\psi(y) p_\psi(x+y),$$

and uses four copies of $|\psi\rangle$. We refer to $q_\psi(x)$ as the Weyl distribution.

To prove Lemma 3.8, we will first need to prove an identity about the projectors onto each measurement outcome. In particular, we will show that they can be expressed as a nice summation over Weyl operators.

Lemma 3.9. *For $a \in \mathbb{F}_2^{2n}$, let $\Pi_a \coloneqq \sum_{x \in \mathbb{F}_2^{2n}} |W_x\rangle\langle W_x| \otimes |W_{x+a}\rangle\langle W_{x+a}|$ be the projector on $2n$ qubits such that $\mathrm{Tr}\left[\Pi_a \, |\psi\rangle\langle\psi|^{\otimes 4}\right] = q_\psi(a)$. Then Π_a can also be expressed as:*

$$\Pi_a = \frac{1}{4^n} \sum_{x \in \mathbb{F}_2^{2n}} (-1)^{[a,x]} W_x^{\otimes 4}.$$

Proof. We will start by getting an expression for $\Pi_0 = \sum_{x \in \mathbb{F}_2^{2n}} W_x^{\otimes 4}$. From there, we can use Fact 2.15 to achieve our desired identity.

It is a well-known fact that the Bell state's Weyl decomposition is

$$|\Phi^+\rangle\langle\Phi^+| = \frac{II + XX - YY + ZZ}{4} = \frac{W_{00}^{\otimes 2} + W_{10}^{\otimes 2} - W_{11}^{\otimes 2} + W_{01}^{\otimes 2}}{4} = \frac{1}{4} \sum_{x \in \mathbb{F}_2^2} W_x^{\otimes 2} (-1)^{b_x}$$

[1] Even when $|\psi\rangle$ is a stabilizer stabilizer state, measuring two copies of $|\psi\rangle$ in the Bell basis returns $x \in \mathbb{F}_2^{2n}$ with probability $p_\psi(x+a)$, where $a \in \mathbb{F}_2^{2n}$ is an unwanted shift. Bell difference sampling essentially cancels out this unwanted shift a. See [Mon17, GNW21] for more detail.

such that $b_x = 1_{x=11}$. We can use this to express $|W_x\rangle\langle W_x|$ as:

$$\begin{aligned}
|W_x\rangle\langle W_x| &= (I \otimes W_x)\,|\Phi^+\rangle\langle\Phi^+|\,(I \otimes W_x) && \text{(Eq. (3.1))}\\
&= (I \otimes W_x)\left(\frac{1}{4}\sum_{y\in\mathbb{F}_2^2} W_y^{\otimes 2}(-1)^{b_y}\right)(I \otimes W_x)\\
&= \frac{1}{4}\sum_{y\in\mathbb{F}_2^2} W_y^{\otimes 2}(-1)^{[x,y]+b_y} && \text{(Fact 2.15)}
\end{aligned}$$

and $|W_x\rangle\langle W_x|^{\otimes 2}$ as:

$$|W_x\rangle\langle W_x|^{\otimes 2} = \frac{1}{16}\sum_{y,z\in\mathbb{F}_2^2} W_y^{\otimes 2}W_z^{\otimes 2}(-1)^{[x,y+z]+b_y+b_z}.$$

Let us return to measurement probabilities. Notice that the probability of Bell difference sampling 0 on a single qubit is $\sum_{x\in\mathbb{F}_2^2}|W_x\rangle\langle W_x|^{\otimes 2}$, which has the following nice identity:

$$\begin{aligned}
\sum_{x\in\mathbb{F}_2^2}|W_x\rangle\langle W_x|^{\otimes 2} &= \sum_{x\in\mathbb{F}_2^2}\left(\frac{1}{16}\sum_{y,z\in\mathbb{F}_2^2} W_y^{\otimes 2}W_z^{\otimes 2}(-1)^{[x,y+z]+b_y+b_z}\right)\\
&= \frac{1}{16}\sum_{y,z\in\mathbb{F}_2^2} W_y^{\otimes 2}W_z^{\otimes 2}(-1)^{b_y+b_z}\left(\sum_{x\in\mathbb{F}_2^2}(-1)^{[x,y+z]}\right)\\
&= \frac{1}{16}\sum_{y,z\in\mathbb{F}_2^2} W_y^{\otimes 2}W_z^{\otimes 2}(-1)^{b_y+b_z}(4\cdot 1_{y=z}) && \text{(Lemma 3.2)}\\
&= \frac{1}{4}\sum_{x\in\mathbb{F}_2^2} W_x^{\otimes 4}.
\end{aligned}$$

Since Π_0 is simply the probability of sampling 0 on all n qubits, we can achieve our initial goal and show that:

$$\Pi_0 = \sum_{x\in\mathbb{F}_2^{2n}}|W_x\rangle\langle W_x|^{\otimes 2} = \bigotimes_{i=1}^{n}\sum_{x\in\mathbb{F}_2^2}|W_x\rangle\langle W_x|^{\otimes 2} = \frac{1}{4^n}\bigotimes_{i=1}^{n}\sum_{x\in\mathbb{F}_2^2}W_x^{\otimes 4} = \frac{1}{4^n}\sum_{x\in\mathbb{F}_2^{2n}}W_x^{\otimes 4}. \tag{3.2}$$

Finally, observe that:

$$\begin{aligned}
\Pi_a &:= \sum_{x \in \mathbb{F}_2^{2n}} |W_x\rangle\langle W_x| \otimes |W_{x+a}\rangle\langle W_{x+a}| \\
&= (I \otimes I \otimes I \otimes W_a)\Big(|W_x\rangle\langle W_x| \otimes |W_x\rangle\langle W_x| \Big)(I \otimes I \otimes I \otimes W_a) \\
&= (I \otimes I \otimes I \otimes W_a)\left(\frac{1}{4^n} \sum_{x \in \mathbb{F}_2^{2n}} W_x^{\otimes 4}\right)(I \otimes I \otimes I \otimes W_a) && \text{(Eq. (3.2))} \\
&= \frac{1}{4^n} \sum_{x \in \mathbb{F}_2^{2n}} (-1)^{[a,x]} W_x^{\otimes 4} && \text{(Fact 2.15)},
\end{aligned}$$

thus completing the proof.

□

Using the identity from Lemma 3.9, we can combine it with our knowledge about the symplectic Fourier transform to prove Lemma 3.8.

Proof of Lemma 3.8. By definition, for all $x \in \mathbb{F}_2^{2n}$, $\operatorname{Tr}\left[\Pi_x |\psi\rangle\langle\psi|^{\otimes 4}\right]$ is the probability of getting x from Bell difference sampling.

$$\begin{aligned}
q_\psi(x) &= \operatorname{Tr}\left[\Pi_x |\psi\rangle\langle\psi|^{\otimes 4}\right] \\
&= \frac{1}{4^n} \operatorname{Tr}\left[\left(\sum_{y \in \mathbb{F}_2^{2n}} (-1)^{[x,y]} W_y^{\otimes 4}\right) |\psi\rangle\langle\psi|^{\otimes 4}\right] && \text{(Lemma 3.9)} \\
&= \frac{1}{4^n} \sum_{y \in \mathbb{F}_2^{2n}} (-1)^{[x,y]} \langle\psi|W_y|\psi\rangle^2 \langle\psi|W_y|\psi\rangle^2 \\
&= \sum_{y \in \mathbb{F}_2^{2n}} (-1)^{[x,y]} p_\psi(y) p_\psi(y) \\
&= \frac{1}{4^n} \sum_{y \in \mathbb{F}_2^{2n}} \widehat{p_\psi}(y) \widehat{p_\psi}(x+y) && \text{(Fact 3.4)} \\
&= \sum_{y \in \mathbb{F}_2^{2n}} p_\psi(y) p_\psi(x+y) && \text{(Fact 3.7)}
\end{aligned}$$

□

3.3 Finding Generators Efficiently via Bell Difference Sampling

Throughout this work, we will want to find important subspaces of $\mathbb{F}_2^{2n}$ through Bell difference sampling. Let's call this important subspace $S \subseteq \mathbb{F}_2^{2n}$ for now. To learn what S is, we want to Bell difference sample independent generators that span S. However, if the probability mass of q_ψ is too heavily concentrated on a proper subspace of S, then we may never get enough independent generators. Assuming this condition is not true (i.e., the "hardest" step is not too hard) then we can upper bound the number of Bell difference samples to find generators of S.

Lemma 3.10. *Let S be some subspace of $\mathbb{F}_2^{2n}$. If $q_\psi(x)$ is such that for all proper subspaces $T \subset S$ with $\dim(T) = \dim(S) - 1$,*

$$\sum_{x \in S \setminus T} q_\psi(x) \geq c$$

then $\frac{2}{c}\big(\dim(S) + \log \frac{1}{\delta}\big)$ Bell difference samples are sufficient to sample generators of S with probability at least $1 - \delta$.

Proof. Let m be the number of samples and let $x_1, ..., x_m \in \mathbb{F}_2^{2n}$ be the results of the Bell difference sampling. Let T_i be the subspace of S spanned by all elements in $\{x_1, \ldots, x_i\} \cap S$. Define the indicator random variable X_i as

$$X_i = \begin{cases} 1 & x_i \in S \setminus T_i \text{ or } T_i = S \\ 0 & \text{otherwise.} \end{cases}$$

Informally, $X_i = 1$ indicates a step at which the algorithm has made progress towards sampling a complete set of generators for S. We need to show that with high probability, $\sum_{i=1}^m X_i \geq \dim(S)$, as this guarantees that $T_m = S$.

By assumption on q_ψ and S, we have that for any assignment of $X_1, \ldots, X_{i-1}$, $\mathbf{E}[X_i | X_1, \ldots, X_{i-1}] \geq c$, because x_i is sampled with probability $q_\psi(x_i)$. Let $\gamma = 1 - \frac{\dim(S)}{cm}$.

Then, by the multiplicative Chernoff bound (Fact 2.2), we have

$$\begin{aligned}
\mathbf{Pr}\left[\sum_{i=1}^{m} X_i < \dim(S)\right] &= \mathbf{Pr}\left[\sum_{i=1}^{m} X_i < (1-\gamma)cm\right] \\
&\leq \exp\left(-\gamma^2 \frac{cm}{2}\right) \\
&= \exp\left(-\left(1 - \frac{2\dim(S)}{cm} + \frac{\dim(S)^2}{c^2m^2}\right)\frac{cm}{2}\right) \\
&\leq \exp\left(-\left(1 - \frac{2\dim(S)}{cm}\right)\frac{cm}{2}\right) \\
&= \exp\left(\dim(S) - \frac{cm}{2}\right) \qquad (3.3)
\end{aligned}$$

Hence, choosing

$$m \geq \frac{2}{c}\left(\dim(S) + \log\frac{1}{\delta}\right)$$

suffices to guarantee that Eq. (3.3) is at most δ. □

3.4 Warm-up: Learning a Stabilizer State

To end this chapter on the basics of Bell Difference Sampling, we show how to use it to efficiently learn stabilizer states as in [Mon17]. This simple algorithm is the both the historical predecessor and the intuitive building block of much of the proceeding chapters, so care should be taken to understand this algorithm properly.

To start we note that for a stabilizer state $|\phi\rangle$, p_ϕ is the uniform distribution over its unsigned stabilizer group.

Lemma 3.11. *For a stabilizer state $|\phi\rangle$, p_ϕ is the uniform distribution over* $\mathrm{Weyl}(|\phi\rangle)$.

Proof. By definition of a stabilizer state, $|\phi\rangle\langle\phi| = \frac{1}{2^n}\sum_{\sigma\in G}\sigma$ for some stabilizer group G. For every x such that $\pm W_x \in G$, we note that $p_\phi(x) = \frac{1}{2^n}|\langle\phi|W_x\phi\rangle|^2 = \frac{1}{2^n}$. In turn, for every x such that $\pm W_x \notin G$, we get $p_\phi(x) = 0$. A fancy way of summarizing this is that $p_\phi(x) = \frac{1}{2^n}1_{x\in\mathrm{Weyl}(|\phi\rangle)}$. This means that p_ϕ is the uniform distribution over $\mathrm{Weyl}(|\phi\rangle)$. □

We now know that p_ϕ is nice and uniform, but what about q_ϕ? We next show that the convolution does nothing to p_ϕ.

Lemma 3.12. *For a stabilizer state $|\phi\rangle$, $q_\phi = p_\phi$.*

Proof.

$$\begin{aligned} q_\phi(x) &= \sum_{a \in \mathbb{F}_2^{2n}} p_\phi(a) p_\phi(a+x) \\ &= \frac{1}{2^n} \sum_{a \in \mathrm{Weyl}(|\phi\rangle)} p_\phi(a+x) && \text{(Lemma 3.11)} \\ &= \frac{1}{2^n} 1_{x \in \mathrm{Weyl}(|\phi\rangle)} \\ &= p_\phi(x) && \text{(Lemma 3.11)} \end{aligned}$$

□

We now have everything we need to learn $|\phi\rangle$. We present the algorithm for learning stabilizer states given in [Mon17] using notation and tools from this dissertation.

Algorithm 1: Learning a stabilizer state [Mon17]

```
Input: $4n + 4\log\frac{1}{\delta}$ copies of $|\phi\rangle$
Promise: $|\phi\rangle$ is a stabilizer state
Output: Generators of $\mathrm{Stab}(|\phi\rangle)$
Let $m = 4n + 4\log\frac{1}{\delta}$
Let $T = \{\}$
repeat $m$ times
    Perform Bell difference sampling to obtain $x \in \mathbb{F}_2^{2n}$
    Add $x$ to $T$
Find generators of $T$, $\{g_1, \cdots g_n\}$ using Gaussian elimination on $T$
Let $G = \{\}$
foreach $g_i$ do
    Add $\mathsf{sgn}(\langle\phi|W_{g_i}|\phi\rangle) \cdot W_{g_i}$ to $G$
return $G$ such that $\frac{1}{2^n}\sum_{\sigma \in \langle G \rangle} \sigma = |\phi\rangle\langle\phi|$
```

We now prove the correctness of Algorithm 1.

Theorem 3.13. *Algorithm 1 learns $|\phi\rangle$ with probability at least $1-\delta$ and uses $O(n+\log\frac{1}{\delta})$ samples and $O(n^3+n^2\log\frac{1}{\delta})$ time.*

Proof. A necessary condition to learning $\mathrm{Stab}(|\phi\rangle)$ is to find the generators of $\mathrm{Weyl}(|\phi\rangle)$. Let $S=\mathrm{Weyl}(|\phi\rangle)$. Because ϕ is a stabilizer state, $\dim(S)=n$. By Lemmas 3.11 and 3.12, our Bell difference samples come uniformly from S, such that for all proper subspaces $T\subset S$,

$$\sum_{x\in S\setminus T} q_\phi(x)\geq\frac{1}{2}.$$

By Lemma 3.10, we only need $4(n+\log\frac{1}{\delta})$ Bell difference samples[2] to get generators of S with probability at least $1-\delta$. By Gaussian elimination, we can find n independent generators $g_1,\cdots,g_n\in\mathbb{F}_2^{2n}$ such that their span is S.

Now all that's left is to find the appropriate phases for $\{W_x : x\in S\}$ to create $\mathrm{Stab}(|\phi\rangle)$. For each generator $g_i\in\mathbb{F}_2^{2n}$ we can simply measure $\langle\phi|W_{g_i}|\phi\rangle$ and take the appropriate sign correction to determine that $\mathsf{sgn}(\langle\phi|W_{g_i}|\phi\rangle)\cdot W_{g_i}\in\mathrm{Stab}(|\phi\rangle)$. We now have n independent generators of $\mathrm{Stab}(|\phi\rangle)$ and so have learned $|\phi\rangle$ with probability at least $1-\delta$. The total number of samples is $m=O(n+\log\frac{1}{\delta})$ and the time complexity is $O(mn^2)=O(n^3+n^2\log\frac{1}{\delta})$ since it is dominated by the cost of Gaussian elimination. □

We remark that since stabilizer states require $\Theta(n^2)$ bits to write down and each sample consists of $2n$ bits, this algorithm is asymptotically tight in sample complexity.

[2] To get rid of the factor of 4, see [Mer07].

Chapter 4

Symplectic Fourier Analysis and Bell Difference Sampling

This chapter explores the relationship between Bell difference sampling and symplectic Fourier analysis. It borrows heavily from [GIKL23c], [GIKL23b] and [GIKL23a], which were all joint work with Sabee Grewal, Vishnu Iyer, and William Kretschmer.

We now prove identities and inequalities related to the characteristic distribution p_ψ and Weyl distribution q_ψ that form the basic proof tools of the remainder of Part I. We emphasize that the results in this chapter hold for *all* pure quantum states.

4.1 Bell Difference Sampling and the Symplectic Complement

We show that the mass on a subspace $T \subseteq \mathbb{F}_2^{2n}$ under p_ψ is proportional to the mass on $T^\perp$ under p_ψ. It should be considered one of the most powerful results in the entirety of Part I. The idea is that Fact 3.7 means that p_ψ must exhibit a lot of structure that is formalized in the following:

Theorem 4.1. *Let $T \subseteq \mathbb{F}_2^{2n}$ be a subspace. Then*

$$\sum_{a \in T} p_\psi(a) = \frac{|T|}{2^n} \sum_{x \in T^\perp} p_\psi(x).$$

Proof.

$$\begin{aligned}
\sum_{a\in T} p_\psi(a) &= \sum_{a\in T}\sum_{x\in\mathbb{F}_2^{2n}} \widehat{p}_\psi(x)(-1)^{[a,x]} \\
&= \frac{1}{2^n}\sum_{a\in T}\sum_{x\in\mathbb{F}_2^{2n}} p_\psi(x)(-1)^{[a,x]} && \text{(Fact 3.7)} \\
&= \frac{|T|}{2^n}\sum_{x\in\mathbb{F}_2^{2n}} p_\psi(x)\cdot 1_{x\in T^\perp} && \text{(Lemma 3.2)} \\
&= \frac{|T|}{2^n}\sum_{x\in T^\perp} p_\psi(x). && \square
\end{aligned}$$

A similar result is true for q_ψ, due to Proposition 3.6. In words, we show that the average probability mass on a subspace T under q_ψ is equal to the squared-ℓ_2-norm of the probability mass on $T^\perp$ under p_ψ. It should be considered just as important as Theorem 4.1.

Theorem 4.2. *Let $T \subseteq \mathbb{F}_2^{2n}$ be a subspace. Then*

$$\frac{1}{|T|}\sum_{a\in T} q_\psi(a) = \sum_{x\in T^\perp} p_\psi(x)^2.$$

Proof.

$$\begin{aligned}
\sum_{a\in T} q_\psi(a) &= \sum_{a\in T}\sum_{x\in\mathbb{F}_2^{2n}} \widehat{q}_\psi(x)(-1)^{[a,x]} \\
&= 4^n\sum_{a\in T}\sum_{x\in\mathbb{F}_2^{2n}} \widehat{p}_\psi(x)^2(-1)^{[a,x]} && \text{(Lemma 3.8, Proposition 3.6.)} \\
&= \sum_{a\in T}\sum_{x\in\mathbb{F}_2^{2n}} p_\psi(x)^2(-1)^{[a,x]} && \text{(Fact 3.7)} \\
&= |T|\sum_{x\in T^\perp} p_\psi(x)^2. && \text{(Lemma 3.2)} \quad \square
\end{aligned}$$

An important consequence below is that q_ψ is “smoother” than p_ψ on subspaces, in that the probability mass on a subspace never increases upon convolution with itself.

Corollary 4.3. *Let $T \subseteq \mathbb{F}_2^{2n}$ be a subspace. Then*

$$\sum_{a\in T} q_\psi(a) \le \sum_{a\in T} p_\psi(a).$$

Proof.

$$\begin{aligned}
\sum_{a\in T} q_\psi(a) &= |T| \sum_{x\in T^\perp} p_\psi(x)^2 && \text{(Theorem 4.2)}\\
&\le \frac{|T|}{2^n} \sum_{x\in T^\perp} p_\psi(x) && (p_\psi(x) \le \frac{1}{2^n})\\
&= \sum_{a\in T} p_\psi(a). && \text{(Theorem 4.1)}
\end{aligned}$$

□

4.2 On the Support of p_ψ and q_ψ

An equally important consequence is that the support of both p_ψ and q_ψ (as well as any further self-convolutions of p_ψ) must lie entirely in the symplectic complement of $\mathrm{Weyl}(|\psi\rangle)$.

Lemma 4.4. *The support of $p_\psi(x)$ is contained in* $\mathrm{Weyl}(|\psi\rangle)^\perp$.

Proof. We show the mass of p_ψ on $\mathrm{Weyl}(|\psi\rangle)^\perp$ is 1.

$$\begin{aligned}
\sum_{x\in \mathrm{Weyl}(|\psi\rangle)^\perp} p_\psi(x) &= \frac{|\mathrm{Weyl}(|\psi\rangle)^\perp|}{2^n} \sum_{x\in \mathrm{Weyl}(|\psi\rangle)} p_\psi(x) && \text{(Theorem 4.1)}\\
&= \frac{|\mathrm{Weyl}(|\psi\rangle)^\perp|}{2^n}\frac{|\mathrm{Weyl}(|\psi\rangle)|}{2^n} && \text{(By definition of Weyl}(|\psi\rangle))\\
&= 1. && \text{(Fact 2.10)}
\end{aligned}$$

□

Corollary 4.5. *The support of $q_\psi(x)$ is contained in* $\mathrm{Weyl}(|\psi\rangle)^\perp$.

Proof. Suppose $x \notin \mathrm{Weyl}(|\psi\rangle)^\perp$. We want to show that $q_\psi(x) = 0$. By the definition of q_ψ and by Lemma 4.4,

$$q_\psi(x) = \sum_{a\in\mathbb{F}_2^{2n}} p_\psi(a)p_\psi(x+a) = \sum_{a\in \mathrm{Weyl}(|\psi\rangle)^\perp} p_\psi(a)p_\psi(x+a),$$

because $p_\psi(a) = 0$ for $a \notin \mathrm{Weyl}(|\psi\rangle)^\perp$. In the right-most sum, since $a \in \mathrm{Weyl}(|\psi\rangle)^\perp$, $x+a \notin \mathrm{Weyl}(|\psi\rangle)^\perp$ if and only if $x \notin \mathrm{Weyl}(|\psi\rangle)^\perp$. So, applying Lemma 4.4 again, $p_\psi(x+a) = 0$ for each term in the sum, implying that the total sum is 0. □

An implication of Corollary 4.5 and Fact 2.10, is that if there exists a subspace $T \subseteq \mathbb{F}_2^{2n}$ such that $\sum_{x\in T} q_\psi(x) = 1$ (this also implies $\sum_{x\in T} p_\psi(x) = 1$ by Corollary 4.3) then $T^\perp \subseteq \mathrm{Weyl}(|\psi\rangle)$.

Corollary 4.6. *Let $T \subseteq \mathbb{F}_2^{2n}$ be a subspace such that $\sum_{x\in T} q_\psi(x) = 1$. Then $T^\perp \subseteq \mathrm{Weyl}(|\psi\rangle)$.*

Proof. By Corollary 4.5, $T \supseteq \mathrm{Weyl}(|\psi\rangle)^\perp$. By Fact 2.10, $T^\perp \subseteq \mathrm{Weyl}(|\psi\rangle)$. □

4.3 Concentration of p_ψ and q_ψ Implies Commutativity

We show that if a subspace of $\mathbb{F}_2^{2n}$ has exceptionally large p_ψ-mass (or q_ψ-mass), then it must be coisotropic. This is important since unsigned stabilizer groups must be isotropic. The idea will be that an important isotropic subgroup of $\mathbb{F}_2^{2n}$ can be associated with a coisotropic subspace (via the symplectic complement) that has large probability mass. Thus, we can learn these important isotropic subspaces by Bell sampling until we are confident we have spanned its symplectic complement.

The first result is that a subspace whose p_ψ-mass is strictly greater than $\frac{3}{4}$ is coisotropic.

Lemma 4.7. *Let H be a subspace of $\mathbb{F}_2^{2n}$ such that*

$$\sum_{x\in H^\perp} p_\psi(x) > \frac{3}{4}.$$

Then H is isotropic.

Proof. Let $M := \{x \in H : 2^n p_\psi(x) > 1/2\}$. By Fact 2.25, every pair of elements in M commutes. Furthermore, every pair of elements in $\langle M\rangle$ (i.e., the span of M) also commute by linearity of the symplectic product. Thus, if we can show that $\langle M\rangle = H$ then H is isotropic.

Suppose for a contradiction that $|M| \leq \frac{|H|}{2}$. Then:

$$\begin{aligned}
\sum_{x \in H^\perp} p_\psi(x) &= \frac{|H^\perp|}{2^n} \sum_{x \in H} p_\psi(x) && \text{(Theorem 4.1)} \\
&= \frac{|H^\perp|}{2^n} \left(\sum_{x \in M} p_\psi(x) + \sum_{x \in H \setminus M} p_\psi(x) \right) \\
&\leq \frac{|H^\perp|}{2^n} \left(\frac{|H|}{2} \cdot \frac{1}{2^n} + \frac{|H|}{2} \cdot \frac{1}{2 \cdot 2^n} \right) \\
&= \frac{|H| \cdot |H^\perp|}{4^n} \cdot \left(\frac{1}{2} + \frac{1}{1} \right) \\
&= \frac{3}{4},
\end{aligned}$$

which contradicts the assumption of the lemma. So, $|M| > \frac{|H|}{2}$. Since a proper subspace of H can have at most $\frac{|H|}{2}$ elements, it follows that $\langle M \rangle = H$ and H is isotropic. □

As one might expect, one can instead use Theorem 4.2 to show that any subspace whose q_ψ-mass is strictly greater than $\frac{5}{8}$ is also coisotropic. The proof follows nearly identically to Lemma 4.7 and is given for completeness.

Lemma 4.8. *Let H be a subspace of $\mathbb{F}_2^{2n}$ such that*

$$\sum_{x \in H^\perp} q_\psi(x) > \frac{5}{8}.$$

Then H is isotropic.

Proof. Let $M := \{x \in H : 2^n p_\psi(x) > 1/2\}$. By Fact 2.25, every pair of elements in M commutes. Furthermore, every pair of elements in $\langle M \rangle$ (i.e., the span of M) also commute by linearity of the symplectic product. Thus, if we can show that $\langle M \rangle = H$ then H is isotropic.

Suppose for a contradiction that $|M| \leq \frac{|H|}{2}$. Then:

$$\begin{aligned}
\sum_{x \in H^\perp} q_\psi(x) &= |H^\perp| \sum_{x \in H} p_\psi(x)^2 && \text{(Theorem 4.2)} \\
&= |H^\perp| \left(\sum_{x \in M} p_\psi(x)^2 + \sum_{x \in H \setminus M} p_\psi(x)^2 \right) \\
&\leq |H^\perp| \left(\frac{|H|}{2} \cdot \frac{1}{4^n} + \frac{|H|}{2} \cdot \frac{1}{4 \cdot 4^n} \right) \\
&= \frac{|H| \cdot |H^\perp|}{4^n} \cdot \left(\frac{1}{2} + \frac{1}{8} \right) \\
&= \frac{5}{8},
\end{aligned}$$

which contradicts the assumption of the lemma. So, $|M| > \frac{|H|}{2}$. Since a proper subspace of H can have at most $\frac{|H|}{2}$ elements, it follows that $\langle M \rangle = H$ and H is isotropic. □

4.4 On the Relationship between Bell Difference Sampling and Stabilizer Fidelity

A crucial part of the analysis of algorithms in Part I will be to relate Bell difference sampling to stabilizer fidelity. Even when stabilizer fidelity is not the important quantity in a particular context, it can still be used as a distance measure to the properties that we do explicitly care about (this will be the case in Chapter 7). In this section, we detail many of the shared proofs/techniques, all of which (explicitly or implicitly) rely on Theorem 4.1 and Theorem 4.2. A simplified summary of the main results are that for a state $|\psi\rangle$ whose stabilizer fidelity is realized by $|\phi\rangle$ then

$$F_{\mathcal{S}}(|\psi\rangle)^2 \leq \sum_{x \in S^*} p_\psi(x) \leq F_{\mathcal{S}}(|\psi\rangle)$$

and

$$F_{\mathcal{S}}(|\psi\rangle)^4 \leq \sum_{x \in S^*} q_\psi(x) \leq F_{\mathcal{S}}(|\psi\rangle)$$

where $S^* := \mathrm{Weyl}(|\phi\rangle)$.

4.4.1 Action of Clifford Circuits on Bell Difference Sampling

It will oftentimes be easier to imagine that we are working with the computational basis. Due to the symmetry of the stabilizer states, one should be able to do this without loss of generality. We provide two basic lemmas that formalize this idea, allowing us to suppose without loss of generality that $|0^n\rangle$ maximizes stabilizer fidelity. See Section 2.6 for preliminary details on Clifford circuits.

The first lemma concerns our ability to map our state (as well as proper subspaces) to $\{I, Z\}^{\otimes n}$.

Lemma 4.9. *Given an n-qubit stabilizer state $|\phi\rangle$, let $S = \mathrm{Weyl}(|\phi\rangle)$ be its unsigned stabilizer group, and let $T \subseteq S$ be a subspace of dimension $n - t$. Then there exists a Clifford circuit C such that $C|\phi\rangle = |0^n\rangle$, $C(S) = 0^n \times \mathbb{F}_2^n$, and $C(T) = 0^{n+t} \times \mathbb{F}_2^{n-t}$.*

Proof. Because the Clifford group acts transitively on stabilizer states, there exists a Clifford circuit C such that $C|\phi\rangle = |0^n\rangle$. Because $S = \{x \in \mathbb{F}_2^{2n} : \langle\phi| W_x |\phi\rangle = \pm 1\}$, this C necessarily maps S to $C(S) = \{x \in \mathbb{F}_2^{2n} : \langle\phi| C^\dagger W_x C |\phi\rangle = \pm 1\} = 0^n \times \mathbb{F}_2^n$. So, it only remains to show that C can be chosen so as to map T to $0^{n+t} \times \mathbb{F}_2^{n-t}$ while preserving these properties. This holds because a CNOT gate between qubits i and j in its action on $\mathbb{F}_2^{2n}$ maps $(0^n, x) \in \mathbb{F}_2^{2n}$ to $(0^n, Mx)$ where $M \in \mathrm{GL}_n(\mathbb{F}_2)$ is an elementary matrix (in particular, a matrix equal to the identity except with the (i, j) entry equal to 1). Hence, CNOT gates between arbitrary qubits generate all of $\mathrm{GL}_n(\mathbb{F}_2)$. So, we can choose CNOT gates so as to map T to an arbitrary subspace of $0^n \times \mathbb{F}_2^n$ of the same dimension, while preserving $|0^n\rangle$ and $0^n \times \mathbb{F}_2^n$. □

The second lemma details how p_ψ changes after a clifford circuit is applied to $|\psi\rangle$.

Lemma 4.10. *Let $|\psi\rangle$ be an n-qubit quantum state, let C be a Clifford circuit, and define $|\psi'\rangle \coloneqq C|\psi\rangle$. Then*

$$p_{\psi'}(x) = p_\psi(C^\dagger(x))$$

for all $x \in \mathbb{F}_2^{2n}$.

Proof.

$$2^n p_{\psi'}(x) = \langle\psi|C^\dagger W_x C|\psi\rangle^2 = 2^n p_\psi(C^\dagger(x)). \qquad \square$$

4.4.2 Identities for Weyl Decompositions of Computational Basis States

Now that Section 4.4.1 allows us to work with computational basis states, we give some useful identities for these states. The techniques will be very similar to doing Boolean Fourier analysis (using the normal inner product over $\mathbb{F}_2$ rather than the symplectic product).

Fact 4.11.

$$|x\rangle\langle x| = \frac{1}{2^k}\sum_{y\in\mathbb{F}_2^k} Z^y(-1)^{x\cdot y}$$

where $x \in \mathbb{F}_2^k$.

Proof. Since $|x\rangle$ is a stabilizer state and there are 2^k Weyl operators in this decomposition, we just need to show that $(-1)^{x\cdot y}Z^y$ stabilizes $|x\rangle$ for all $y \in \mathbb{F}_2^k$.

$$(-1)^{x\cdot y}Z^y|x\rangle = (-1)^{x\cdot y}(-1)^{x\cdot y}|x\rangle = |x\rangle \qquad \square$$

Another way of seeing this is that since $|x\rangle\langle x|$ is a $2^k \times 2^k$ diagonal matrix with a single element and Z^y form the parity functions over the diagonal, this is simply the Boolean Fourier decomposition of the indicator function

We now prove a relation between sums over products of basis state projections and sums products of Pauli-Z strings.

Fact 4.12.

$$\sum_{x\in\mathbb{F}_2^k} |x\rangle\langle x| \otimes |x\rangle\langle x| = \frac{1}{2^k}\sum_{x\in\mathbb{F}_2^k} Z^x \otimes Z^x.$$

Proof.

$$\begin{aligned}\sum_{x\in\mathbb{F}_2^k} |x\rangle\langle x| \otimes |x\rangle\langle x| &= \sum_{x\in\mathbb{F}_2^k} \frac{1}{4^k}\left(\sum_{a\in\mathbb{F}_2^k} Z^a(-1)^{a\cdot x}\right)\left(\sum_{b\in\mathbb{F}_2^k} Z^b(-1)^{b\cdot x}\right) && \text{(Fact 4.11)}\\ &= \frac{1}{4^k}\sum_{a,b\in\mathbb{F}_2^k} Z^a\otimes Z^b \sum_{x\in\mathbb{F}_2^k}(-1)^{(a+b)\cdot x}\\ &= \frac{1}{2^k}\sum_{a\in\mathbb{F}_2^k} Z^a\otimes Z^a. && \square\end{aligned}$$

4.4.3 Stabilizer Fidelity Implies Concentration of p_ψ

Let $|\psi\rangle$ be an arbitrary quantum state and let $|\phi\rangle$ be a stabilizer state that maximizes stabilizer fidelity with $|\psi\rangle$. It is easy to see that if $S^* := \mathrm{Weyl}(|\phi\rangle)$ and ψ is itself a stabilizer state (such that $|\psi\rangle = |\phi\rangle$), then the sum of p_ψ over S^* will be 1. Informally, it should be the case that if the stabilizer fidelity of $|\psi\rangle$ is still large, then $\sum_{x\in S^*} p_\psi(x)$ should also remain close to 1. Now, we show that the p_ψ-mass on S^* is bounded below by the squared stabilizer fidelity of $|\psi\rangle$.

Lemma 4.13. *Given an n-qubit state $|\psi\rangle$, let $|\phi\rangle$ be a stabilizer state that maximizes the stabilizer fidelity with $|\psi\rangle$, and let $S^* = \mathrm{Weyl}(|\phi\rangle)$. Then*

$$\sum_{x\in S^*} p_\psi(x) \geq F_{\mathcal{S}}(|\psi\rangle)^2.$$

Proof. Since $|\phi\rangle$ maximizes the stabilizer fidelity with $|\psi\rangle$, we can write $F_{\mathcal{S}}(|\psi\rangle) = |\langle\phi|\psi\rangle|^2$. Let C be a Clifford circuit from Lemma 4.9 such that $C\,|\phi\rangle = |0^n\rangle$ and $C(S^*) = 0^n \times \mathbb{F}_2^n$ (the choice of T is irrelevant). Now let $|\psi'\rangle = C\,|\psi\rangle$. Based on Lemma 4.10,

$$\sum_{x\in S^*} p_\psi(x) = \sum_{x\in C(S^*)} p_{\psi'}(x) = \sum_{x\in\mathbb{F}_2^n} p_{\psi'}(0^n, x).$$

It remains to lower bound $\sum_{x \in \mathbb{F}_2^n} p_{\psi'}(0^n, x)$.

$$\begin{aligned}
\sum_{x \in \mathbb{F}_2^n} p_{\psi'}(0^n, x) &= \frac{1}{2^n} \sum_{x \in \mathbb{F}_2^n} \langle \psi' | Z^x | \psi' \rangle^2 \\
&\geq \frac{1}{4^n} \left(\sum_{x \in \mathbb{F}_2^n} \langle \psi' | Z^x | \psi' \rangle \right)^2 \\
&= \frac{1}{4^n} \left(2^n \langle \psi' | 0^n \rangle \langle 0^n | \psi' \rangle \right)^2 && \text{(Fact 4.11)} \\
&= |\langle \psi | C^\dagger | 0^n \rangle|^4 \\
&= |\langle \psi | \phi \rangle|^4 = F_{\mathcal{S}}(|\psi\rangle)^2.
\end{aligned}$$

The second line follows from Cauchy-Schwarz.

Since we know that $\sum_{x \in \mathbb{F}_2^n} p_{\psi'}(0^n, x) \geq F_{\mathcal{S}}(|\psi\rangle)^2$, this tells us that $\sum_{x \in S^*} p_\psi(x) \geq F_{\mathcal{S}}(|\psi\rangle)^2$ as well. □

We can also generalize this result to arbitrary subspaces of S^*. This will become relevant later in Chapter 6.

Corollary 4.14. *Given an n-qubit state $|\psi\rangle$, let $|\phi\rangle$ be a stabilizer state that maximizes the stabilizer fidelity, and let $S^* = \mathrm{Weyl}(|\phi\rangle)$. Let $T \subseteq S^*$ be a subspace of S^*. Then*

$$\sum_{x \in T} p_\psi(x) \geq \frac{|T|}{2^n} F_{\mathcal{S}}(|\psi\rangle)^2.$$

Proof.

$$\begin{aligned}
\sum_{x \in T} p_\psi(x) &= \frac{|T|}{2^n} \sum_{x \in T^\perp} p_\psi(x) && \text{(Theorem 4.1)} \\
&\geq \frac{|T|}{2^n} \sum_{x \in S^*} p_\psi(x) && (\forall x, p_\psi(x) \geq 0) \\
&= \frac{|T|}{2^n} F_{\mathcal{S}}(|\psi\rangle)^2, && \text{(Lemma 4.13)}
\end{aligned}$$

where we have used the fact that $S^* \subseteq T^\perp$. □

Finally, for completeness we state the following related result for q_ψ by using Theorem 4.2.

Corollary 4.15. *Given an n-qubit state $|\psi\rangle$, let $|\phi\rangle$ be a stabilizer state that maximizes the stabilizer fidelity, and let $S^* = \mathrm{Weyl}(|\phi\rangle)$. Let $T \subseteq S^*$ be a subspace of S^*. Then*

$$\sum_{x \in T} q_\psi(x) \geq \frac{|T|}{2^n} F_{\mathcal{S}}(|\psi\rangle)^4.$$

Proof.

$$\begin{aligned}
\sum_{x\in T} q_\psi(x) &= |T| \sum_{x \in T^\perp} p_\psi(x)^2 && \text{(Theorem 4.2)} \\
&\geq |T| \sum_{x \in S^*} p_\psi(x)^2 && (\forall x, p_\psi(x) \geq 0) \\
&\geq \frac{|T|}{2^n} \left(\sum_{x \in S^*} p_\psi(x) \right)^2 && \text{(Cauchy-Schwarz)} \\
&= \frac{|T|}{2^n} F_{\mathcal{S}}(|\psi\rangle)^4, && \text{(Lemma 4.13)}
\end{aligned}$$

where we again use the fact that $S^* \subseteq T^\perp$ as in the proof of Lemma 4.13. □

4.4.4 Concentration of p_ψ Implies Stabilizer Fidelity

We will now show that the converse relations are also true, in that if the p_ψ mass is large over an isotropic subspace, than this also implies something about stabilizer fidelity. The first instance of such a result comes from [GNW21], where the isotropic subspace needs to also be Lagrangian.

Proposition 4.16 ([GNW21] Theorem 3.3)**.** *Let $S \subset \mathbb{F}_2^{2n}$ be some Lagrangian subspace.*

$$F_{\mathcal{S}}(|\psi\rangle) \geq \sum_{x \in S} p_\psi(x).$$

We will give a slightly different proof from [GNW21] that will allow us to generalize more easily to results for isotropic subspaces that are not Lagrangian later in Chapter 7.

First we prove that the p_ψ-mass on an isotropic subspace has a nice operational interpretation.

Lemma 4.17. *Let $|\psi\rangle$ be an n-qubit state, and let $T = 0^{n+t} \times \mathbb{F}_2^{n-t}$. Upon measuring the last $n-t$ qubits in the computational basis on 2 copies of $|\psi\rangle$, the probability of observing the same string $x \in \mathbb{F}_2^{n-t}$ twice (i.e., the collision probability) is*

$$2^t \sum_{x \in T} p_\psi(x).$$

Proof. The probability of observing some $x \in \mathbb{F}_2^{n-t}$ twice is

$$\begin{aligned}
\sum_{x \in \mathbb{F}_2^{n-t}} \langle \psi | \left(I^{\otimes t} \otimes |x\rangle\langle x| \right) |\psi\rangle^2 &= \langle \psi |^{\otimes 2} \left(\sum_{x \in \mathbb{F}_2^{n-t}} I^{\otimes t} \otimes |x\rangle\langle x| \otimes I^{\otimes t} \otimes |x\rangle\langle x| \right) |\psi\rangle^{\otimes 2} \\
&= \langle \psi |^{\otimes 2} \left(\frac{1}{2^{n-t}} \sum_{x \in \mathbb{F}_2^{n-t}} I^{\otimes t} \otimes Z^x \otimes I^{\otimes t} \otimes Z^x \right) |\psi\rangle^{\otimes 2} \\
&= \frac{1}{2^{n-t}} \sum_{x \in \mathbb{F}_2^{n-t}} \langle \psi | I^{\otimes t} \otimes Z^x | \psi \rangle^2 \\
&= 2^t \sum_{x \in T} p_\psi(x).
\end{aligned}$$

The third step follows from Fact 4.12 by treating the $I^{\otimes t}$ as constants. □

It will turn out that the collision probability lower-bounds the stabilizer fidelity. Intuitively, if the collision probability is very high, then $|\psi\rangle$ must be close to a computational basis state, which is a special kind of stabilizer state. Since the collision probability is related to p_ψ via Lemma 4.17, we should then be able to lower-bound the stabilizer fidelity by p_ψ. We now formalize this intuition in our proof of Proposition 4.16.

Proof of Proposition 4.16. Let C be a Clifford circuit from Lemma 4.9 such that $C|\phi\rangle = |0^n\rangle$ and $C(S) = 0^n \times \mathbb{F}_2^n$ (the choice of T is irrelevant). Now let $|\psi'\rangle = C|\psi\rangle$. From Lemma 4.10,

$$\sum_{x \in S} p_\psi(x) = \sum_{x \in C(S^*)} p_{\psi'}(x) = \sum_{x \in \mathbb{F}_2^n} p_{\psi'}(0^n, x).$$

If we can show that

$$|\langle\psi|\phi\rangle|^2 = |\langle\psi|C|0^n\rangle|^2 = \max_{x\in\mathbb{F}_2^n}|\langle\psi'|x\rangle|^2 \geq \sum_{x\in\mathbb{F}_2^n} p_{\psi'}(0,x),$$

then we are done. First, we can always perform the following decomposition $|\psi'\rangle = \sum_{x\in\mathbb{F}_2^n}\alpha_x\,|x\rangle$ such that the sum of the $|\alpha_x|^2$ is 1. Using this decomposition, we find:

$$\begin{aligned} &= \max_{x\in\mathbb{F}_2^n}|\langle\psi'|x\rangle|^2 && = \max_{x\in\mathbb{F}_2^n}|\alpha_x|^2 \\ &= \max_{x\in\mathbb{F}_2^n}|\alpha_x|^2 \cdot \sum_{x\in\mathbb{F}_2^{n-t}}|\alpha_x|^2 && \\ &\geq \sum_{x\in\mathbb{F}_2^n}|\alpha_x|^4. && \end{aligned}$$

Observe that $\sum_{x\in\mathbb{F}_2^n}|\alpha_x|^4$ is precisely the collision probability when measuring all n qubits of $|\psi\rangle$ in the computational basis. Hence, by Lemma 4.17 with $t = 0$,

$$\sum_{x\in\mathbb{F}_2^n}|\alpha_x|^4 = \sum_{x\in S} p_\psi(x). \qquad \square$$

We will now informally state a generalization of Proposition 4.16 to all isotropic subspaces. While the statement is a bit convoluted, the idea is that if an isotropic subspace T has near maximal p_ψ-mass then it is close to a state $|\widehat{\psi}\rangle$ whose unsigned stabilizer group is T (i.e., acts like a stabilizer state relative to T). As such, if T is in fact a Lagrangian subspace, the only such $|\widehat{\psi}\rangle$ are stabilizer states and we recover Proposition 4.16. We defer the formal statement and proof to Corollary 7.9, where we will also need to include computational requirements.

Corollary 4.18 (Informal version of Corollary 7.9)**.** *Let T be an isotropic subspace of dimension $n - t$. Then there exists a state $|\widehat{\psi}\rangle$ with $T \subseteq \mathrm{Weyl}(|\widehat{\psi}\rangle)$ such that*

$$|\langle\psi|\psi'\rangle|^2 \geq \sum_{x\in T} p_\psi(x).$$

Finally, we note that, due to Corollary 4.3, all of these lower bounds also hold for q_ψ.

Chapter 5

Tolerant Property Testing of Stabilizer States

This chapter is based on both [GIKL23c] and Section 6 of [GIKL23b], which were joint work with Sabee Grewal, Vishnu Iyer, and William Kretschmer. The work is presented non-chronologically with each section being from one or the other. Some of the preliminary results were moved to Chapters 2 to 4.

In this chapter, we give a tolerant property testing algorithm for stabilizer states. In the tolerant property testing model [PRR06], which generalizes ordinary property testing [RS96, GGR98], a tester must accept objects that are at most ε_1-close to having some property and reject objects that are at least ε_2-far from having that same property for $0 \leq \varepsilon_1 < \varepsilon_2 \leq 1$. The standard property testing model is recovered when $\varepsilon_1 = 0$, and the relaxed completeness condition generally makes tolerant testing a much harder problem. Nonetheless, the tolerant testing model is natural to consider in certain error models, such as in the presence of imprecise quantum gates.

Our algorithm takes copies of an n-qubit quantum state $|\psi\rangle$ and decides whether $|\psi\rangle$ has stabilizer fidelity at least α_1 or at most α_2, promised that one of these is the case. Note that we have taken $\alpha_1 \coloneqq 1 - \varepsilon_1$ and $\alpha_2 \coloneqq 1 - \varepsilon_2$ for notational simplicity.

Theorem 5.1 (Informal version of Theorem 5.6)**.** *Let $|\psi\rangle$ be an n-qubit pure state, and Let $\alpha_1, \alpha_2 \in [0,1]$ such that $\alpha_2 < \frac{4\alpha_1^6 - 1}{3}$, and define $\gamma \coloneqq \alpha_1^6 - \frac{3\alpha_2+1}{4}$. There is an algorithm that uses $O(1/\gamma^2)$ copies of $|\psi\rangle$, $O(n/\gamma^2)$ time, and decides whether $|\psi\rangle$ has stabilizer fidelity at least α_1 or at most α_2, promised that one of these is the case.*

While our algorithm does not work for all settings of ε_1 and ε_2—giving such an algorithm is an open problem—our algorithm does significantly improve over prior work. In Section 5.3, we compare the parameter regimes in which our algorithm works to the existing literature and show those regimes visually in Fig. 5.1.

We note that this tolerant testing algorithm results solely from an improvement of the analysis of the stabilizer state property testing algorithm due to Gross, Nezami, and Walter [GNW21] (hereafter, the "GNW algorithm"). It is also the basis for the distinguishing algorithm in Section 8.2.

5.1 The GNW Algorithm

We briefly describe the GNW algorithm, which works as follows. Perform Bell difference sampling on the input state to get a string $x \in \mathbb{F}_2^{2n}$. Then measure the corresponding Weyl operator W_x twice and accept if the result is the same across both measurements. The algorithm uses six copies of the input state.

[GIKL23c] introduced the following statistic:

$$\eta \coloneqq \mathop{\mathbf{E}}_{x \sim q_\psi(x)} [2^n p_\psi(x)] = 4^n \sum_{x \sim \mathbb{F}_2^{2n}} p_\psi(x)^3,$$

and showed that it can be efficiently estimated. Let p_{accept} denote the acceptance probability of the GNW algorithm. It is easy to show that $\eta = 2p_{\mathsf{accept}} - 1$ (see [GNW21, Page 19]). [GNW21] showed that for any pure quantum state $|\psi\rangle$,

$$2\eta - 1 \leq F_{\mathcal{S}}(|\psi\rangle).$$

Note that for a stabilizer state $\eta = 1$, since q_ψ will only be supported on $\mathrm{Weyl}(|\psi\rangle)$ (see Lemma 3.11). This gave an efficient testing algorithm that could distinguish between stabilizer states and states with stabilizer fidelity bounded away from 1.

We greatly improve on the analysis of this algorithm, improving both the completeness and soundness. As a first step, we relate η to the Fourier coefficients of p_ψ. Note that this analysis closely resembles the BLR linearity test [BLR93] (see also [O'D14, Section 1.6]).

Fact 5.2. *Let $|\psi\rangle$ be an n-qubit pure state. Then,*

$$\eta = 4^n \sum_{x\in\mathbb{F}_2^{2n}} p(x)^3.$$

Proof.

$$\begin{aligned}
\eta &= \mathop{\mathbf{E}}_{x\sim q_\psi}\left[|\langle\psi|W_x|\psi\rangle|^2\right] \\
&= 2^n \mathop{\mathbf{E}}_{x\sim q_\psi}\left[p_\psi(x)\right] \\
&= 2^n \sum_{x\in\mathbb{F}_2^{2n}} p_\psi(x) q_\psi(x) \\
&= 8^n \sum_{x\in\mathbb{F}_2^{2n}} p_\psi(x)(p_\psi * p_\psi)(x) \\
&= \frac{32^n}{4^n} \sum_{x\in\mathbb{F}_2^{2n}} \left[p_\psi(x)(p_\psi * p_\psi)(x)\right] \\
&= 32^n \sum_{x\in\mathbb{F}_2^{2n}} \widehat{p_\psi}(x)\widehat{p_\psi * p_\psi}(x)) && \text{(Fact 3.4)} \\
&= 32^n \sum_{x\in\mathbb{F}_2^{2n}} \widehat{p_\psi}(x)^3. && \text{(Proposition 3.6)} \\
&= 4^n \sum_{x\in\mathbb{F}_2^{2n}} p(x)^3 && \text{(Fact 3.7)}
\end{aligned}$$

□

5.1.1 Improved Completeness Analysis

Intuitively, η measures how concentrated the $p_\psi(x)$ are. Based on Lemma 3.11, we can again see that $\eta = 1$ for a stabilizer state, since the $p(x)$ are maximally concentrated. At a high level, if $|\psi\rangle$ is "close" to a stabilizer state then $p_\psi(x)$ should be "somewhat" concentrated. We will formalize this idea by lower-bounding η by stabilizer fidelity using the relationship developed in Section 4.4.3.

Proposition 5.3.

$$\eta \geq F_{\mathcal{S}}(|\psi\rangle)^6$$

Proof. Let $|\phi\rangle$ be a stabilizer state that maximizes the stabilizer fidelity, and let $S^* = \mathrm{Weyl}(|\phi\rangle)$.

$$\begin{aligned}
\eta &= 4^n \sum_{x \in \mathbb{F}_2^{2n}} p_\psi(x)^3 && \text{(Fact 5.2)} \\
&\geq 4^n \sum_{x \in S^*} p_\psi(x)^3 \\
&\geq \left(\sum_{x \in S^*} p_\psi(x) \right)^3 && \text{(Cauchy-Schwarz)} \\
&\geq F_{\mathcal{S}}(|\psi\rangle)^6 && \text{(Lemma 4.13)}
\end{aligned}$$

The second line follows from the fact that $p_\psi(x) \geq 0$ for all $x \in \mathbb{F}_2^{2n}$. □

5.1.2 Improved Soundness Analysis

At a high level, our proof is similar to [GNW21] and uses the results found in Section 4.4.4. The improvement comes from using the identity $\eta = 4^n \sum_x p_\psi(x)^3$, which was unknown to [GNW21]. Instead, they (implicitly) used the bound $\eta \leq 2^n \sum_x p_\psi(x)^2$ in their proof.

Proposition 5.4. *Let $|\psi\rangle$ be an n-qubit pure state. Then*

$$\frac{4\eta - 1}{3} \leq F_{\mathcal{S}}(|\psi\rangle).$$

Proof. Let $M \coloneqq \{x \in \mathbb{F}_2^{2n} : 2^n p_\psi(x) > 1/2\}$. Then

$$\begin{aligned}
\sum_{x \in M} p_\psi(x) &= \mathop{\mathbf{Pr}}_{x \sim p_\psi}[x \in M] \\
&= \mathop{\mathbf{Pr}}_{x \sim p_\psi}[2^n p(x) > 1/2] && \text{(Definition of } M\text{)} \\
&= \mathop{\mathbf{Pr}}_{x \sim p_\psi}[4^n p^2(x) > 1/4] \\
&= 1 - \mathop{\mathbf{Pr}}_{x \sim p_\psi}[4^n p^2(x) \le 1/4] \\
&= 1 - \mathop{\mathbf{Pr}}_{x \sim p_\psi}[1 - 4^n p^2(x) \ge 3/4] \\
&\ge 1 - \frac{4}{3}\left(1 - \mathop{\mathbf{E}}_{x \sim p_\psi}[4^n p^2(x)]\right) && \text{(Markov's Inequality)} \\
&= 1 - \frac{4}{3}(1 - \eta) \\
&= \frac{4\eta - 1}{3}.
\end{aligned}$$

To complete the proof, we note that M must commute by Fact 2.25. Because $p_\psi(x) \ge 0$, we can arbitrarily extend M to the Lagrangian subspace $S \supseteq M$ such that $\sum_{x \in S} p_\psi(x) \ge \frac{4\eta-1}{3}$. Finally, we apply Proposition 4.16 to lower-bound the stabilizer fidelity by

$$F_{\mathcal{S}}(|\psi\rangle) \ge \sum_{x \in S} p_\psi(x) \ge \frac{4\eta - 1}{3}. \qquad \square$$

5.2 Tolerantly Testing Stabilizer States

In the previous section, we proved that for all quantum states $|\psi\rangle$,

$$\frac{4\eta - 1}{3} \le F_{\mathcal{S}}(|\psi\rangle) \le \eta^{1/6}.$$

To simplify notation, let $\alpha_1 \coloneqq 1 - \varepsilon_1$ and $\alpha_2 \coloneqq 1 - \varepsilon_2$. Observe that if $F_{\mathcal{S}}(|\psi\rangle) \ge \alpha_1$ then $\eta \ge \alpha_1^6$, and if $F_{\mathcal{S}}(|\psi\rangle) \le \alpha_2$ then $\eta \le \frac{3\alpha_2+1}{4}$. This is the basis of our testing algorithm. Specifically, as long as

$$\alpha_1^6 - \frac{3\alpha_2 + 1}{4} \ge \frac{1}{\mathsf{poly}(n)},$$

then we can efficiently distinguish the two cases simply by estimating η. For the remainder of this section, define

$$\gamma \coloneqq \alpha_1^6 - \frac{3\alpha_2 + 1}{4}.$$

A more general form of the algorithm is stated as Algorithm 2.

Algorithm 2: η distinguishing

```
Input: $48\log(2/\delta)/\gamma^2$ copies of $|\psi\rangle$
Promise: Either case (i): $\eta \geq \beta_1$ or case (ii): $\eta \leq \beta_2$, for $\alpha_1, \alpha_2 \in [0,1]$ such that
         $\gamma > 0$
Output: 1 if case (i) holds and 0 if case (ii) holds, with probability at least $1 - \delta$
1 Let $m = \frac{8\log(2/\delta)}{\gamma^2}$.
2 repeat $m$ times
3     Perform Bell difference sampling to obtain $W_x \sim q_\psi$.
4     Perform the measurement $W_x^{\otimes 2}$ on $|\psi\rangle^{\otimes 2}$. Let $X_i \in \{\pm 1\}$ denote the
        measurement outcome.
5 Set $\widehat{\eta} = \frac{1}{m}\sum_i X_i$. Output 1 if $\widehat{\eta} > \beta_1 - \frac{\gamma}{2}$ and 0 otherwise.
```

Lemma 5.5. *For $\gamma > 0$ such that $\gamma = \beta_1 - \beta_2$, Algorithm 2 distinguishes between states with $\eta \geq \beta_1$ and states with $\eta \leq \beta_2$. It uses $48\log(2/\delta)/\gamma^2$ copies of the input state, $O(n\log(1/\delta)/\gamma^2)$ time, and succeeds with probability at least $1-\delta$.*

Proof. Algorithm 2 fails when $|\widehat{\eta} - \eta| \geq \gamma/2$. By the definition of $\eta = \mathbf{E}_{x \sim q_\psi}[2^n p_\psi(x)]$, each X_i is an unbiased estimator of η and so is $\widehat{\eta}$. Therefore, by Hoeffding's inequality (Fact 2.1),

$$\mathbf{Pr}[\text{Algorithm 2 fails}] = \mathbf{Pr}[|\widehat{\eta} - \eta| \geq \gamma/2] \leq 2e^{-m\gamma^2/8} = \delta.$$

The number of copies follows from the fact that Bell difference sampling consumes 4 copies of the input state, the measurement in Step 4 of Algorithm 2 consumes 2 copies of the input state, and that the loop is repeated m times. The running time is clearly $O(mn)$. □

We now apply our relationships between η and $F_{\mathcal{S}}$.

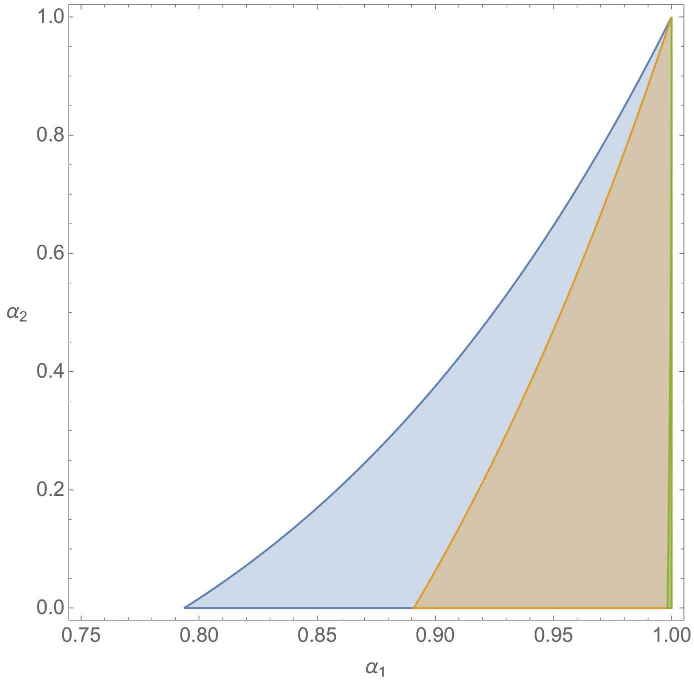


Figure 5.1: The shaded regions indicate the parameter regimes of α_1 and α_2 that are permissible by the GNW algorithm (green) and Algorithm 2 (blue). The orange region shows the parameter regime that is permissible if Algorithm 2 is analyzed using the looser bound $2\eta - 1 \leq F_{\mathcal{S}}(|\psi\rangle)$ from [GNW21]. Thus, the difference between the orange and blue regions illustrates the improvement due to Proposition 5.4.

Theorem 5.6. *For $\gamma > 0$ such that $\gamma = \alpha_1^6 - \frac{3\alpha_2+1}{4}$, Algorithm 2 distinguishes between states with $F_{\mathcal{S}}(|\psi\rangle) \geq \alpha_1$ and states with $F_{\mathcal{S}}(|\psi\rangle) \leq \alpha_2$. It uses $48\log(2/\delta)/\gamma^2$ copies of the input state, $O(n\log(1/\delta)/\gamma^2)$ time, and succeeds with probability at least $1-\delta$.*

Proof. Set $\beta_1 = \alpha_1^6$ and $\beta_2 = \frac{3\eta+1}{4}$ by Propositions 5.3 and 5.4 respectively. Then apply Lemma 5.5. □

5.3 Parameter Regime Discussion

We conclude this chapter by comparing the regime in which our algorithm works with prior work (namely, the GNW algorithm). We first establish the values of α_1 and α_2 in which the GNW algorithm works. As mentioned above, the GNW algorithm proved that for any quantum state $|\psi\rangle$, $2\eta - 1 \leq F_{\mathcal{S}}(|\psi\rangle)$. Additionally, since the GNW algorithm uses 6 copies of the input state and accepts stabilizer states with probability 1, it follows that $1 - 6\sqrt{1 - F_{\mathcal{S}}(|\psi\rangle)} \leq p_{\mathsf{accept}}$, where we use the subadditivity of the trace distance (see Fact 2.6), and the ability to convert to trace distance from fidelity and back for pure states using Fact 2.7. Finally, using the fact that $\eta = 2p_{\mathsf{accept}} - 1$, we get $F_{\mathcal{S}}(|\psi\rangle) \leq \frac{1}{144}(2\eta - \eta^2 + 143)$. Repeating the analysis from Section 5.2, we get that the GNW algorithm works as long as

$$1 - 12\sqrt{1 - \alpha_1} > \frac{\alpha_2 + 1}{2},$$

whereas, as shown earlier, our algorithm works as long as

$$\alpha_1^6 > \frac{3\alpha_2 + 1}{4}.$$

This is a significant improvement, which is shown visually in Fig. 5.1.

5.4 On the Tightness of Our Completeness Analysis

We now argue that the first part of Proposition 5.3 is polynomially-close to optimal. We begin by explicitly computing the stabilizer extent and stabilizer fidelity of Clifford magic states. The two technical ingredients involved in the computation are due to Bravyi et al. [BBC+19].

Fact 5.7 ([BBC+19, Proposition 2]). *Let $|\psi\rangle$ be a Clifford magic state. Then, $\xi(|\psi\rangle) = F_{\mathcal{S}}(|\psi\rangle)^{-1}$.*

Fact 5.8 ([BBC+19, Proposition 1]). *Let $\{|\psi_1\rangle, |\psi_2\rangle, \ldots, |\psi_L\rangle\}$ be any set of states such that each state $|\psi_j\rangle$ describes a system of at most 3 qubits. Then,*

$$\xi(|\psi_1\rangle \otimes |\psi_2\rangle \otimes \ldots \otimes |\psi_L\rangle) = \prod_i \xi(|\psi_i\rangle).$$

It is well known that the k-fold tensor product of $|T\rangle := \frac{|0\rangle + e^{i\pi/4}|1\rangle}{\sqrt{2}}$ is a Clifford magic state. Using the facts above, we can exactly compute the stabilizer extent and stabilizer fidelity of $|T^{\otimes n}\rangle$.

Fact 5.9.

$$\xi(|T^{\otimes n}\rangle) = \left(\cos\frac{\pi}{8}\right)^{-2n} \quad \textit{and} \quad F_{\mathcal{S}_m}(|T^{\otimes n}\rangle) = \left(\cos\frac{\pi}{8}\right)^{2n}.$$

Proof. By Fact 5.8, the stabilizer extent of $|T^{\otimes n}\rangle$ is simply the stabilizer extent of $|T\rangle$ raised to the power n. By Fact 5.7, the stabilizer extent is the inverse of the stabilizer fidelity. Hence, the result follows simply by showing that the stabilizer fidelity of $|T\rangle$ is $\cos^2\frac{\pi}{8}$, which can be verified by explicit calculation over the 6 different 1-qubit stabilizer states. □

Next, we compute η for the state $|T^{\otimes n}\rangle$.

Claim 5.10. *Let $|\psi\rangle = |T^{\otimes n}\rangle$ and define $\eta := \mathbf{E}_{x\sim q_\psi}[2^n p_\psi(x)]$. Then, $\eta = (5/8)^n$.*

Proof. We begin by writing out $|T\rangle\langle T|$ as a sum of Pauli matrices. By definition,

$$|T\rangle\langle T| = \frac{1}{2}\left(I + \frac{1}{\sqrt{2}}X + \frac{1}{\sqrt{2}}Y\right).$$

We wish to compute $\sum_{x\in\mathbb{F}_2^{2n}} p_\psi(x)^3$. We know that every such Pauli with nonzero $p_\psi(x)$ is a tensor product combination of I, X, and Y, so we enumerate over the number of indices where an X or Y appear.

$$\sum_{x\in\mathbb{F}_2^{2n}} p_\psi(x)^3 = \frac{1}{2^{3n}}\sum_{k=0}^{n}\binom{n}{k}\frac{1}{2^{3k}}\cdot 2^k = \frac{1}{8^n}\sum_{k=0}^{n}\binom{n}{k}\frac{1}{4^k} = \left(\frac{5}{32}\right)^n.$$

Thus, by Fact 5.2,

$$\eta = 4^n \sum_{x\in\mathbb{F}_2^{2n}} p_\psi(x)^3 = \left(\frac{5}{8}\right)^n.$$ □

Combining Claim 5.10 with Proposition 5.3, we have

$$F_{\mathcal{S}}(|\psi\rangle) \leq \eta^{1/c} = \left(\frac{5}{8}\right)^{n/c}$$

for $c = 6$. Improving our completeness amounts to lowering the value of c. But, from Fact 5.9, we know that $F_{\mathcal{S}}(|T^{\otimes n}\rangle) = \left(\cos\frac{\pi}{8}\right)^{2n}$. Combining the two statements gives

$$\left(\cos\frac{\pi}{8}\right)^{2n} \leq (5/8)^{n/c}.$$

$c = \frac{\log \frac{5}{8}}{2 \log \cos(\frac{\pi}{8})} \approx 2.97$ is then minimum c that does not violate this inequality. Hence, one cannot hope for much more than a quadratic improvement in our completeness analysis.

5.5 Improvements via State Preparation Unitary

When given access to a state preparation unitary for $|\psi\rangle$ (and its inverse), denoted by U and $U^\dagger$, we can improve the sample and time complexities of our algorithm to $O(\log(1/\delta)/\gamma)$ and $O(n\log(1/\delta)/\gamma)$, respectively, at the cost of $O(\log(1/\delta)/\gamma)$ queries to U and $U^\dagger$.

Access to U and $U^\dagger$ allows us to run quantum amplitude estimation (QAE) as a subroutine in our algorithm. Recall the well-known result of Brassard, Høyer, Mosca, and Tapp:

Theorem 5.11 (Quantum Amplitude Estimation (Theorem 12 in [BHMT02]))**.** *Let Π be a projector and $|\psi\rangle$ be an n-qubit pure state such that $\langle\psi|\Pi|\psi\rangle = \eta$. Given access to the unitary transformations $R_\Pi = 2\Pi - I$ and $R_\psi = 2|\psi\rangle\langle\psi| - I$, there exists a quantum algorithm that outputs $\widehat{\eta}$ such that*

$$|\widehat{\eta} - \eta| \leq \frac{2\pi\sqrt{\eta(1-\eta)}}{m} + \frac{\pi^2}{m^2}$$

with probability at least $\frac{8}{\pi^2}$. The algorithm makes m calls to R_Π and R_ψ.

Corollary 5.12. *Let Π, $|\psi\rangle$, R_Π, and R_ψ be the same as in Theorem 5.11. There exists a quantum algorithm that outputs $\widehat{\eta}$ such that*

$$|\widehat{\eta} - \eta| \leq \varepsilon$$

with probability at least $\frac{8}{\pi^2}$. The algorithm makes no more than

$$\pi\frac{\sqrt{\eta(1-\eta)+\varepsilon}}{\varepsilon}$$

calls to R_Π and R_ψ.

Proof. By Theorem 5.11, this will require m queries, where m is a solution to the following quadratic equation:

$$\frac{2\pi\sqrt{\eta(1-\eta)}}{m} + \frac{\pi^2}{m^2} \leq \varepsilon \Rightarrow m \geq \pi\frac{\sqrt{\eta(1-\eta)+\varepsilon}}{\varepsilon} \geq \pi\frac{\sqrt{\eta(1-\eta)}+\sqrt{\eta(1-\eta)+\varepsilon}}{2\varepsilon}.$$

□

With that, we are ready to explain the modifications to Algorithm 2 that achieves a quadratic speedup in the dependency on γ. Recall the Bell difference sampling projectors on measurement outcome $x \in \mathbb{F}_2^{2n}$ as[1]

$$\Pi_x := \sum_{y \in \mathbb{F}_2^{2n}} |W_y\rangle\langle W_y| \otimes |W_{x+y}\rangle\langle W_{x+y}|$$

such that $q_\psi(x) = \|\Pi_x |\psi^{\otimes 4}\rangle\|$.[2] We can also perform the projective measurement $P_{\psi,x} := W_x |\psi\rangle\langle\psi| W_x = W_x U |0\rangle\langle 0| U^\dagger W_x$, where this measurement is performed by applying W_x, $U^\dagger$, and then measuring in the computational basis. We can entangle Π_x and $P_{\psi,x}$ to form the following projector:

$$M = \sum_{x \in \mathbb{F}_2^{2n}} \Pi_x \otimes P_{\psi,x}.$$

Building M involves controlled applications of W_x according to the Bell difference sampling outcome. Observe that

$$\langle\psi^{\otimes 5}|M|\psi^{\otimes 5}\rangle = \sum_{x \in \mathbb{F}_2^{2n}} \langle\psi^{\otimes 4}|\Pi_x|\psi^{\otimes 4}\rangle \cdot \langle\psi|P_{\psi,x}|\psi\rangle = \mathop{\mathbf{E}}_{x\sim q_\psi}\left[|\langle\psi|W_x|\psi\rangle|^2\right] = \eta.$$

Hence, we can run QAE with the input projector M and the input state $|\psi^{\otimes 5}\rangle$, and the output will be an estimate of η whose accuracy depends on m, the number of total calls to R_Π and R_ψ.

[1] See Section 3.2 for details and see Eq. (3.1) for a definition of $|W_x\rangle$.

[2] Indeed, this is the way Gross, Nezami, and Walter [GNW21] introduce Bell difference sampling.

5.5.1 Additive Distance

We now show how to use Quantum Amplitude Estimation to achieve a quadratic speedup in terms of γ. Like Section 5.2, we start with the more general problem of distinguishing η then use Propositions 5.3 and 5.4 to relate it to fidelity.

Lemma 5.13. *Let $\beta_1 > \beta_2$ be parameters in $[0,1]$ let $\gamma = \beta_1 - \beta_2$. Let $|\psi\rangle$ be an unknown n-qubit pure state prepared by a unitary U. There exists a quantum algorithm that distinguishes whether $|\psi\rangle$ is a state with η at least β_1 or a state with η less than β_2, promised that one of these is the case. The algorithm uses $O\left(\log(1/\delta)\frac{\sqrt{\gamma+\max\{\beta_1(1-\beta_1),\beta_2(1-\beta_2\}}}{\gamma}\right)$ applications of either U or $U^\dagger$ and time $O\left(n\log(1/\delta)\frac{\sqrt{\gamma+\max\{\beta_1(1-\beta_1),\beta_2(1-\beta_2\}}}{\gamma}\right)$, and distinguishes the two cases with success probability at least $1-\delta$.*

Proof. The algorithm is the same as Algorithm 2, but with estimation of η done using state preparation unitaries.

Proving the sample complexity bound will mimic Theorem 5.6. Suppose $|\psi\rangle$ is a state with η at least β_1. For our algorithm to succeed, recall from the proof of Theorem 5.6 that

$$|\eta - \widehat{\eta}| \leq |\eta - \frac{\beta_1+\beta_2}{2}| \Rightarrow \widehat{\eta} \geq \beta_1 - \gamma/2 = \frac{\beta_1+\beta_2}{2}.$$

Therefore, we can run QAE with a fixed value of m (to be specified later) for an estimate of η whose accuracy is within $\pm\left(\eta - \frac{\beta_1+\beta_2}{2}\right)$. By Corollary 5.12,

$$m \geq \pi\frac{\sqrt{\eta(1-\eta)+\eta-\frac{\beta_1+\beta_2}{2}}}{\eta-\frac{\beta_1+\beta_2}{2}} \tag{5.1}$$

queries suffice. The chosen value of m must work for all $\eta \in [\beta_1, 1]$. Note that Eq. (5.1) is monotonically decreasing for $\eta \in [\beta_1, 1)$, and is therefore maximized by β_1 within that range. To succeed with probability at least $\frac{8}{\pi^2}$,

$$m \geq \pi\frac{\sqrt{\beta_1(1-\beta_1)+\gamma}}{\gamma} \geq \pi\frac{\sqrt{\beta_1(1-\beta_1)+\beta_1-\frac{\beta_1+\beta_2}{2}}}{\beta_1-\frac{\beta_1+\beta_2}{2}}$$

calls to R_Π and R_ψ suffices.

Now suppose $|\psi\rangle$ is a state with $\eta \leq \beta_2$. Using Corollary 5.12 as long as we have

$$m \geq \pi \frac{\sqrt{\beta_2(1-\beta_2)+\gamma}}{\gamma} \geq \pi \frac{\sqrt{\beta_2(1-\beta_2)+\frac{\beta_1+\beta_2}{2}-\beta_2}}{\frac{\beta_1+\beta_2}{2}-\beta_2} \geq \pi \frac{\sqrt{\eta(1-\eta)+\frac{\beta_1+\beta_2}{2}-\eta}}{\frac{\beta_1+\beta_2}{2}-\eta}$$

queries to R_Π and R_ψ, we obtain the correct answer with probability at least $\frac{8}{\pi^2}$. In the inequalities above we use similar reasoning about monotonicity in $[0, \beta_2]$ similar to the completeness case.

We will simply take the larger of these two lower bounds such that both conditions are met. Since R_Π and R_ψ use a constant number of calls to U and $U^\dagger$, the total number of calls is $O(\frac{\sqrt{\gamma+\max\{\beta_1(1-\beta_1),\beta_2(1-\beta_2\}}}{\gamma})$. Hoeffding's inequality (Fact 2.1) can be used to bring the success probability from 3/4 to $1-\delta$ using $6\ln(1/\delta)$ repetitions. The runtime includes an extra factor of $O(n)$, due to the linear cost of both preparing W_x and the Bell difference sampling projector, giving a $O\left(n\log(1/\delta)\frac{\sqrt{\gamma+\max\{\beta_1(1-\beta_1),\beta_2(1-\beta_2\}}}{\gamma}\right)$ time complexity. □

Using our knowledge of how stabilizer fidelity can be used to bound η, we then recover the following quadratic speedup in terms of γ.

Corollary 5.14. *For $\gamma > 0$ such that $\gamma = \alpha_1^6 - \frac{3\alpha_2+1}{4}$, let $|\psi\rangle$ be an unknown n-qubit pure state prepared by a unitary U. There exists a quantum algorithm that distinguishes whether $|\psi\rangle$ is a state with stabilizer fidelity at least α_1 or a state with fidelity less than α_2, promised that one of these is The case. The algorithm uses $O(\log(1/\delta)/\gamma)$ applications of either U or $U^\dagger$ and time $O(n\log(1/\delta)/\gamma)$, and distinguishes the two cases with success probability at least $1-\delta$.*

Proof. By Proposition 5.3 we can set $\beta_1 = \alpha_1^6$, and by Proposition 5.4 we can set $\beta_2 = \frac{3\alpha_2+1}{4}$. Finally, we note that for all $x \in [0,1], x(1-x) \leq \frac{1}{4}$, giving us the desired bounds when applying Lemma 5.13.

□

5.5.2 Multiplicative Distance

Let us now imagine that β_1 and β_2 are multiplicatively related in our η distinguishing problem. With just black-box access to the state, Lemma 5.5, one would expect a scaling of $O(1/\beta_1^2)$ sample complexity. We show how to use a state preparation unitary to achieve a quartic speedup of $O(1/\sqrt{\beta_1})$ as compared to black-box access to quantum states. The first quadratic speedup simply comes from naïve application of Quantum Amplitude Estimation like with Corollary 5.14. The second speedup comes from the realization that the outcomes of the random variables are now $\{0, 1\}$ rather than $\{\pm 1\}$. In the right regimes this causes the variance of $\widehat{\eta}$ to decrease significantly, allowing for faster estimation.

Corollary 5.15. *Let $\mu < 1$ be an arbitrary positive constant bounded away from 1 (i.e., $\frac{1}{2}, \frac{2}{3}, \frac{99}{100}$, etc.) and let $|\psi\rangle$ be an unknown n-qubit pure state prepared by a unitary U. There exists a quantum algorithm that distinguishes whether $|\psi\rangle$ is a state with η at least β or a state with η less than $\mu\beta$, promised that one of these is the case. The algorithm uses $O\left(\log(1/\delta)/\sqrt{\beta}\right)$ applications of either U or $U^\dagger$ and time $O\left(n \log(1/\delta)/\sqrt{\beta}\right)$, and distinguishes the two cases with success probability at least $1 - \delta$.*

Proof. If we apply Lemma 5.13 with $\alpha_1 = \beta$ and $\alpha_2 = \mu\alpha$, we note that $\gamma = \frac{1-\mu}{2}\beta = \Theta(\beta)$. Using the fact that $\beta \geq \beta(1-\beta)$ for $\beta \in [0, 1]$, we get the desired bounds.

□

5.6 Discussion and Open Problems

Can tighter bounds between η and stabilizer fidelity be proven? In Section 5.4, we proved that one can hope for at most a roughly quadratic improvement in the bound $F_{\mathcal{S}}(|\psi\rangle)^6 \leq \eta$. Progress in this direction would extend the parameter regimes for which our property testing algorithm works (see Fig. 5.1). Likewise, can the soundness case be improved? It seems that either a better understanding of higher moments of p_ψ or an entirely new proof technique would be required.

In addition to η, are there other statistics related to stabilizer fidelity (or any other stabilizer complexity measure) that can be estimated efficiently? We note that, assuming the existence of quantum-secure one way functions, the results of [ABF+22] show that no statistics can differentiate states with fidelity inverse super-polynomial from states with extremely inverse exponential fidelity (specifically Haar random states). See Chapter 8 for details.

Chapter 6

Stabilizer State Approximations

This chapter is based on Section 5 of [GIKL23b], which was joint work with Sabee Grewal, Vishnu Iyer, and William Kretschmer. Some of the preliminary results were moved to Chapters 2 to 4.

We give an algorithm for estimating stabilizer fidelity (Definition 2.35). Recall that it is simply the maximum of $|\langle\phi|\psi\rangle|^2$ over all stabilizer states $|\phi\rangle$. Assuming $|\psi\rangle$ has stabilizer fidelity at least τ, our algorithm returns a succinct description of a stabilizer state (i.e., generators of its stabilizer group) that witnesses overlap at least $\tau - \varepsilon$ with $|\psi\rangle$.

Theorem 6.1 (Informal version of Theorem 6.9)**.** *Let $|\psi\rangle$ be an n-qubit pure state and fix $\tau > \varepsilon > 0$. If the stabilizer fidelity of $|\psi\rangle$ is at least τ, there is an algorithm that returns a stabilizer state $|\phi\rangle$ that satisfies $|\langle\phi|\psi\rangle|^2 \geq \tau - \varepsilon$. The algorithm uses $O(n/(\varepsilon^2\tau^4))$ copies of $|\psi\rangle$ and $\exp\left(O(n/\tau^4)\right)/\varepsilon^2$ time.*

To our knowledge, this is the first nontrivial algorithm to approximate an arbitrary quantum state with a stabilizer state. Indeed, we are not aware of any prior algorithm better than a brute-force search over all stabilizer states, which takes $2^{O(n^2)}$ time and $O(n^2)$ samples.[1] Thus our algorithm offers a substantial improvement in the regime of $\tau = \omega(n^{-1/4})$. Arguably, the most interesting setting of parameters is constant τ, in which case we have a quadratic improvement in sample complexity and a superpolynomial improvement in time complexity.

[1] The polynomial sample complexity follows from a straightforward application of the classical shadows framework [HKP20]. See [Gro06, Corollary 21] for a proof that there are $2^{O(n^2)}$ many stabilizer states.

Observe that, because we output a witness of stabilizer fidelity at least $\tau - \varepsilon$ with high probability, assuming a state with fidelity τ exists, our algorithm can be used as a subroutine to *estimate* stabilizer fidelity and, moreover, find a stabilizer state that witnesses this. More precisely, if the goal is to estimate stabilizer fidelity to accuracy $\pm\varepsilon$, then one can break $[0, 1]$ into intervals of width ε and perform a binary search procedure using our algorithm. Overall, this takes $O(n/\varepsilon^6)$ samples and $\exp(O(n/\varepsilon^4))$ time.

As an application, our stabilizer fidelity estimation algorithm could be used to search for better stabilizer decompositions of magic states. Recall that magic states are states that, when injected into Clifford circuits, allow for the simulation of universal quantum computation [BK05]. The best-known algorithms for simulating quantum circuits dominated by Clifford gates use decompositions of magic states into linear combinations of stabilizer states, and have a runtime that scales polynomially in the complexity of the decomposition [BBC+19]. Hence, better stabilizer decompositions of magic states yield faster algorithms. These decompositions are often obtained by writing the tensor product of a small number of magic states (usually on the order of 10 qubits) as a slightly larger number of stabilizer states [BSS16, Koc22]. Therefore, if our algorithm could be made practical for (say) $n \approx 15$ qubits, there is reason to believe that running our algorithm on magic states, combined with a meta-algorithm such as matching pursuit [MZ94], could find better stabilizer decompositions of magic states and, as a result, improve the runtime of near-Clifford simulation.

Finally, we remark that the problem we solve is similar in spirit to the agnostic probably approximately correct (PAC) learning framework [Val84, KSS92]. In the agnostic PAC model, a learner is given labeled training data $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ from some unknown distribution $\mathcal{D}$, as well as some concept class $\mathcal{C}$ to choose a hypothesis from. The goal of the learner is to find a hypothesis function $h \in \mathcal{C}$ that approximates the best fit for the training data, even though no function in $\mathcal{C}$ will necessarily fit the training data perfectly. In an analogous fashion, our algorithm finds a stabilizer state $|\phi\rangle$ that approximates the best fit for $|\psi\rangle$ over the set of stabilizer states, which need not contain $|\psi\rangle$. We note that Aaronson studied PAC learning of quantum states in the so-called realizable setting [Aar07]. However, agnostic PAC learning of quantum states has not yet appeared in the literature.

Our presentation is split into two parts. First, in Section 6.1, we prove a useful lemma regarding q_ψ on $S^* = \mathrm{Weyl}(|\phi\rangle)$, where $|\phi\rangle$ is the stabilizer state that maximizes stabilizer fidelity with $|\psi\rangle$. At a high level, we argue that any sample from q_ψ has a good enough chance of "making progress" towards learning a complete set of generators for S^*. Formally, we prove that the q_ψ-mass on S^* is not heavily concentrated on proper subspaces of S^*, so that when we sample an element of S^*, we obtain an element of S^* that is linearly independent of the previous samples with a reasonable probability. Second, in Section 6.2, we state our algorithm, prove its correctness, and analyze its sample and time complexities.

6.1 Stabilizer Fidelity Implies Anticoncentration of q_ψ

We show a series of anticoncentration[2] results (for both p_ψ and q_ψ) on proper subspaces of S^*. Just as with Section 4.4, for these next lemmas we will find it more convenient to assume without loss of generality (because of Lemmas 4.9 and 4.10) that the state maximizing fidelity is $|0^n\rangle$, which conceptually simplifies the computations.

We start by showing that the c_ψ values (see Definition 2.16) are anticoncentrated.

Lemma 6.2. *Let $|\psi\rangle$ be an n-qubit state. Suppose the fidelity $|\langle\psi|\phi\rangle|^2$ is maximized by $|\phi\rangle = |0^n\rangle$ over stabilizer states $|\phi\rangle$. Let $S^* = 0^n \times \mathbb{F}_2^n = \mathrm{Weyl}(|0^n\rangle)$, and let $T = 0^{n+1} \times \mathbb{F}_2^{n-1}$ be a maximal subspace of S^*. Then*

$$\sum_{x \in S^* \setminus T} c_\psi(x) \geq 2^{\frac{n}{2}-1}\left(\sqrt{3}-1\right) F_{\mathcal{S}}(|\psi\rangle).$$

[2] We remark that Section 4.4.3 is focused on *concentration* instead.

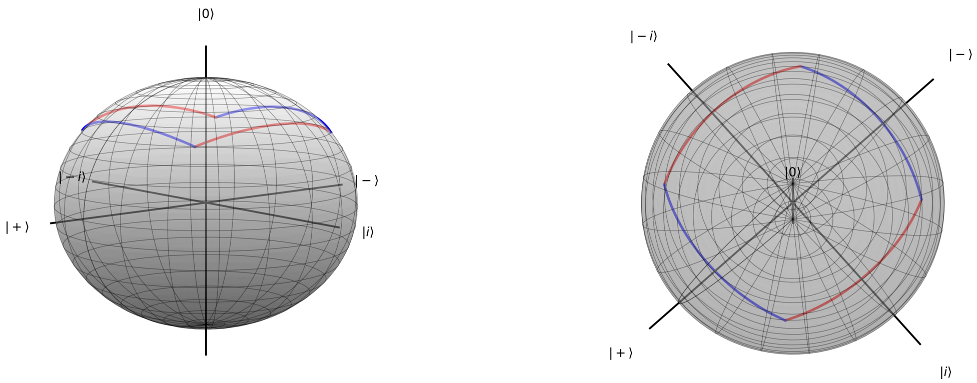


Figure 6.1: An illustration of the argument in the proof of Lemma 6.2. Consider the (possibly unnormalized) state $\alpha_0 |0^n\rangle + \alpha_1 |10^{n-1}\rangle$. We can visualize the first qubit of this state on the Bloch sphere. The surface enclosed by the red and blue curve is exactly the set of points on the sphere for which $|0\rangle$ is the closest stabilizer state. By our assumption that the stabilizer fidelity of $|\psi\rangle$ is maximized by $|0^n\rangle$, $\alpha_0 |0\rangle + \alpha_1 |1\rangle$ must lie on this surface, up to normalization. The corners of this surface (the intersection of a blue curve with a red curve) represent a choice of α_0 and α_1 that minimizes α_0.

Proof. We can express the sum as

$$\begin{aligned}\sum_{x \in S^* \setminus T} c_\psi(x) &= \frac{1}{\sqrt{2^n}} \sum_{x \in 1 \times \mathbb{F}_2^{n-1}} \mathrm{Tr}\left[|\psi\rangle\langle\psi| \, W_{0^n,x}\right] \\ &= 2^{\frac{n}{2}-1} \mathrm{Tr}\left[|\psi\rangle\langle\psi| \left(Z \otimes |0\rangle\langle 0|^{n-1}\right)\right] \\ &= 2^{\frac{n}{2}-1} \left(|\alpha_0|^2 - |\alpha_1|^2\right),\end{aligned}$$

where α_0 is the amplitude of $|\psi\rangle$ on $|0^n\rangle$ and α_1 is its amplitude on $|10^{n-1}\rangle$. Note that $|\alpha_0|^2 = F_{\mathcal{S}}(|\psi\rangle)$, by assumption. Thus, we need to show that $|\alpha_1|$ cannot be too big compared to $|\alpha_0|$, or else it would contradict the maximality of $|\langle\psi|\phi\rangle|^2$ at $|\phi\rangle = |0^n\rangle$. We give a visual proof of this fact in Figure 6.1, along with an algebraic proof below.

Choose the global phase on $|\psi\rangle$ to assume without loss of generality that α_0 is positive-real and $\alpha_1 = |\alpha_1|e^{i\theta}$. We may write:

$$\begin{aligned}
|\left(\langle +| \otimes \langle 0^{n-1}|\right) |\psi\rangle|^2 &= \frac{1}{2}|\alpha_0 + \alpha_1|^2 = \frac{1}{2}\left((\alpha_0 + |\alpha_1|\cos\theta)^2 + |\alpha_1|^2 \sin^2\theta\right) \\
|\left(\langle -| \otimes \langle 0^{n-1}|\right) |\psi\rangle|^2 &= \frac{1}{2}|\alpha_0 - \alpha_1|^2 = \frac{1}{2}\left((\alpha_0 - |\alpha_1|\cos\theta)^2 + |\alpha_1|^2 \sin^2\theta\right) \\
|\left(\langle i| \otimes \langle 0^{n-1}|\right) |\psi\rangle|^2 &= \frac{1}{2}|\alpha_0 + i\alpha_1|^2 = \frac{1}{2}\left((\alpha_0 - |\alpha_1|\sin\theta)^2 + |\alpha_1|^2 \cos^2\theta\right) \\
|\left(\langle -i| \otimes \langle 0^{n-1}|\right) |\psi\rangle|^2 &= \frac{1}{2}|\alpha_0 - i\alpha_1|^2 = \frac{1}{2}\left((\alpha_0 + |\alpha_1|\sin\theta)^2 + |\alpha_1|^2 \cos^2\theta\right).
\end{aligned}$$

All of these values need to be less than $|\alpha_0|^2$, as otherwise $|\psi\rangle$ would have larger fidelity with one of the above states. Due to symmetry of both sin and cos, we will only consider $\theta \in [0, \frac{\pi}{2}]$ such that the only relevant equations to consider are the first and last. This allows us to write the largest of the above inner products as

$$\frac{1}{2}\left(\alpha_0^2 + |\alpha_1|^2 + 2\alpha_0|\alpha_1| \cdot \max(\cos\theta, \sin\theta)\right),$$

which is minimized for $\theta = \pi/4$. Plugging that back in and comparing to α_0^2 leads to

$$\alpha_0^2 \geq \frac{1}{2}\left(\alpha_0^2 + |\alpha_1|^2 + \sqrt{2}\alpha_0|\alpha_1|\right),$$

and solving for the maximum $|\alpha_1|$ gives $|\alpha_1| \leq \left(2 - \sqrt{3}\right)^{1/2}|\alpha_0|$. Hence, $|\alpha_0|^2 - |\alpha_1|^2 \geq 1 - (2 - \sqrt{3})\alpha_0^2 = \left(\sqrt{3} - 1\right) F_{\mathcal{S}}(|\psi\rangle)$. Therefore,

$$\sum_{x \in S^* \setminus T} c_\psi(x) = 2^{\frac{n}{2}-1}\left(|\alpha_0|^2 - |\alpha_1|^2\right) \geq 2^{\frac{n}{2}-1}\left(\sqrt{3} - 1\right) F_{\mathcal{S}}(|\psi\rangle). \qquad \square$$

We can now use Cauchy-Schwarz to show that p_ψ must naturally also be anticoncentrated.

Lemma 6.3. *Let $|\psi\rangle$ be an n-qubit state. Suppose the fidelity $|\langle\psi|\phi\rangle|^2$ is maximized by $|\phi\rangle = |0^n\rangle$ over stabilizer states $|\phi\rangle$. Let $S^* = 0^n \times \mathbb{F}_2^n = \mathrm{Weyl}(|0^n\rangle)$, and let $T = 0^{n+1} \times \mathbb{F}_2^{n-1}$ be a maximal subspace. Then*

$$\sum_{x \in S^* \setminus T} p_\psi(x) \geq \left(2 - \sqrt{3}\right) F_{\mathcal{S}}(|\psi\rangle)^2.$$

Proof.

$$\begin{aligned}
\sum_{x\in S^*\setminus T} p_\psi(x) &\geq \frac{1}{2^{n-1}}\left(\sum_{x\in S^*\setminus T} |c_\psi(x)|\right)^2 && \text{(Cauchy-Schwarz)}\\
&\geq \frac{1}{2^{n-1}}\left(\sum_{x\in S^*\setminus T} c_\psi(x)\right)^2 && \text{(Triangle Inequality)}\\
&\geq \frac{1}{2^{n-1}}\left(2^{\frac{n}{2}-1}\left(\sqrt{3}-1\right)F_{\mathcal{S}}(|\psi\rangle)\right)^2 && \text{(Lemma 6.2)}\\
&= \left(2-\sqrt{3}\right)F_{\mathcal{S}}(|\psi\rangle)^2. && \square
\end{aligned}$$

Due to the structure of the convolution, we now combine *both* concentration and anticoncentration of p_ψ to give anticoncentration of q_ψ.

Lemma 6.4. *Let $|\psi\rangle$ be an n-qubit state. Suppose the fidelity $|\langle\psi|\phi\rangle|^2$ is maximized by $|\phi\rangle = |0^n\rangle$ over stabilizer states $|\phi\rangle$. Let $S^* = 0^n \times \mathbb{F}_2^n = \mathrm{Weyl}(|0^n\rangle)$, and let $T = 0^{n+1}\times\mathbb{F}_2^{n-1}$ be a maximal subspace. Then*

$$\sum_{x\in S^*\setminus T} q_\psi(x) \geq \frac{2-\sqrt{3}}{2}F_{\mathcal{S}}(|\psi\rangle)^4.$$

Proof. We can write

$$\begin{aligned}
\sum_{x\in S^*\setminus T} q_\psi(x) &= \sum_{x\in S^*\setminus T}\sum_{t\in\mathbb{F}_2^{2n}} p_\psi(t)p_\psi(x+t)\\
&\geq \sum_{t\in T} p_\psi(t)\sum_{x\in S^*\setminus T} p_\psi(x+t)\\
&= \left(\sum_{t\in T} p_\psi(t)\right)\left(\sum_{x'\in S^*\setminus T} p_\psi(x')\right) && (t+S^*\setminus T = S^*\setminus T)
\end{aligned}$$

Now apply Corollary 4.14 and Lemma 6.3 respectively and we get

$$\sum_{x\in S^*\setminus T} q_\psi(x) \geq \frac{|T|}{2^n}F_{\mathcal{S}}(|\psi\rangle)^2\cdot\left(2-\sqrt{3}\right)F_{\mathcal{S}}(|\psi\rangle)^2 = \frac{2-\sqrt{3}}{2}F_{\mathcal{S}}(|\psi\rangle)^4$$

as desired. □

Finally, we use Lemmas 4.9 and 4.10 once again to generalize the result for arbitrary maximizing stabilizer states.

Lemma 6.5. *Given an n-qubit state $|\psi\rangle$, let $|\phi\rangle$ be a stabilizer state that maximizes the stabilizer fidelity, and let $S^* = \mathrm{Weyl}(|\phi\rangle)$. Let $T \subset S^*$ be a proper subspace of S^*. Then*

$$\sum_{x \in S^* \setminus T} q_\psi(x) \geq \frac{2-\sqrt{3}}{2} F_{\mathcal{S}}(|\psi\rangle)^4.$$

Proof. Use Lemma 4.9 to choose a Clifford circuit such that $C|\phi\rangle = |0^n\rangle$, $C(S^*) = 0^n \times \mathbb{F}_2^n$, and $C(T) \subseteq 0^{n+1} \times \mathbb{F}_2^{n-1}$. Let $|\psi'\rangle = C|\psi\rangle$. Then by Lemma 4.10:

$$\sum_{x \in S^* \setminus T} q_\psi(x) = \sum_{x \in C(S^* \setminus T)} q_{\psi'}(x) \geq \sum_{x \in 0^n \times 1 \times \mathbb{F}_2^{n-1}} q_{\psi'}(x).$$

By Lemma 6.4, this sum is lower bounded by

$$\sum_{x \in 0^n \times 1 \times \mathbb{F}_2^{n-1}} q_{\psi'}(x) \geq \frac{2-\sqrt{3}}{2} F_{\mathcal{S}}(|\psi'\rangle)^4 = \frac{2-\sqrt{3}}{2} F_{\mathcal{S}}(|\psi\rangle)^4.$$ □

6.2 The Algorithm

Our algorithm for estimating stabilizer fidelity uses the powerful classical shadows framework [HKP20] to improve its sample complexity.

Theorem 6.6 (Classical shadows algorithm [HKP20])**.** *Let ρ be an unknown n-qubit mixed state. Then there exists a quantum algorithm that first performs $m_{\mathsf{shadow}} = O(\log(K/\delta)/\varepsilon^2)$ random Clifford measurements on independent copies of ρ. Then, later given K different observables $O_1, O_2, \ldots, O_K$ in an online fashion, where each O_i is a rank-1 projector, the algorithm uses the measurement results to output estimates $\widehat{o}_1, \ldots, \widehat{o}_K$, such that with probability at least $1-\delta$, for every $i \in [K]$, $|\widehat{o}_i - \mathrm{Tr}(O_i\rho)| \leq \varepsilon$. Moreover, if O_i is a projector onto a stabilizer state, then each $\widehat{o}_i$ can be computed from the measurement results by a classical algorithm that takes time $O(n^2 m_{\mathsf{shadow}})$.*

For the "moreover" part of Theorem 6.6, see the remarks on Page 1053 of [HKP20].

We also require an algorithm, due to [TTT06], for computing all of the maximal cliques in a graph.

Theorem 6.7 (Computing maximal cliques [TTT06]). *Given an undirected graph G with n vertices, there is a classical algorithm that outputs a list of all of the maximal cliques in G in time $O(3^{n/3})$.*

Note that this implies that the number of maximal cliques is at most $O(3^{n/3})$.

We are now ready to describe the fidelity estimation algorithm. At a high level, it uses Bell difference sampling to obtain a list of candidate Lagrangian subspaces generated by the sampled Weyl operators. Then, it iterates through the candidate groups to find the stabilizer state with largest fidelity, using classical shadows to perform the estimation.

We first argue that with high probability, one of the maximal cliques generates the Lagrangian subspace corresponding to a state that maximizes stabilizer fidelity.

Lemma 6.8. *Given an n-qubit state $|\psi\rangle$, let $|\phi\rangle$ be a stabilizer state that maximizes the stabilizer fidelity, and let $S^* = \mathrm{Weyl}(|\phi\rangle)$. Suppose $|\langle\phi|\psi\rangle|^2 \geq \tau$. Then choosing $m_{\mathsf{clique}} \geq \frac{8+4\sqrt{3}}{\tau^4}(n + \log(1/\delta))$ is sufficient to guarantee that with probability at least $1-\delta$, the Bell difference sampling step of algorithm Algorithm 3 samples a complete set of generators for S^*.*

Proof. By Lemmas 3.10 and 6.5 with $c = \frac{2-\sqrt{3}}{2}\tau^4$, we only need $\frac{8+4\sqrt{3}}{\tau^4}(n+\log\frac{1}{\delta})$ Bell difference samples to find generators of S^* with probability at least $1-\delta$. □

Now we have everything needed to prove the correctness of Algorithm 3.

Theorem 6.9. *Let $|\psi\rangle$ be an n-qubit state with $F_{\mathcal{S}}(|\psi\rangle) \geq \tau$. Then choosing*

$$m_{\mathsf{clique}} = \frac{8+4\sqrt{3}}{\tau^4}(n + \log(2/\delta)) \qquad\qquad m_{\mathsf{shadow}} = O\left(\frac{n+\log(1/\delta)}{\varepsilon^2\tau^4}\right)$$

Algorithm 3: Estimating Stabilizer Fidelity

```
Input: $m_{\mathsf{shadow}} + 4m_{\mathsf{clique}}$ copies of $|\psi\rangle$
Promise: $|\psi\rangle$ has stabilizer fidelity at least $\tau$
Output: A stabilizer state $|\phi\rangle$ such that $|\langle\phi|\psi\rangle|^2 \geq \tau - \varepsilon$ with probability at
        least $1 - \delta$
Initialize an empty graph $G$
repeat $m_{\mathsf{clique}}$ times
    Using 4 copies of $|\psi\rangle$, perform Bell difference sampling to obtain $x \in \mathbb{F}_2^{2n}$
    Add a vertex for $x$ in $G$ and connect it to all vertices $y$ in $G$ such that
      $[x, y] = 0$.
repeat $m_{\mathsf{shadow}}$ times
    Choose a random Clifford circuit $U$
    Measure $U|\psi\rangle$ in the computational basis and store the result
foreach maximal clique $(v_1, \dots, v_k) \in G$ computed using Theorem 6.7 do
    Compute $S \coloneqq \langle v_1, \dots, v_k\rangle$ via Gaussian elimination
    if $|S| = 2^n$ then
        foreach stabilizer state $|\phi\rangle$ with $\mathrm{Weyl}(|\phi\rangle) = S$ do
            Let $\widehat{o}_\phi$ be the estimator of $|\langle\psi|\phi\rangle|^2$ computed using the algorithm in
              Theorem 6.6
return whichever $|\phi\rangle$ maximizes $\widehat{o}_\phi$
```

suffices to guarantee that with probability at least $1-\delta$*, Algorithm 3 outputs a state* $|\phi\rangle$ *satisfying* $|\langle\phi|\psi\rangle|^2 \geq \tau - \varepsilon$.

Proof. Choose the failure probability in Lemma 6.8 to be at most $\delta/2$. Choose the parameters in Theorem 6.6 so that the additive error in the estimates is $\varepsilon/2$ and the failure probability is at most $\delta/2$; this requires choosing $K = 2^n \cdot O\left(3^{m_{\mathsf{clique}}/3}\right)$ and thus $m_{\mathsf{shadow}} = O(\log(K/\delta)/\varepsilon^2) = O((n + m_{\mathsf{clique}} + \log(1/\delta))/\varepsilon^2) = O(m_{\mathsf{clique}}/\varepsilon^2)$.

We assume henceforth that both Theorem 6.6 and Lemma 6.8 do not fail, which happens with probability at least $1-\delta$ over the samples.

Letting $|\varphi\rangle$ be the state maximizing stabilizer fidelity and $S^* = \mathrm{Weyl}(|\varphi\rangle)$, Lemma 6.8 guarantees that the algorithm samples a complete set of generators for S^*. These generators are necessarily contained in some maximal clique of G because they all commute, and moreover, the subspace spanned by this clique must equal S^* because S^* equals its symplectic complement (so the maximal clique cannot contain any elements not in S^*).

By Theorem 6.6, the estimate $\widehat{o}_\varphi$ is at least $\tau - \varepsilon/2$, so $\max_\phi \widehat{o}_\phi \geq \tau - \varepsilon/2$. Thus, the state $|\phi\rangle$ that maximizes the estimate $\widehat{o}_\phi$ (and is output by the algorithm) has $|\langle\phi|\psi\rangle|^2 \geq \widehat{o}_\phi - \varepsilon/2 \geq \widehat{o}_\varphi - \varepsilon/2 \geq \tau - \varepsilon$. □

Finally, we briefly comment on the runtime of Algorithm 3. The runtime is dominated by iterating through all of the maximal cliques, iterating through all of the stabilizer states $|\phi\rangle$ such that $\mathrm{Weyl}(|\phi\rangle) = S$, and computing $\widehat{o}_\phi$. There are at most $O\left(3^{m_{\mathsf{clique}}/3}\right)$ maximal cliques, by Theorem 6.7. There are exactly 2^n stabilizer states in each basis. Finally, Theorem 6.6 guarantees that computing each $\widehat{o}_\phi$ from the classical shadows takes time $O(n^2 m_{\mathsf{shadow}})$. Thus the overall time complexity is at most

$$O\left(3^{m_{\mathsf{clique}}/3} \cdot 2^n \cdot n^2 m_{\mathsf{shadow}}\right).$$

Plugging in the bounds on m_{clique} and m_{shadow} gives

$$\exp\left(O\left(\frac{n+\log(1/\delta)}{\tau^4}\right)\right)\cdot\frac{n^3+n^2\log(1/\delta)}{\varepsilon^2\tau^4},$$

which further simplifies to

$$\exp\left(O\left(\frac{n+\log(1/\delta)}{\tau^4}\right)\right)\cdot\frac{1}{\varepsilon^2}$$

by absorbing the rightmost term into the big-O in the exponent.

6.3 Discussion and Open Problems

A natural direction for future work is to improve the performance of our algorithms or to prove (conditional or unconditional) lower bounds. In particular, can the exponential running time of Algorithm 3 be improved upon, or is stabilizer fidelity estimation computationally hard? We are optimistic that the exponential factors in our runtime analysis could be made much smaller in practice, because our bound on the sample complexity of finding a complete set of generators is probably far from optimal.

We also remark that, at least superficially, our problem of finding the nearest stabilizer state resembles the closest vector problem (CVP): given a lattice L and a target vector, find the nearest lattice point to the target vector. In our problem, we are given a target vector, and we want to find the nearest stabilizer state to the target vector. While not a lattice, the stabilizer states are "evenly spread" across the complex unit sphere due to their 3-design property [KG15, Web16]. CVP is known to be NP-hard to solve approximately to within any constant and some almost-polynomial factors [vEB81, ABSS97, DKS98]. Is there a formal connection between these two problems?

One can view the output of Algorithm 3 as an approximation of the input state by a nearby stabilizer state. Following this theme, a natural objective is to design similar approximation algorithms relative to other classes of quantum states such as product states or matchgate states. We note that it is even open to design a time-efficient algorithm that, given copies of an n-qubit quantum state, outputs the nearest state from the set $\{|0\rangle, |1\rangle, |+\rangle, |-\rangle, |i\rangle, |-i\rangle\}^{\otimes n}$, which is a subset of stabilizer states. In addition to potentially improving Clifford+T simulation algorithms, are there other applications for these types of state approximation algorithms? As mentioned previously, something like matching pursuit [MZ94] could potentially lead to better magic state decompositions. In addition, are there other applications for these types of state approximation algorithms?

Chapter 7

Efficient Learning of Quantum States Prepared With Few Non-Clifford Gates

This chapter is based on [GIKL23a]. It was joint work with Sabee Grewal, Vishnu Iyer, and William Krestchmer. Some of the preliminary results were moved to Chapters 2 to 4.

7.1 Introduction

Quantum state tomography is the task of constructing a classical description of a quantum state, given copies of the state. This task—whose study dates back to the 1950s [Fan57]—is fundamentally important in quantum theory, and finds applications in the verification of quantum technologies and in experiments throughout physics, among other things. For a thorough history and motivation, we refer the reader to [DPS03, BCG13].

The optimal number of copies to perform quantum state tomography on a d-dimensional quantum *mixed* state is $\Theta(d^2)$ using entangled measurements [OW16, HHJ+17] and $\Theta(d^3)$ using single-copy measurements [KRT17, HHJ+17, CHL+22]. For a quantum *pure* state, $\Theta(d)$ copies are necessary and sufficient [BM99]. Alas, since the dimension d grows exponentially with the system size, the number of copies consumed by state tomography algorithms quickly becomes impractical, and, indeed, learning systems of even 10 qubits can require millions of measurements [SXL+17].

There have been several approaches to circumvent the exponential scaling of quantum state tomography, which we discuss further in Section 7.1.2. For example, one can try to recover less information about the state, or make additional assumptions about the state. While these results have drastically improved copy complexities relative to general quantum state tomography, many of them remain computationally inefficient.

In this chapter, we present a pure state tomography algorithm whose copy and time complexities scale in the complexity of a circuit that prepares the state. More specifically, we assume that the circuit is described by a gate set consisting of Clifford gates (i.e., Hadamard, phase, and CNOT gates) as well as single-qubit non-Clifford gates. We re-iterate that such gate sets are well-studied in quantum information because they are universal for quantum computation [Shi03], and have a number of desirable properties for quantum error correction and fault tolerance [Kni04, BK05, BBB+23], classical simulation of quantum circuits [BBC+19], and efficient implementation of approximate t-designs [HMMH+20].

Our main result is a tomography algorithm that scales *polynomially* in the number of qubits and *exponentially* in the number of non-Clifford gates needed to prepare the state.

Theorem 7.1 (Informal version of Theorem 7.14)**.** *Let $|\psi\rangle$ be an n-qubit quantum state that can be prepared by Clifford gates and t single-qubit non-Clifford gates. There exists an algorithm that uses $\mathsf{poly}(n, 2^t, 1/\varepsilon)$ time and copies of $|\psi\rangle$, and outputs a classical description of $|\psi\rangle$ that is ε-close in trace distance to $|\psi\rangle$ with high probability.*

Hence, our algorithm learns in polynomial time any quantum state that can be prepared by Clifford gates and $O(\log(n))$ single-qubit non-Clifford gates. Although our algorithm is no longer efficient when t exceeds $\omega(\log(n))$, it still remains more efficient than standard pure state tomography as long as t is asymptotically smaller than n.

7.1.1 Main Ideas

Our algorithm in fact learns a more general class of states, namely: quantum states with *stabilizer dimension* at least $n - t$ (Definition 2.37). Informally, a quantum state has stabilizer dimension $n-t$ if it is stabilized by an abelian group of 2^{n-t} Pauli operators. (Recall that an operator U stabilizes a quantum state $|\psi\rangle$ when $U|\psi\rangle = |\psi\rangle$.) Quantum states prepared by Clifford gates and $t/2$ non-Clifford gates fall into this class because they have stabilizer dimension at least $n-t$ (Lemma 7.3).

Our first observation is that learning $|\psi\rangle$ reduces to learning $\mathrm{Weyl}(|\psi\rangle)$ (Definition 2.22). In particular, we show in Lemma 7.6 that given a set of generators for $\mathrm{Weyl}(|\psi\rangle)$, we can efficiently construct a Clifford circuit C such that $C|\psi\rangle = |\varphi\rangle|x\rangle$, where $|x\rangle$ is a computational basis state on $n-t$ qubits and $|\varphi\rangle$ is a general state on t qubits. This construction builds on standard techniques for manipulating stabilizer tableaux, which appeared e.g., in the Aaronson-Gottesman algorithm [AG04]. In some sense, this step "compresses" the non-Cliffordness of the state into the first t qubits.[1] Once we know C, we can easily learn $|x\rangle$ by measuring $C|\psi\rangle$, and can learn $|\varphi\rangle$ using a tomography algorithm on t qubits, which takes $2^{O(t)}$ time [BM99].

To learn $\mathrm{Weyl}(|\psi\rangle)$, we again utilize *Bell difference sampling*. A key property of Bell difference sampling that we showed in Chapter 3 is that the support is always constrained to $\mathrm{Weyl}(|\psi\rangle)^{\perp}$ (Corollary 4.5). This suggests a natural approach to try to compute $\mathrm{Weyl}(|\psi\rangle)$: Bell difference sample repeatedly, and then take our estimate $\widehat{\mathrm{Weyl}(|\psi\rangle)}$ to be the symplectic complement of the sampled Pauli operators.

A priori, it is not at all clear why this strategy could work, because in general $\widehat{\mathrm{Weyl}(|\psi\rangle)}$ may be much larger than $\mathrm{Weyl}(|\psi\rangle)$. A key technical step in our proof amounts to showing that, after $\mathsf{poly}(n, 1/\varepsilon)$ Bell difference samples, with high probability, $|\psi\rangle$ must be ε-close to a state whose unsigned stabilizer group is $\widehat{\mathrm{Weyl}(|\psi\rangle)}$. In other words, if after sufficiently many samples $\widehat{\mathrm{Weyl}(|\psi\rangle)}$ is larger than G, then this witnesses that $|\psi\rangle$ is close to a state with stabilizer dimension $n-\widehat{t}$ for some $\widehat{t} < t$. So, we can use the aforementioned

[1] Similar techniques of compressing non-Cliffordness have appeared in [ABNOGm22, LOLH22].

stabilizer tableau algorithm (Lemma 7.6) on $\widehat{\mathrm{Weyl}(|\psi\rangle)}$ to find a Clifford circuit C such that $C|\psi\rangle \approx |\varphi\rangle|x\rangle$, where $|\varphi\rangle$ has $\widehat{t} < t$ qubits.

As a byproduct of this step in our proof, we obtain an algorithm for *property testing* stabilizer dimension, which may be of independent interest.

Theorem 7.2 (Informal version of Theorem 7.10)**.** *Let $k \geq 1$, let $|\psi\rangle$ be an n-qubit quantum state, and suppose that either (1) $|\psi\rangle$ has stabilizer dimension at least k, or (2) $|\psi\rangle$ has fidelity at most $1-\varepsilon$ with all such states. There is an algorithm that distinguishes these two cases using $O(n/\varepsilon)$ copies of $|\psi\rangle$ and $O(n^3/\varepsilon)$ time.*

Notably, this property testing algorithm is efficient for *all* choices of the stabilizer dimension k, unlike our learning algorithm. Much like Theorem 5.6, Theorem 7.2 can also be used to show that certain states are not computationally pseudorandom (using the definition of [JLS18]). In particular, that Haar-random states are efficiently distinguishable from states with nonzero stabilizer dimension (see Section 8.3 for the details and proof).

To recap, the steps in our learning algorithm are as follows: (1) Bell difference sample repeatedly, (2) compute the symplectic complement $\widehat{\mathrm{Weyl}(|\psi\rangle)}$ of the sampled Pauli operators, (3) use $\widehat{\mathrm{Weyl}(|\psi\rangle)}$ to find a Clifford circuit C such that $C|\psi\rangle \approx |\varphi\rangle|x\rangle$, and (4) learn $|\varphi\rangle|x\rangle$. While some aspects of the analysis are technical, the algorithm itself is quite simple and could be amenable to implementation on near-term devices. Indeed, the only quantum parts of the algorithm involve measuring pairs of qubits, applying Clifford circuits, measuring in the computational basis, and performing tomography on a t-qubit state. So, for example, the resource requirements of our algorithm are quite comparable to those of the classical shadows protocol [HKP20].

7.1.2 Related Work

There is a long line of work devoted to developing near-Clifford *simulation* algorithms [AG04, BG16, RLCK19, BBC+19, QPG21], which classically simulate quantum circuits dominated by Clifford gates. These algorithms scale polynomially in the number of qubits and Clifford gates and exponentially in the number of non-Clifford gates. The main contribution of this work is to complement these classical simulation algorithms with a learning algorithm that scales comparably with respect to the number of non-Clifford gates.

There are a few other classes of quantum states for which time-efficient tomography algorithms are known. Among these are stabilizer states [Mon17], non-interacting fermion sates [AG23], matrix product states [CPF+10], and certain classes of phase states [ABDY22]. As a result of our work, the class of quantum states prepared by Clifford gates and $O(\log(n))$ non-Clifford gates joins this list. We note that our result strictly generalizes Montanaro's algorithm for learning stabilizer states (see Algorithm 1).

Lai and Cheng [LC22] gave an algorithm that learns a quantum state that is prepared via Clifford gates and a few T-gates, where the T-gate is the non-Clifford unitary $T = |0\rangle\langle 0| + e^{i\pi/4}|1\rangle\langle 1|$. However, their algorithm only works if the circuit U that prepares the input state meets two conditions. Firstly, U must be written as $C_1 T^v C_2$, where C_1 and C_2 are Clifford circuits, and $T^v = T^{v_1} \otimes \cdots \otimes T^{v_n}$ for a string $v \in \mathbb{F}_2^n$ of Hamming weight $O(\log(n))$ (i.e., there is a single layer of $O(\log(n))$ T-gates between two Clifford circuits). Secondly, the X-matrix of the stabilizer tableau of $C_1|0^n\rangle$ must be full rank, but only when restricted to the qubits that the T-gates act on. Suffice it to say, their algorithm works in a highly restricted setting. In contrast, our algorithm works for *any* quantum state prepared using Clifford gates and $O(\log(n))$ arbitrary non-Clifford gates (not just the T-gate), and with the non-Clifford gates allowed to be placed anywhere in the circuit. Our work therefore applies to a substantially larger set of states than Lai and Cheng's algorithm.

Besides Lai and Cheng [LC22], other authors have explored the complexity of some related learning problems that involve Clifford+T circuits. For example, [HIN$^+$22] observes that, given samples from the measurement distribution of a circuit comprised of Clifford gates and a single T-gate, learning this distribution can be as hard as the learning parities with noise (LPN) problem. Leone, Oliviero, Lloyd, and Hamma [LOLH22] give algorithms for learning the dynamics of a quantum circuit U comprised of Clifford gates and few T-gates, given oracle access to U. Both of these results are incomparable to our algorithm, because the inputs and outputs of the respective learning tasks are different from ours.

In another direction, one can reduce the computational complexity of learning by only estimating certain properties of quantum states, instead of producing an entire description of the state. For example, consider the *shadow tomography* problem [Aar19, AR19, BO21] where, given a list of known two-outcome observables and copies of an unknown quantum state, the goal is to estimate the expectation value of each observable with respect to the unknown state. Aaronson [Aar19] showed that shadow tomography requires a number of copies that scales polylogarithmically in both the number of two-outcome observables and the Hilbert space dimension, but the algorithm is not computationally efficient. More recently, Huang, Kueng, and Preskill [HKP20] introduced *classical shadows*, a shadow tomography algorithm that could be amenable to near-term quantum devices. Just like prior work, classical shadows uses exponentially fewer copies of the input state relative to state tomography, but, in general, is not computationally efficient.[2]

7.2 Stabilizer Dimension of t-doped Clifford Circuits

To set the stage, we first show that the output state $|\psi\rangle$ of a t-doped Clifford circuit, where $t < n/2$, induces a distribution q_ψ that is supported over a subspace of dimension

[2] There are certain settings where classical shadows is computationally efficient; see [HKP20] for more detail.

at most $2n - 2$. This means that we will be dealing with states with nonzero stabilizer dimension. This also serves to motivate the importance of states with nonzero stabilizer dimension, as they are a superset of the states produced by t-doped Clifford circuits for small t.

Lemma 7.3. *Let $|\psi\rangle$ be the output state of a t-doped Clifford circuit. Then the stabilizer dimension of $|\psi\rangle$ is at least $n - 2t$.*

Proof. We proceed by induction on t. In the base case $t = 0$, so $|\psi\rangle$ is a stabilizer state and has stabilizer dimension n.

For the inductive step, let $t > 0$. Write $|\psi\rangle = CU|\varphi\rangle$, where $|\varphi\rangle$ is the output of a $(t-1)$-doped Clifford circuit, U is a single-qubit gate, and C is a Clifford circuit. Because the stabilizer dimension is unchanged by Clifford gates, it suffices to show that the stabilizer dimension of $U|\varphi\rangle$ is at least $n - 2t$.

Let $S = \mathrm{Weyl}(|\varphi\rangle)$, which by the induction assumption has dimension at least $n - 2(t-1)$. Observe that for any $x \in S$, if the Weyl operator W_x commutes with U, then:

$$\langle\varphi|U^\dagger W_x U|\varphi\rangle = \langle\varphi|W_x|\varphi\rangle = \pm 1.$$

Hence, letting $T \coloneqq \{x \in S : UW_xU^\dagger = W_x\}$, we see that the stabilizer dimension of $U|\varphi\rangle$ is at least the dimension of T. But $|T| \geq |S|/4$, because T contains all elements x of S for which W_x restricts to the identity on the qubit to which U is applied. Thus, the stabilizer dimension of $U|\varphi\rangle$ is at least $n - 2t$, as desired. □

We remark that the stabilizer dimension lower bound in Lemma 7.3 can be improved to $n - t$ in the case that all of the non-Clifford gates are diagonal (for example, if all of the non-Clifford gates are T-gates). This is because diagonal gates commute with both I and Z.

As a consequence of Lemma 7.3, we can use Corollary 4.5 to show that the support of q_ψ must lie in some proper subspace if t is small enough.

Corollary 7.4. *Let $|\psi\rangle$ be the output state of a t-doped Clifford circuit. Then the support of q_ψ is a subspace of dimension at most $n + 2t$.*

Proof. By Lemma 7.3, the dimension of $\mathrm{Weyl}(|\psi\rangle)$ is at least $n - 2t$, implying the dimension of $\mathrm{Weyl}(|\psi\rangle)^\perp$ is at most $n + 2t$. The result follows from Corollary 4.5, which says that the support of q_ψ must lie in $\mathrm{Weyl}(|\psi\rangle)^\perp$. □

7.3 Linear Algebra Subroutines

Our tomography algorithm uses two linear algebraic subroutines, which we describe below. First, we give an algorithm for computing the symplectic complement of a subspace.

Lemma 7.5. *Given a set of m vectors whose span is a subspace $H \subseteq \mathbb{F}_2^{2n}$, there is an algorithm that outputs a basis for $H^\perp$ in $O(mn \cdot \min(m, n))$ time.*

Proof. The algorithm works as follows. First, construct a $m \times 2n$ matrix whose rows are the m elements of H given as input. Then swap the left and right $m \times n$ block submatrices, and denote the resulting matrix by M. Observe that for a nonzero vector v, $Mv = 0$ only when the symplectic product between v and all vectors in H is 0. Hence, v is in $H^\perp$, and the nullspace of M is precisely $H^\perp$. Finding a basis for the nullspace of M can be done via Gaussian elimination, which takes $O(mn \cdot \min(m, n))$ time. □

Next, we explain how to find a Clifford circuit whose action on $\mathbb{F}_2^{2n}$ maps an arbitrary d-dimensional isotropic subspace of $\mathbb{F}_2^{2n}$ to the subspace $0^{2n-d} \times \mathbb{F}_2^d$. We note that while the existence of such a Clifford circuit is not difficult to show (cf. Lemma 4.9), an explicit and efficient construction requires a bit more effort.

Lemma 7.6. *Given a set of m vectors whose span is a d-dimensional isotropic subspace $H \subset \mathbb{F}_2^{2n}$, there exists an efficient algorithm that outputs a Clifford circuit C such that*

$C(H) = 0^{2n-d} \times \mathbb{F}_2^d$. The algorithm runs in $O(mn \cdot \min(m,n))$ time, and the circuit size of C (i.e., the number of gates) is $O(nd)$.

Proof. We will explain the algorithm and then prove its correctness. To begin, run Gaussian elimination on the set of m vectors to get a basis for H such that, when written as a $d \times 2n$ matrix $M = (m_{i,j})$, the matrix M is in row echelon form. This process takes $O(m^2 n)$ time. The subspace spanned by the rows of M is precisely the subspace H.

The matrix M is essentially a *stabilizer tableau*, and therefore Clifford gates have the following effect on M (for additional detail see, e.g., [AG04]):

- Applying the Hadamard gate on the ith qubit corresponds to swapping the ith and $(n+i)$th columns of M.
- Applying the phase gate on the ith qubit corresponds to adding the ith column of M to the $(n+i)$th column of M.
- Applying the CNOT gate with control qubit i and target qubit j corresponds to adding the ith column of M to the jth column M and adding the $(n+j)$th column of M to the $(n+i)$th column of M.

Additionally, row operations do not change the subspace spanned by the rows of M and therefore can be done freely.

Our job now is to find a sequence of Hadamard, phase, and CNOT gates that maps M to a matrix whose rows span the subspace $0^{2n-d} \times \mathbb{F}_2^d$; in particular, a matrix with the following form

$$\left(0 \,\middle|\, 0 \quad I\right), \tag{7.1}$$

where the first 0 is a $d \times n$ matrix of all 0's, the second is a $d \times (n-d)$ matrix of all 0's, and the last is a $d \times d$ identity matrix.

The remainder of the algorithm works as follows.

1. For each row $i \in [d]$ of M:
   (a) For each $j \in \{i, \ldots, n\}$, apply phase and Hadamard gates so that either $m_{i,j} = 1$ and $m_{i,n+j} = 0$ or both are 0.
   (b) If $m_{i,i} = 0$, then find a $k \in \{i+1, \ldots, n\}$ for which $m_{i,k} = 1$.[3] Apply a CNOT with control qubit i and target qubit k so that $m_{i,i} = 1$.
   (c) For each $j \in \{i+1, \ldots, n\}$, if $m_{i,j} = 1$, apply CNOT with control qubit i and target qubit j.
   (d) For $j \in \{i+1, \ldots, d\}$, set $m_{j,i} = 0$.[4]
2. Apply a Hadamard gate to each of the first d qubits.
3. For $i \in \{0, \ldots, d-1\}$, apply a CNOT with control qubit $n-i$ and target qubit $d-i$. Then apply a CNOT with control qubit $d-i$ and target qubit $n-i$.

Let C denote the Clifford circuit described by the above process. The algorithm concludes by outputting C. We apply $O(n)$ Clifford gates and do at most $O(d)$ row-sum operations per row. Thus, looping over the matrix M takes $O(nd)$ time, and therefore the overall running time is $O(mn \cdot \min(m,n))$, due to the Gaussian elimination step at the beginning of the algorithm. The size of the circuit is at most $O(nd)$.

To prove correctness, we must argue that $C(H) = 0^{2n-d} \times \mathbb{F}_2^d$, or equivalently, that the algorithm above maps M to a matrix as in Eq. (7.1).

[3] At least one such k must exist, for if it didn't, then the ith row would be all 0's, which is impossible since the rows of M are linearly independent.

[4] This corresponds to adding the ith row to the jth row, which, as mentioned earlier, does not change the subspace spanned by the rows of M.

First, we show that Step 1 of the algorithm maps M to a matrix of the form

$$\begin{pmatrix} I & 0 \mid 0 \end{pmatrix},$$

where I is the $d \times d$ identity matrix. It is clear that after the first iteration of Step 1 completes, $m_{1,1} = 1$ and the remaining entries of the first row and column are 0's. By way of induction, assume that this is true after the first $i - 1$ iterations, so that the resulting matrix looks as follows:

$$\left(\begin{array}{c|c|c|c}
\multicolumn{2}{c|}{\overbrace{\begin{matrix}1 & & \\ & \ddots & \\ & & 1\end{matrix} \;\Bigg|\; \begin{matrix}0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0\end{matrix}}^{n \text{ columns}}} & \multicolumn{2}{c}{\begin{matrix}0 & \cdots & \cdots & \cdots & 0 \\ \vdots & & & & \vdots \\ 0 & \cdots & \cdots & \cdots & 0\end{matrix}} \\ \hline
\underbrace{\begin{matrix}0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0\end{matrix}}_{i-1 \text{ columns}} & m_{i,i} \qquad\qquad & \underbrace{\begin{matrix}0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0\end{matrix}}_{i-1 \text{ columns}} & \qquad\qquad
\end{array}\right). \tag{7.2}$$

In the top row, from left to right, the first block is the $(i-1) \times (i-1)$ identity matrix, then an $(i-1) \times (n-i+1)$ block of all 0's, and finally an $(i-1) \times 2n$ block of all 0's. In the bottom row, from left to right, the first block is a $(d-i+1) \times (i-1)$ matrix of all 0's, then a $(d-i+1) \times (n-i+1)$ block being processed by the algorithm, then a $(d-i+1) \times (i-1)$ block of all 0's, and finally a $(d-i+1) \times (n-i+1)$ block being processed by the algorithm.

We will argue that after the ith iteration of Step 1 finishes, the matrix will have the form of Eq. (7.2) but with $m_{i,i} = 1$ and the rest of the ith row and ith column cleared to 0. First, observe that the third block in the second row must be all 0's if the first block of the top row is the identity matrix because the subspace spanned by the rows is isotropic (and applying Clifford gates will not affect that by their preservation of commutation relations among Paulis). It is also clear that the operations performed in the ith iteration will set $m_{i,i} = 1$, set $m_{i,j} = 0$ for $j \in \{i+1, \ldots, 2n\}$, and set $m_{j,i} = 0$ for $j \in \{i+1, \ldots, d\}$. Therefore, we just need to argue that the ith iteration does not reintroduce 1's into the blocks of 0's or affect the $(i-1) \times (i-1)$ identity matrix in the first block of the first row. Observe that neither of these can happen as long as Hadamard gates and CNOT gates are not applied to the first $i-1$ qubits in the ith iteration. Indeed, our algorithm does not apply any gates to the first $i-1$ qubits, so the structure of matrix in Eq. (7.2) is preserved. Therefore, once Step 1 terminates, the resulting matrix will by a $d \times d$ identity matrix in the first block and the remaining entries of the matrix will be 0's.

The layer of Hadamard gates in Step 2 maps the $d \times d$ identity matrix to the right block of the matrix, i.e.,

$$\left(0 \mid I \quad 0\right).$$

Finally, the CNOT gates in Step 3 move the identity matrix to the rightmost-side of the tableau, matching the goal shown in Eq. (7.1). This can be verified by explicit calculation. We note that the CNOT gates move the identity matrix by starting with the rightmost column and then proceeding leftward, which is critical for correctness when $d > n/2$. Hence, C performs the desired mapping. □

7.4 On Subspaces with Large q_ψ-mass

We prove two lemmas that form the starting point for our property testing and tomography algorithms. First, we show that, by Bell difference sampling a sufficient number of times, one can efficiently learn a coisotropic subspace $H^\perp$ that accounts for a large fraction of the q_ψ-mass (and therefore, the p_ψ-mass, by Corollary 4.3). The coisotropic property will naturally follow from Lemmas 4.7 and 4.8 if the fraction of the probability mass is large enough. We will use the symplectic complement H of $H^\perp$ as a sort of proxy for $\mathrm{Weyl}(|\psi\rangle)$ (Definition 2.22), i.e., the set of Weyl operators that stabilizer $|\psi\rangle$ up to sign.

Second, we show that, if an $(n+t)$-dimensional (coisotropic) subspace $H^\perp$ *does* account for a large fraction of the p_ψ-mass, then the state $|\psi\rangle$ is close in fidelity to a state of the form $C\,|\varphi\rangle\,|x\rangle$, where C is a Clifford circuit determined by H, $|x\rangle$ is an $(n-t)$-qubit basis state, and $|\varphi\rangle$ is a t-qubit state. Furthermore, we explain how to use Lemma 7.6 to efficiently construct the Clifford circuit C given H.

7.4.1 Sampling Lemma

We begin by showing that Bell difference sampling allows one to approximate the support of q_ψ. The proof is similar in nature to that of Lemma 3.10.

Lemma 7.7. *Let $|\psi\rangle$ be an n-qubit quantum pure state, and let $H^\perp$ denote the subspace of $\mathbb{F}_2^{2n}$ spanned by m samples drawn from q_ψ. Then, for $\varepsilon, \delta \in (0,1)$,*

$$\sum_{x \in H^\perp} q_\psi(x) \geq 1 - \varepsilon$$

with probability at least $1-\delta$ as long as

$$m \geq \frac{2\log\frac{1}{\delta} + 4n}{\varepsilon}.$$

Proof. For samples $x_1, \ldots, x_m \in \mathbb{F}_2^{2n}$ drawn from q_ψ, let $H_i^\perp = \langle x_1, \ldots, x_i \rangle$ be the subspace spanned by the first i samples for arbitrary $0 \leq i \leq m$. Define the indicator random variable X_i as

$$X_i = \begin{cases} 1 & \text{if } x_i \in \mathbb{F}_2^{2n} \setminus H_{i-1}^\perp \text{ or } \sum_{x \in H_{i-1}^\perp} q_\psi(x) \geq 1 - \varepsilon. \\ 0 & \text{otherwise.} \end{cases}$$

Intuitively, $X_i = 1$ indicates that either the sample x_i has increased the span of $H_i^\perp$ (i.e., $H_{i-1}^\perp \subset H_i^\perp$) or that $H_{i-1}^\perp$ already accounts for a $(1-\varepsilon)$-fraction of the mass of q_ψ. If $H_{i-1}^\perp$ does account for a $(1-\varepsilon)$-fraction of the mass of q_ψ, then $X_i = 1$ with probability

1. If not, then the probability mass on $\mathbb{F}_2^{2n} \setminus H_{i-1}^{\perp}$ is at least ε, and therefore, $X_i = 1$ with probability at least ε. In both cases, $\mathbf{Pr}[X_i = 1] \geq \varepsilon$, and therefore $\mu := \mathsf{E}[\sum_i X_i] \geq m\varepsilon$.[5]

Once $\sum_{i=1}^{m} X_i \geq 2n$, the subspace $H_m^{\perp}$ must account for a $(1-\varepsilon)$-fraction of the mass of q_ψ. This is because the dimension of $\mathbb{F}_2^{2n}$ is $2n$ and so the span can only expand $2n$ times. Set $\gamma := 1 - \frac{2n}{\mu}$. By a Chernoff bound (Fact 2.2),

$$\begin{aligned}\mathbf{Pr}\left[\sum_{i=1}^{m} X_i \leq 2n\right] &= \mathbf{Pr}\left[\sum_{i=1}^{m} X_i \leq (1-\gamma)\mu\right] \\ &\leq \exp\left(-\frac{\mu}{2}\gamma^2\right) \\ &= \exp\left(-\frac{\mu}{2} - \frac{2n^2}{\mu} + 2n\right) \\ &\leq \exp\left(-\frac{m\varepsilon}{2} + 2n\right).\end{aligned}$$

Hence, as long as

$$m \geq \frac{2\log\frac{1}{\delta} + 4n}{\varepsilon},$$

$H_m^{\perp}$ will account for a $(1-\varepsilon)$-fraction of the q_ψ-mass with probability at least $1-\delta$. □

7.4.2 Product State Structure

We prove a critical relation between the p_ψ-mass on isotropic subspaces and stabilizer dimension by generalizing Proposition 4.16. In particular, we show that if an $(n-t)$-dimensional isotropic subspace has large p_ψ-mass, then there is a Clifford circuit C (that can be constructed efficiently) that maps $|\psi\rangle$ to a product state $|\varphi\rangle|x\rangle$, where $|x\rangle$ is an $(n-t)$-qubit basis state and $|\varphi\rangle$ is an arbitrary t-qubit state.

Just like Proposition 4.16, it will be easier to assume that the isotropic subspace is a subset of $\{I, Z\}^{\otimes n}$. We now state our main lemma in that context.

[5] Actually, this bound on $\mathbf{Pr}[X_i = 1]$ is loose until $\dim H_i^{\perp}$ exceeds n, because subspaces of dimension less than n cannot account for more than half of the q_ψ mass. Accounting for this should yield a constant factor improvement in the sample complexity.

Lemma 7.8. *Let $T = 0^{n+t} \times \mathbb{F}_2^{n-t}$, and suppose that*

$$\sum_{x \in T} p_\psi(x) \geq \frac{1-\varepsilon}{2^t}.$$

Then there exists an $(n-t)$-qubit computational basis state $|x\rangle$ and a t-qubit quantum state

$$|\varphi\rangle := \frac{(I \otimes \langle x|)\, |\psi\rangle}{\|(I \otimes \langle x|)\, |\psi\rangle\|_2},^{6}$$

such that the fidelity between $|\varphi\rangle\,|x\rangle$ and $|\psi\rangle$ is at least $1-\varepsilon$.

Proof. We can always write $|\psi\rangle = \sum_{x \in \mathbb{F}_2^{n-t}} \alpha_x \,|\varphi_x\rangle\,|x\rangle$ where $\sum_{x \in \mathbb{F}_2^{n-t}} |\alpha_x|^2 = 1$. If we can show that

$$\max_{x \in \mathbb{F}_2^{n-t}} |\langle \psi|\, (|\varphi_x\rangle\,|x\rangle)|^2 \geq 1-\varepsilon,$$

then we are done, by taking $|\varphi\rangle = |\varphi_x\rangle$. First,

$$\begin{aligned} \max_{x \in \mathbb{F}_2^{n-t}} |\langle \psi|\, (|\varphi_x\rangle\,|x\rangle)|^2 &= \max_{x \in \mathbb{F}_2^{n-t}} |\alpha_x|^2 \\ &= \max_{x \in \mathbb{F}_2^{n-t}} |\alpha_x|^2 \cdot \sum_{x \in \mathbb{F}_2^{n-t}} |\alpha_x|^2 \\ &\geq \sum_{x \in \mathbb{F}_2^{n-t}} |\alpha_x|^4. \end{aligned}$$

Observe that $\sum_x |\alpha_x|^4$ is precisely the collision probability when measuring the last $n-t$ qubits of $|\psi\rangle$ in the computational basis. Hence, by Lemma 4.17,

$$\sum_{x \in \mathbb{F}_2^{n-t}} |\alpha_x|^4 = 2^t \sum_{x \in T} p_\psi(x) \geq 1-\varepsilon. \qquad \square$$

Finally, we generalize the previous lemma using the Clifford mapping algorithm from Lemma 7.6.

[6] This is to say that $|\varphi\rangle$ is obtained by postselecting on measuring the last $n-t$ qubits of $|\psi\rangle$ to be $|x\rangle$.

Corollary 7.9. *Let T be an isotropic subspace of dimension $n-t$, and suppose that*

$$\sum_{x \in T} p_\psi(x) \geq \frac{1-\varepsilon}{2^t}.$$

Then there exists a state $|\widehat{\psi}\rangle$ with $T \subseteq \mathrm{Weyl}(|\widehat{\psi}\rangle)$ such that the fidelity between $|\widehat{\psi}\rangle$ and $|\psi\rangle$ is at least $1-\varepsilon$.

In particular, $|\widehat{\psi}\rangle = C^\dagger |\varphi\rangle |x\rangle$, where $|x\rangle$ is an $(n-t)$-qubit basis state,

$$|\varphi\rangle \coloneqq \frac{(I \otimes \langle x|)C|\psi\rangle}{\|(I \otimes \langle x|)C|\psi\rangle\|_2}$$

is a t-qubit quantum state, and C is a Clifford circuit that can be constructed efficiently.

Proof. Let C be the Clifford circuit mapping T to $0^{n+t} \times \mathbb{F}_2^{n-t}$ described in Lemma 7.6. Define $|\phi\rangle \coloneqq C|\psi\rangle$. Then, by Lemma 4.10,

$$\sum_{x \in T} p_\psi(x) = \sum_{x \in C(T)} p_\phi(x) \geq \frac{1-\varepsilon}{2^t}.$$

Therefore, by Lemma 7.8, $C|\psi\rangle$ is $(1-\varepsilon)$-close in fidelity to a state $|\varphi\rangle|x\rangle$, where $|x\rangle$ is an $(n-t)$-qubit basis state and

$$|\varphi\rangle \coloneqq \frac{(I \otimes \langle x|)C|\psi\rangle}{\|(I \otimes \langle x|)C|\psi\rangle\|_2}$$

is a t-qubit quantum pure state. Since fidelity is unitarily invariant, the fidelity between $|\psi\rangle$ and $|\widehat{\psi}\rangle = C^\dagger|\varphi\rangle|x\rangle$ is also at least $1-\varepsilon$. Clearly, $C(T) \subseteq \mathrm{Weyl}(|\varphi\rangle|x\rangle)$, and therefore, by Lemma 4.10, $T \subseteq \mathrm{Weyl}(|\widehat{\psi}\rangle)$. □

7.5 Property Testing Stabilizer Dimension

As a first application, we present an efficient algorithm for property testing stabilizer dimension. Recall that a property tester for a class $\mathcal{Q}$ of quantum states takes copies of a state $|\psi\rangle$ as input and determines whether $|\psi\rangle \in \mathcal{Q}$ or $|\psi\rangle$ is ε-far from all such states (according to some measure of distance), promised that one of these is the case. Our algorithm efficiently

```
Algorithm 4: Property Testing Stabilizer Dimension
    Input: $\frac{8\log(1/\delta)+16n}{\varepsilon}$ copies of $|\psi\rangle$, $k \in [n]$, $\varepsilon \in (0, 3/8)$, and $\delta \in (0, 1]$
    Promise: $|\psi\rangle$ has stabilizer dimension at least $k$ or is $\varepsilon$-far in fidelity from all
             such states
    Output: 1 if $|\psi\rangle$ has stabilizer dimension at least $k$, 0 otherwise, with
            probability at least $1 - \delta$
    Perform Bell difference sampling to draw $\frac{2\log(1/\delta)+4n}{\varepsilon}$ samples from $q_\psi$.
    Denote the span of the samples by $H^\perp$. Let $\widehat{k} = \dim H = 2n - \dim H^\perp$.
    Return 1 if $\widehat{k} \geq k$ and 0 otherwise.
```

tests whether an input state has stabilizer dimension at least k or has fidelity less than $1-\varepsilon$ with all such states.

Theorem 7.10. *Let $|\psi\rangle$ be an n-qubit quantum state. Algorithm 4 determines whether $|\psi\rangle$ has stabilizer dimension at least k or has fidelity at most $1-\varepsilon$ with all such states, promised that one of these is the case. The algorithm uses $\frac{16n+8\log(1/\delta)}{\varepsilon}$ copies of $|\psi\rangle$ and $O\left(\frac{n^3+n^2\log(1/\delta)}{\varepsilon}\right)$ time, and succeeds with probability at least $1-\delta$.*

Proof. First, suppose that $|\psi\rangle$ has stabilizer dimension at least k. By Lemma 4.4, q_ψ is supported on a subspace of dimension at most $2n-k$. So, no matter what, $\dim H^\perp$ will never exceed $2n-k$, and therefore $\widehat{k}$ will always be at least k. Hence, Algorithm 4 accepts with probability 1.

Now suppose that $|\psi\rangle$ has fidelity less than $1-\varepsilon$ with every state of stabilizer dimension at least k. By Corollary 4.3 and Lemma 7.7, with probability at least $1-\delta$, the subspace H satisfies

$$\sum_{x \in H^\perp} p_\psi(x) \geq \sum_{x \in H^\perp} q_\psi(x) \geq 1 - \varepsilon.$$

Assuming this occurs, applying Theorem 4.1 gives

$$\sum_{x \in H} p_\psi(x) = \frac{|H|}{2^n} \sum_{x \in H^\perp} p_\psi(x) \geq \frac{1-\varepsilon}{2^{n-\widehat{k}}}.$$

Since $\sum_{x \in H^\perp} q_\psi(x) \geq 1 - \varepsilon > 5/8$, H is isotropic by Lemma 4.8. As such, by Corollary 7.9, there exists a state of stabilizer dimension at least $\widehat{k}$ that has fidelity at least $1-\varepsilon$ with $|\psi\rangle$. By assumption, we must have $\widehat{k} < k$, and thus Algorithm 4 rejects with probability at least $1-\delta$.

Overall, we find that in either case, with probability at least $1-\delta$, the algorithm succeeds. It remains to bound the runtime. Computing $\dim H^\perp$ first requires computing a basis for $H^\perp$, and, to do so, we have to run Gaussian elimination. This takes $O(mn^2)$ time, where $m = O\left(\frac{n+\log(1/\delta)}{\varepsilon}\right)$ is the number of Bell difference samples taken by the algorithm, and dominates the running time. □

7.6 Tomography of High-Stabilizer-Dimension States

We present our tomography algorithm that learns a classical description of a quantum state, promised that the state has stabilizer dimension at least $n-t$. As corollaries, we explain how our algorithm learns the outputs of t-doped Clifford circuits and that our algorithm is efficient (both in copy and time complexity) when $t = O(\log(n))$.

Our tomography algorithm builds on the property testing algorithm presented in Section 7.5. That is, we begin by taking enough Bell difference samples to find a coisotropic subspace that accounts for a large fraction of the q_ψ-mass. After that, we use Corollary 7.9 to show that there is a Clifford circuit C such that $C|\psi\rangle$ is close to a state of the form $|\varphi\rangle|x\rangle$ (and we use Lemma 7.6 to construct the Clifford circuit efficiently). From there, one can easily learn the basis state $|x\rangle$. Finally, we run a pure state tomography algorithm to learn the arbitrary state $|\varphi\rangle$.

Definition 7.11 (Pure state tomography copy and time complexities)**.** *Let $N_{n,\varepsilon,\delta}$ and $M_{n,\varepsilon,\delta}$ denote the copy and time complexities, respectively, of a pure state tomography algorithm to learn a classical description of an n-qubit state to trace distance at most ε with probability at least $1-\delta$.*

We will state our copy and time complexities in terms of $N_{n,\varepsilon,\delta}$ and $M_{n,\varepsilon,\delta}$, but, for concreteness, we mention two pure state tomography algorithms. The first is the state-of-the-art when it comes to scaling in the dimension $d = 2^n$ of the system (though, it does not match the best-known lower bound of $\Omega(d)$). The second scales poorly in the dimension, but achieves the optimal dependence on ε and δ.

Theorem 7.12 ([FBaK21]). *Given access to copies of an n-qubit pure state $|\psi\rangle$, there is an algorithm that uses $O\left(2^n n \log(1/\delta)\varepsilon^{-4}\right)$ copies, $O\left(4^n n^3 \log(1/\delta)\varepsilon^{-5}\right)$ time, and outputs a state $|\widehat{\psi}\rangle$ that is ε-close to $|\psi\rangle$ in trace distance with probability at least $1-\delta$. The algorithm only requires applying random Clifford circuits and classical post-processing.*

Theorem 7.13 ([AG23, Section 5]). *Given access to copies of an n-qubit pure state $|\psi\rangle$, there is an algorithm that uses $O\left(16^n \log(1/\delta)\varepsilon^{-2}\right)$ copies, $O\left(32^n \log(1/\delta)\varepsilon^{-2}\right)$ time, and outputs a state $|\widehat{\psi}\rangle$ that is ε-close to $|\psi\rangle$ in trace distance with probability at least $1-\delta$. The algorithm only requires applying Clifford circuits and classical post-processing.*

We are now ready to state our algorithm.

```
Algorithm 5: Tomography of High-Stabilizer-Dimension States
    Input: $O\left(\frac{n+\log(1/\delta)}{\varepsilon^2} + N_{t,\frac{\varepsilon}{2},\frac{\delta}{3}}\right)$ copies of $|\psi\rangle$ and $\varepsilon, \delta \in (0,1]$
    Promise: $|\psi\rangle$ has stabilizer dimension at least $n - t$
    Output: $|\widehat{\psi}\rangle$ such that $\mathrm{TD}(|\psi\rangle, |\widehat{\psi}\rangle) \leq \varepsilon$ with probability at least $1-\delta$
  1 Perform Bell difference sampling to draw $\frac{8\log(3/\delta)+16n}{\varepsilon^2}$ samples from $q_\psi$.
  2 Use the algorithm in Lemma 7.5 to compute the symplectic complement of the
      subspace spanned by the samples. Denote this subspace by $H$ and set
      $\widehat{t} = n - \dim H$.
  3 Use the algorithm in Lemma 7.6 to construct a Clifford circuit $C$ that maps $H$
      to $0^{n+\widehat{t}} \times \mathbb{F}_2^{n-\widehat{t}}$.
  4 Measure the last $n - \widehat{t}$ qubits of $2N_{\widehat{t},\frac{\varepsilon}{2},\frac{\delta}{3}} + 24\log(3/\delta)$ copies of $C|\psi\rangle$. Let
      $\widehat{x} \in \mathbb{F}_2^{n-\widehat{t}}$ be the majority result. Reserve the first $\widehat{t}$ qubits of the states for
      which $\widehat{x}$ is observed, and, otherwise, discard the qubits.
  5 Run a pure state tomography algorithm on the $\widehat{t}$-qubit states reserved in the
      prior step to learn a classical description of the state to trace distance $\varepsilon/2$ with
      probability at least $1-\delta/3$. Let $|\widehat{\varphi}\rangle$ be the output.
  6 Return $|\widehat{\psi}\rangle = C^\dagger |\widehat{\varphi}\rangle |\widehat{x}\rangle$.
```

Theorem 7.14. *Let $|\psi\rangle$ be an n-qubit quantum state with stabilizer dimension at least $n-t$. Given copies of $|\psi\rangle$ as input, Algorithm 5 outputs a classical description of a state $|\widehat{\psi}\rangle$ such that $\mathrm{TD}(|\psi\rangle, |\widehat{\psi}\rangle) \leq \varepsilon$ with probability at least $1-\delta$. The algorithm uses*

$$O\left(\frac{n+\log(1/\delta)}{\varepsilon^2} + N_{t,\frac{\varepsilon}{2},\frac{\delta}{3}}\right)$$

copies of $|\psi\rangle$ and

$$O\left(\frac{n^3+n^2\log(1/\delta)}{\varepsilon^2} + M_{t,\frac{\varepsilon}{2},\frac{\delta}{3}}\right)$$

time.

Proof. By Lemma 7.7, except with probability at most $\delta/3$, the Bell difference sampling phase of the algorithm in Line 1 finds a subspace $H^\perp \subseteq \mathbb{F}_2^{2n}$ such that

$$\sum_{x\in H^\perp} q_\psi(x) \geq 1 - \frac{\varepsilon^2}{4} \geq \frac{3}{4}.$$

Assuming this occurs, by Lemma 4.8, $H^\perp$ is coisotropic. From Corollary 4.3, we know that

$$\sum_{x\in H^\perp} p_\psi(x) \geq 1 - \frac{\varepsilon^2}{4}.$$

In the next step, the algorithm computes H, and we denote $\widehat{t} = n - \dim H$. Using Theorem 4.1, we have that

$$\sum_{x\in H} p_\psi(x) = \frac{|H|}{2^n}\sum_{x\in H^\perp} p_\psi(x) \geq \frac{1-\varepsilon^2/4}{2^{\widehat{t}}}.$$

The algorithm then produces a Clifford circuit C according to Lemma 7.6 such that C maps H to $0^{n+\widehat{t}} \times \mathbb{F}_2^{n-\widehat{t}}$. By Corollary 7.9, there exists an $(n-\widehat{t})$-qubit basis state $|x\rangle$ and a $\widehat{t}$-qubit state

$$|\varphi\rangle := \frac{(I\otimes\langle x|)C\,|\psi\rangle}{\|(I\otimes\langle x|)C\,|\psi\rangle\|_2},$$

such that the fidelity between $C\,|\psi\rangle$ and $|\varphi\rangle\,|x\rangle$ is at least $1-\varepsilon^2/4$, implying that

$$\mathrm{TD}(|\psi\rangle, C^\dagger\,|\varphi\rangle\,|x\rangle) \leq \frac{\varepsilon}{2}. \tag{7.3}$$

Note that the definition of $|\varphi\rangle$ means that $|\varphi\rangle$ is the state on $\widehat{t}$ qubits conditioned on measuring $|x\rangle$ on the last $n-\widehat{t}$ qubits of $C\,|\psi\rangle$. We use this property in Line 4 of Algorithm 5, whose purpose is two-fold. We determine the $(n-\widehat{t})$-qubit basis state $|x\rangle$ by taking a majority vote over measurement outcomes, and also reserve copies of $|\varphi\rangle$ to pass to the pure state tomography algorithm in the proceeding step. Suppose we have m copies of the state $C\,|\psi\rangle$, and let $x_1, \ldots, x_m$ denote the $(n-\widehat{t})$-bit measurement outcome upon measuring the last $n-\widehat{t}$ qubits in the computational basis. Define the indicator random variable X_i as

$$X_i = \begin{cases} 1 & \text{if } x_i = x. \\ 0 & \text{otherwise.} \end{cases}$$

Because the fidelity between $C\,|\psi\rangle$ and $|\varphi\rangle\,|x\rangle$ is at least $1 - \varepsilon^2/4$ and $\varepsilon \in (0,1]$, $\mathbf{Pr}[X_i = 1] \geq 0.75$ and $\mu \coloneqq \mathsf{E}[\sum_i X_i] \geq 0.75m$. By a Chernoff bound (Fact 2.2),

$$\begin{aligned} \mathbf{Pr}\left[\sum_{i=1}^{m} X_i \leq \frac{m}{2}\right] &\leq \mathbf{Pr}\left[\sum_{i=1}^{m} X_i \leq \left(1 - \frac{1}{3}\right)\mu\right] \\ &\leq \exp\left(-\frac{\mu}{18}\right) \\ &\leq \exp\left(-\frac{m}{24}\right). \end{aligned}$$

Hence, as long as,

$$m \geq 24\log(1/\delta),$$

over half of the m samples will be x with probability at least $1-\delta$. In our case, taking m to be $2N_{\widehat{t},\frac{\varepsilon}{2},\frac{\delta}{3}} + 24\log(3/\delta)$ suffices for the majority result $\widehat{x}$ to equal x, and for $N_{\widehat{t},\frac{\varepsilon}{2},\frac{\delta}{3}}$ copies of the $\widehat{t}$-qubit state $|\varphi\rangle$ to be reserved for the pure state tomography algorithm in the proceeding step of the algorithm, except with probability at most $\delta/3$. Assuming this occurs, with probability at least $1-\delta/3$ the tomography algorithm in Line 5 returns a state $|\widehat{\varphi}\rangle$ that is within trace distance at most $\varepsilon/2$ from $|\varphi\rangle$.

In the final step of the algorithm, we return $C^\dagger |\widehat{\varphi}\rangle |\widehat{x}\rangle$. Then we have

$$\begin{aligned} d_{\mathrm{Tr}}(|\psi\rangle, C^\dagger |\widehat{\varphi}\rangle |\widehat{x}\rangle) &\leq d_{\mathrm{Tr}}(|\psi\rangle, C^\dagger |\varphi\rangle |x\rangle) + d_{\mathrm{Tr}}(|\varphi\rangle |x\rangle, |\widehat{\varphi}\rangle |\widehat{x}\rangle) \\ &\leq \frac{\varepsilon}{2} + d_{\mathrm{Tr}}(|\varphi\rangle |x\rangle, |\widehat{\varphi}\rangle |\widehat{x}\rangle) \\ &\leq \varepsilon. \end{aligned}$$

The first step follows by the triangle inequality and the fact that trace distance is unitarily invariant. The second step follows from Eq. (7.3). The final step follows assuming $\widehat{x} = x$ and $d_{\mathrm{Tr}}(|\varphi\rangle, |\widehat{\varphi}\rangle) \leq \varepsilon/2$.

By applying a union bound over the "bad" events (namely, Lines 1, 4 and 5 of the algorithm), we have that the overall success probability of the algorithm is at least $1 - \delta$.

To conclude, we bound the copy and time complexities of the algorithm. By Lemma 4.4, the support of q_ψ is a subspace of dimension at most $n + t$, so $\dim H^\perp \leq n + t$ and therefore $\widehat{t} \leq t$. Thus, the tomography subroutine in Line 5 uses $O\left(N_{t,\frac{\varepsilon}{2},\frac{\delta}{3}} + \log(1/\delta)\right)$ copies, and Line 1 uses $O\left((\log(1/\delta) + n)/\varepsilon^2\right)$ copies. Hence, in total, the algorithm uses

$$O\left(\frac{n + \log(1/\delta)}{\varepsilon^2} + N_{t,\frac{\varepsilon}{2},\frac{\delta}{3}}\right)$$

copies of the input state. As for time complexity, Lines 2 and 5 dominate the running time. The overall time complexity is therefore

$$O\left(\frac{n^3 + n^2\log(1/\delta)}{\varepsilon^2} + M_{t,\frac{\varepsilon}{2},\frac{\delta}{3}}\right). \qquad \square$$

The dependence on t comes entirely from a pure state tomography algorithm (e.g., Theorem 7.12 or Theorem 7.13) in Line 5 of Algorithm 5. Hence, one can upgrade this part of the algorithm with improved pure state tomography algorithms when/if they are discovered. Additionally, one can interpret our result as a reduction to pure state tomography on t-qubits, where the reduction uses $O((n + \log(1/\delta))/\varepsilon^2)$ copies and $O((n^3 + n^2\log(1/\delta))/\varepsilon^2)$ time.

We remark that Lines 1 and 2 are generalizations of [Mon17] (see Theorem 3.13). Notably, since $\mathrm{Weyl}(|\phi\rangle)$ is Lagrangian in the case that $|\phi\rangle$ is a stabilizer state, the symplectic complement did not need to be computed in that $\mathrm{Weyl}(|\phi\rangle)^{\perp} = \mathrm{Weyl}(|p\rangle\, hi)$.

By Lemma 7.3, t-doped Clifford circuits have stabilizer dimension at least $n - 2t$, which yields the following corollary.

Corollary 7.15. *Let $|\psi\rangle$ be an n-qubit quantum state prepared by a t-doped Clifford circuit. Then $O\left(\frac{n+\log(1/\delta)}{\varepsilon^2} + N_{2t,\frac{\varepsilon}{3},\frac{\delta}{3}}\right)$ copies of $|\psi\rangle$ and $O\left(\frac{n^3+n^2\log(1/\delta)}{\varepsilon^2} + M_{2t,\frac{\varepsilon}{2},\frac{\delta}{3}}\right)$ time suffice for Algorithm 5 to output a classical description $|\widehat{\psi}\rangle$ that is within trace distance ε of $|\psi\rangle$ with probability at least $1-\delta$.*

Finally, by taking $t = O(\log(n))$, our algorithm is efficient in both copy and time complexity.

Corollary 7.16. *Let $|\psi\rangle$ be an n-qubit quantum state prepared by a $O(\log(n))$-doped Clifford circuit. Then $\mathsf{poly}(n, \log(1/\delta), 1/\varepsilon)$ copies of $|\psi\rangle$ and time suffice for Algorithm 5 to output a classical description $|\widehat{\psi}\rangle$ that is within trace distance ε of $|\psi\rangle$ with probability at least $1-\delta$.*

7.7 Open Problems

Our tomography algorithm efficiently learns a classical description of a quantum state $|\psi\rangle$ that can be prepared with Clifford gates and $O(\log(n))$ non-Clifford gates. However, our algorithm is not a *proper* learner—in other words, the circuit that the algorithm outputs to approximate the state does not necessarily decompose into few non-Clifford gates.

Question 7.17. *Given an n-qubit state $|\psi\rangle$ that is the output of an $O(\log n)$-doped Clifford circuit, is it possible to construct an $O(\log n)$-doped Clifford circuit C such that $C\,|0^n\rangle$ is ε-close to $|\psi\rangle$?*

We note that an efficient algorithm for stabilizer states is known (namely, just run Algorithm 5 with $t = 0$). We also note that this task becomes trivial if polynomially many non-Clifford gates are allowed in C because one can use Algorithm 5 to produce C, $|x\rangle$, and $|\varphi\rangle$, and then construct a circuit with at most $2^{O(\log n)} = \mathsf{poly}(n)$ general gates that outputs $|\varphi\rangle$ [SBM06].

As a subroutine, Algorithm 5 uses pure state tomography to recover a classical description of a pure state, and therefore the copy and time complexities of our algorithm can be improved if faster pure state tomography algorithms are developed. Developing a pure state tomography algorithm that achieves the optimal $O(2^n/\varepsilon^2)$ copy and time complexities is an interesting and important direction for future work.

Finally, is there an efficient algorithm for learning states with large stabilizer dimension that uses only *single-copy* measurements? Bell difference sampling, which involves two consecutive two-copy measurements of the unknown state, is the only part of our algorithm that uses multiple copies of the state at a time. We remark that Aaronson and Gottesman [AG08] (see also [Aar22, Section 23.3]) gave an algorithm for learning stabilizer states from single-copy measurements.

Chapter 8

Pseudorandom Quantum States and Stabilizer Complexity

This chapter is based on [GIKL23c] and Section 4 of [GIKL23b], which were joint work with Sabee Grewal, Vishnu Iyer, and William Kretschmer. The lower bound based on stabilizer extent/fidelity is due to [GIKL23c]. The improved lower bound based on stabilizer dimension is due to [GIKL23b]. Some of the preliminary results are contained in Chapters 2 to 5 and 7.

This chapter concerns pseudorandom quantum states, first defined by Ji, Liu, and Song [JLS18], which are quantum states that appear indistinguishable from Haar-random states to computationally bounded adversaries. That is, an ensemble of n-qubit states is defined to be pseudorandom if no $\mathsf{poly}(n)$-time quantum algorithm can distinguish copies of a state drawn from the ensemble from copies of a Haar-random state, except with advantage negligible in n.

As a quantum analogue of cryptographic pseudorandom generators, pseudorandom states have recently attracted much attention in quantum cryptography and complexity theory. They suffice to build a wide range of cryptographic primitives, including quantum commitments, secure multiparty communication, one-time digital signatures, and more [JLS18, AQY22, MY22b, BCQ23, MY22a, HMY22]. The language of pseudorandom states has also been found to play a key role in resolving some paradoxes at the heart of black hole physics [BFV20, Bra22]. Finally, and perhaps most surprisingly, there is recent evidence to suggest that pseudorandom states can be constructed without assuming the existence of one-way functions [Kre21, KQST22].

Collectively, these results have motivated recent works that seek to characterize what computational properties or resources are required of pseudorandom states. For example, [ABF$^+$22] investigates the possibility of building pseudorandom quantum states with limited entanglement, and prove the existence of pseudorandom state ensembles with entanglement entropy substantially smaller than n, assuming the existence of quantum-secure one-way functions.

It is a well known fact that cryptography and learning are natural counterparts to one another. Hardness for learning is often derived from assumed cryptographically hard problems. In turn, an object that is easy to learn, in the computational sense, is not suitable for cryptographic purposes. As such, learning algorithms give natural lower-bounds on the kinds of resources needed to construct cryptographic objects. In this chapter in particular, we study quantum pseudorandom states from the perspective of stabilizer complexity and consider the number of non-Clifford gates in a circuit as a resource, similar to size or depth. This will be done entirely using the existing algorithms of Algorithms 2 and 4.

Theorem 8.1 (Informal version of Theorem 8.12)**.** *Let $|\psi\rangle$ be an unknown n-qubit pure state. There is an efficient algorithm that distinguishes whether $|\psi\rangle$ is Haar-random or a state with stabilizer fidelity at least τ, promised that one of these is the case. In particular, the algorithm uses $O(\tau^{-12}\log\frac{1}{\delta})$ copies of $|\psi\rangle$ and $O(n\tau^{-12}\log\frac{1}{\delta})$ time to succeed with probability at least $1-\delta$.*

The key idea is that the statistic measured in Algorithm 2 is related to how concentrated the $p_\psi(x)$ are and for a Haar random state the $p_\psi(x)$ should not be concentrated at all with overwhelming probability over the Haar measure.

Theorem 8.1 also generalizes to distinguishing states with low stabilizer extent from Haar-random. To the best of our knowledge, prior to our work, it was even unknown whether states of stabilizer extent at most a *constant* could be efficiently distinguished from Haar-random. We also emphasize that the contrast between our positive learning result and the hardness result of [HIN$^+$22] stems in part from the differing access models: we assume access to copies of the quantum state, whereas [HIN$^+$22] considers algorithms that only have outcomes of standard basis measurements of the state.

As a simple corollary, we prove a $\omega(\log n)$ lower bound on the number of T-gates required to prepare computationally pseudorandom quantum states by showing that the stabilizer extent decreases by at most a constant multiplicative factor for each T gate.

Corollary 8.2 (Corollary 8.14)**.** *Any family of Clifford+T circuits that produces an ensemble of n-qubit computationally pseudorandom quantum states must use at least $\omega(\log n)$ T-gates.*

We then give an exponential improvement on this lower bound based on stabilizer dimension (Definition 2.37). We start with the following result based on Algorithm 4.

Theorem 8.3 (Informal version of Theorem 8.17)**.** *Let $|\psi\rangle$ be an unknown n-qubit pure state. There is an efficient algorithm that distinguishes whether $|\psi\rangle$ is Haar-random or a state with stabilizer dimension at least 1, promised that one of these is the case. In particular, the algorithm uses $O(n+\log\frac{1}{\delta})$ copies of $|\psi\rangle$ and $O(n^3+n^2\log\frac{1}{\delta})$ time to succeed with probability at least $1-\delta$.*

We then observe that the stabilizer dimension is at least $n-2t$ for t single-qubit non-Clifford gates. In turn, a Haar random state has not only stabilizer dimension zero with high probability, it is also very far from any state with non-zero stabilizer dimension with high probability.

Corollary 8.4 (Informal version of Corollary 8.18)**.** *Any family of Clifford circuits that produces an ensemble of n-qubit computationally pseudorandom quantum states must use at least $n/2$ auxiliary non-Clifford single-qubit gates.*

In the special case that the non-Clifford gates are all diagonal (e.g., T-gates), our lower bound improves to n.

Under plausible computational assumptions, Corollary 8.4 is tight up to constant factors. In particular, the existence of linear-time quantum-secure pseudorandom functions (which are believed to exist [IKOS08, FLY22, GIKL23c]) implies the existence of linear-time constructible pseudorandom states [BS19, GIKL23c], which of course have at most $O(n)$ non-Clifford gates. The key idea is that Haar-random states have minimal stabilizer dimension (Definition 2.37) with overwhelming probability, whereas each single-qubit non-Clifford gate decreases the stabilizer dimension by at most 2.

Under slightly weaker computational assumptions, Theorem 8.1 is also tight, in terms of stabilizer fidelity rather than T gate count. Assuming the existence of quantum-secure one way functions (that need not be linear-time constructible) [ABF+22] show how to construct pseudorandom quantum states with arbitrary inverse super-polynomial stabilizer fidelity.

8.1 Stabilizer Extent of t-doped Clifford Circuits

To prove lower bounds, via stabilizer fidelity, on the number of T-gates necessary to prepare pseudorandom quantum states, we need to lower bound the stabilizer fidelity of a quantum state prepared by some number of non-Clifford gates. This will be analogous to Section 7.2, where the stabilizer dimension was also lower bounded in the number of non-Clifford gates.

Recall that a t-doped circuit is a quantum circuit comprised only of Clifford gates (i.e., Hadamard, Phase, and CNOT) and at most t single-qubit non-Clifford gates. We will will now *upper bound* the stabilizer extent, which immediately give *lower bounds* on stabilizer fidelity as well by Claim 2.36.

Claim 8.5. *For* $|\psi\rangle = \alpha|v\rangle + \beta|w\rangle$*,*

$$\xi(|\psi\rangle) \leq \left(|\alpha|\sqrt{\xi(|v\rangle)} + |\beta|\sqrt{\xi(|w\rangle)}\right)^2.$$

Proof. Let $|v\rangle = \sum_i c_i|\phi_i\rangle$ and $|w\rangle = \sum_j d_j|\varphi_j\rangle$ be the minimal decompositions in terms of stabilizer extent (i.e., $(\sum_i |c_i|)^2 = \xi(|v\rangle)$). Since $|\psi\rangle = \alpha|v\rangle + \beta|w\rangle = \alpha\sum_i c|\phi_i\rangle + \beta\sum_j d|\varphi_j\rangle$, we have a stabilizer decomposition of $|\psi\rangle$. The stabilizer extent of this decomposition is at most

$$\left(\sum_i |\alpha c_i + \beta d_i|\right)^2 \leq \left(|\alpha|\sum_i |c_i| + |\beta|\sum_i |d_i|\right)^2 = \left(|\alpha|\sqrt{\xi(v)} + |\beta|\sqrt{\xi(w)}\right)^2. \qquad \square$$

We now prove our stabilizer extent upper bound based on how non-Clifford the circuit is. Note that we give bounds specifically in terms of the T gate, but that the result can be generalized to all t-doped circuits by adjusting the parameters of the decomposition in the proof of Lemma 8.6.

Lemma 8.6. *Let* C *be a* t*-doped Clifford circuit such that all of the non-Clifford gates are the* T *gate. Let* $|\psi\rangle = C|0^n\rangle$*. Then,*

$$\xi(|\psi\rangle) \leq \left(1 + \frac{1}{\sqrt{2}}\right)^t.$$

Proof. We observe that a t-doped Clifford circuit can be broken into layers of Clifford circuits, followed by a single non-Clifford gate, followed by more layers of Clifford circuits, and so on. Since Clifford circuits preserve stabilizer extent, we only need to show that the T-gate increases the stabilizer extent of any state by at most a constant multiplicative factor. Since the SWAP gate is a Clifford operation, we assume without loss of generality that each T-gate is applied to the first qubit.

We proceed by induction on the layers of the circuit. In the first layer, when no T-gates have been applied, the bound is trivially true because the stabilizer extent of any stabilizer state is 1. Now, assume that, after applying some portion of the circuit C' to $|0^n\rangle$ with $t-1$

T-gates, we get the state $|\varphi\rangle$. Observe that the T-gate can be expressed as $\cos(\pi/8)e^{i\pi/8}I + \sin(\pi/8)e^{i13\pi/8}Z$. Thus, $(T \otimes I^{\otimes n-1})|\varphi\rangle = \cos(\pi/8)e^{i\pi/8}|\varphi\rangle + \sin(\pi/8)e^{i13\pi/8}(Z \otimes I^{\otimes n-1})|\varphi\rangle$. Since $Z \otimes I^{\otimes n-1}$ is a Clifford operation, $(Z \otimes I^{\otimes n-1})|\varphi\rangle$ has the same extent as $|\varphi\rangle$. Therefore, applying Claim 8.5,

$$\xi(|\psi\rangle) \leq (\cos(\pi/8) + \sin(\pi/8))^2\, \xi(|\varphi\rangle) \leq \left(1 + \frac{1}{\sqrt{2}}\right)^t. \qquad \square$$

We remark that due to Fact 5.9, this exponential scaling in t is tight, though the actual base of the exponent is likely very loose.

8.1.1 Bell difference sampling for Haar-Random States

Haar-random states, due to behaving like gaussian random vectors, concentrate extremely well. This makes the Bell difference sampling distribution very nice with high probability over the Haar measure. We formalize this idea by showing that all of the $p_\psi(x)$ for $x \neq 0$ cannot be to big with overwhelmingly high probability.

We first require the following lemma, which shows that the Weyl measurements are concentrated around 0. The building block is a concentration inequality known as Lévy's Lemma.

Lemma 8.7 (Lévy's Lemma, see e.g., [Ger13]). *Let $\mathbb{S}^N$ denote the set of all N-dimensional pure quantum states, and let $f : \mathbb{S}^N \to \mathbb{R}$ be L-Lipschitz, meaning that $|f(|\psi\rangle) - f(|\varphi\rangle)| \leq L \cdot \||\psi\rangle - |\varphi\rangle\|_2$. Then:*

$$\Pr_{|\psi\rangle \sim \mu_{\text{Haar}}}[|f(|\psi\rangle) - \mathbf{E}[f]| \geq \varepsilon] \leq 2\exp\left(-\frac{N\varepsilon^2}{9\pi^3 L^2}\right).$$

Lemma 8.8. *For any n-qubit Weyl operator W_x, the function $f_x : \mathbb{S}^{2^n} \to \mathbb{R}$ defined by $f_x(|\psi\rangle) = \langle\psi| W_x |\psi\rangle$ is 2-Lipschitz.*

Proof. Write $W_x = \Pi_+ - \Pi_-$ where Π_+ and Π_- are the projectors onto the positive and negative eigenspaces of W_x, respectively. Then,

$$
\begin{aligned}
|f_x(|\psi\rangle) - f_x(|\varphi\rangle)| &= |\langle\psi| W_x |\psi\rangle - \langle\varphi| W_x |\varphi\rangle| \\
&= |\langle\psi| \Pi_+ |\psi\rangle - \langle\varphi| \Pi_+ |\varphi\rangle - \langle\psi| \Pi_- |\psi\rangle + \langle\varphi| \Pi_- |\varphi\rangle| \\
&\le |\langle\psi| \Pi_+ |\psi\rangle - \langle\varphi| \Pi_+ |\varphi\rangle| + |\langle\psi| \Pi_- |\psi\rangle + \langle\varphi| \Pi_- |\varphi\rangle| \\
&= |\ \|\Pi_+ |\psi\rangle\|_2 - \|\Pi_+ |\varphi\rangle\|\|_2 + \|\|\Pi_- |\psi\rangle\|_2 - \|\Pi_- |\varphi\rangle\|_2| \\
&\le \|\Pi_+(|\psi\rangle - |\varphi\rangle)\|_2 + \|\Pi_-(|\psi\rangle - |\varphi\rangle)\|_2 \\
&\le 2\||\psi\rangle - |\varphi\rangle\|_2,
\end{aligned}
$$

where the third and fifth lines apply the triangle inequality, and the fourth and sixth lines use the fact that Π_+ and Π_- are projectors. □

Corollary 8.9. *Let W_x be any n-qubit Weyl operator in which $x \neq 0$ (i.e., $W_x \neq I$). Then:*

$$
\mathop{\mathbf{Pr}}_{|\psi\rangle \sim \mu_{\text{Haar}}} [|\langle\psi|W_x|\psi\rangle| \ge \varepsilon] \le 2 \exp\left(-\frac{2^n \varepsilon^2}{36\pi^3}\right).
$$

Proof. Define $f_x(|\psi\rangle) = \langle\psi| W_x |\psi\rangle$ as in Lemma 8.8. By Lemma 8.8, we know that f_x is 2-Lipschitz. Additionally, observe that $\mathbf{E}[f] = 0$ over the Haar measure because exactly half of the eigenvalues of W_x are 1 and the other half are -1. Then the corollary follows from Lemma 8.7. □

Corollary 8.10.

$$
\mathop{\mathbf{Pr}}_{|\psi\rangle \sim \mu_{\text{Haar}}} [\exists x \neq 0 : |\langle\psi|W_x|\psi\rangle| \ge \varepsilon] \le 2^{2n+1} \exp\left(-\frac{2^n \varepsilon^2}{36\pi^3}\right).
$$

Proof. This follows from Corollary 8.9 and a union bound over all 2^{2n} possible Weyl operators. □

Note that if $\varepsilon \ge \frac{1}{\text{poly}(n)}$, then the probability bound in Corollary 8.10 is doubly-exponentially small.

We have now shown that, with high probability, all Weyl measurements (except $W_x = I$) are close to 0. As a result, we can think of p_ψ (and critically q_ψ as well, due to Corollary 4.3) as well-spread and similar to the uniform distribution over $\mathbb{F}_2^{2n}$ due to the lack of heavy-hitters.

8.2 Distinguishing From Haar-Random Based on Stabilizer Fidelity

To efficiently distinguish a state with low stabilizer complexity (meaning, a state with low stabilizer extent or non-negligible stabilizer fidelity) from a Haar-random one, we require a property or statistic of the state that distinguishes it from Haar-random. It turns out that η from Algorithm 2 is actually one! We have already proven that $\eta \geq F_{\mathcal{S}}(|\psi\rangle)^6$ (Proposition 5.3), so now we need to show that η is small for a Haar random state with high probability over the Haar measure.

Lemma 8.11. *Let $|\psi\rangle$ be an n-qubit pure state. If the stabilizer fidelity of $|\psi\rangle$ is at least τ, then*

$$\eta \coloneqq \mathop{\mathbf{E}}_{x \sim q_\psi} \left[|\langle \psi | W_x | \psi \rangle|^2 \right] \geq \tau^6.$$

In contrast, if $|\psi\rangle$ is Haar-random, then, with probability at least

$$1 - 2^{2n+1} \exp\left(-\frac{2^{5n/6}}{36\pi^3} \left(\frac{2^n - 2^{n/2}}{4^n - 1} \right)^{1/3} \right)$$

over the Haar measure,

$$\eta \coloneqq \mathop{\mathbf{E}}_{x \sim q_\psi} \left[|\langle \psi | W_x | \psi \rangle|^2 \right] \leq 2^{-n/2}.$$

Proof. The first part is already stated as Proposition 5.3.

Now suppose $|\psi\rangle$ is a Haar-random state. By Corollary 8.10, for all $W_x \neq I$,

$|\langle\psi|W_x|\psi\rangle|^2 = 2^n p(x) \leq \varepsilon^2$ with probability $1 - 2^{2n+1}\exp\left(-\frac{2^n\varepsilon^2}{36\pi^3}\right)$. Therefore by Fact 5.2,

$$\begin{aligned}\eta &= 4^n \sum_{x\in\mathbb{F}_2^{2n}} p(x)^3 \\ &= 4^n\left(\frac{1}{8^n} + \sum_{x\neq 0} p(x)^3\right) \\ &\leq \frac{1+(4^n-1)\varepsilon^6}{2^n},\end{aligned}$$

with probability at least $1 - 2^{2n+1}\exp\left(-\frac{2^n\varepsilon^2}{36\pi^3}\right)$. By setting $\epsilon^2 = \frac{1}{2^{n/6}}\left(\frac{2^n-2^{n/2}}{4^n-1}\right)^{1/3}$, we get

$$\eta \leq \frac{1}{2^{n/2}}$$

with probability at least $1 - 2^{2n+1}\exp\left(-\frac{2^{5n/6}}{36\pi^3}\left(\frac{2^n-2^{n/2}}{4^n-1}\right)^{1/3}\right)$.

□

We note that while the expression is complicated, the failure probability is doubly-exponentially for suitably large n.

Theorem 8.12. *Let $|\psi\rangle$ be an unknown n-qubit pure state. Algorithm 2 distinguishes whether $|\psi\rangle$ is Haar-random or a state with stabilizer fidelity at least τ, promised that one of these is the case. The algorithm uses $O\left(\tau^{-12}\log\frac{1}{\delta}\right)$ copies of $|\psi\rangle$ and $O(n\tau^{-12}\log\frac{1}{\delta})$ time, and distinguishes the two cases with success probability at least $1-\delta$.*

Let us assume that the second part of Lemma 8.11 is true. Our algorithm then amounts to estimating the quantity η via Algorithm 2 with $\beta_1 = \tau^6$ and $\beta_2 = 2^{-n/2}$ to probability $\delta/2$. For appropriately large n and $\tau = O(\mathsf{poly}(n))$, the gap $\gamma = \beta_1 - \beta_2 = O(\tau^6)$. Thus Lemma 5.5 tells us that for $\tau = O(\mathsf{poly}(n)$ and sufficiently large n, we can distinguish between low-stabilizer-complexity and Haar-random states using $O\left(\frac{1}{\tau^{12}}\log\frac{1}{\delta}\right)$ samples and $O\left(\frac{n}{\tau^{12}}\log\frac{1}{\delta}\right)$ time. By the union bound, our failure probability is at most

$\delta/2 + 2^{2n+1} \exp\left(-\frac{2^{5n/6}}{36\pi^3}\left(\frac{2^n - 2^{n/2}}{4^n - 1}\right)^{1/3}\right)$ which is less than δ for suitably large n (and δ that is not doubly-exponentially small).

All of these results also apply to states with stabilizer extent at most τ^{-1}, since by Claim 2.36, such states have stabilizer fidelity at least τ.

Corollary 8.13. *Let $|\psi\rangle$ be an unknown n-qubit pure state. Algorithm 2 distinguishes whether $|\psi\rangle$ is Haar-random or a state with stabilizer extent at most τ^{-1}, promised that one of these is the case. The algorithm uses $O\left(\tau^{-12}\log\frac{1}{\delta}\right)$ copies of $|\psi\rangle$ and $O(n\tau^{-12}\log\frac{1}{\delta})$ time, and distinguishes the two cases with success probability at least $1-\delta$.*

The above result immediately implies that output states of Clifford+T circuits with few T-gates cannot be computationally pseudorandom.

Corollary 8.14. *Any family of Clifford+T circuits that produces an ensemble of n-qubit computationally pseudorandom quantum states must use at least $\omega(\log n)$ T-gates.*

Proof. Consider any ensemble of states wherein each state in the ensemble is the output of some Clifford+T circuit with at most $K\log n$ T-gates. By Lemma 8.6, the stabilizer extent of any such state $|\psi\rangle$ is at most $n^{\alpha K}$ for $\alpha \leq 0.7716$. By Corollary 8.13, on input copies of $|\psi\rangle$, the algorithm from Theorem 8.12 takes $O(n^{12\alpha K+1}) \leq \mathsf{poly}(n)$ time and outputs 1 with probability at least $\frac{2}{3}$. On the other hand, if $|\psi\rangle$ is a Haar-random state then the same algorithm outputs 1 with probability at most $\frac{1}{3}$. As such, the algorithm's distinguishing advantage between the ensemble and the Haar measure is non-negligible. This is to say that the ensemble cannot be pseudorandom under the definition of [JLS18]. □

8.3 Distinguishing From Haar-Random Based on Stabilizer Dimension

Similar to how Algorithm 2 can be used to test for Haar randomness, so too can Algorithm 4. In particular (just like the fidelity setting's completeness case), we already know that if the stabilizer dimension is greater than zero, then the support of q_ψ can have dimension at most $2n-1$. This would give us perfect completeness, and indeed Corollary 7.4 tells us that if we use less than $n/2$ non-Clifford gates than this will be the case.

For soundness, we need to show that if $|\psi\rangle$ is Haar-random, then q_ψ is well-supported over the entirety of $\mathbb{F}_2^{2n}$ in the sense that every proper subspace of $\mathbb{F}_2^{2n}$ contains a bounded fraction of the q_ψ mass. This implies that sampling from q_ψ gives $2n$ linearly independent elements of $\mathbb{F}_2^{2n}$ after a reasonable number of iterations. The idea is to combine the concentration of p_ψ (Corollary 8.10) with the fact (Theorem 4.2) that the q_ψ mass on a subspace is proportional to its p_ψ^2 mass on the symplectic complement to obtain the following.

Lemma 8.15. *Let $|\psi\rangle$ be a Haar-random n-qubit state. Then with probability all subspaces $T \subseteq \mathbb{F}_2^{2n}$ of dimension $2n-1$ simultaneously satisfy*

$$\sum_{x\in T} q_\psi(x) \leq \frac{2}{3},$$

except with probability at most

$$2^{2n+1}\exp\left(-\frac{2^n}{36\sqrt{3}\pi^3}\right).$$

Proof. Let T be any subspace of dimension $2n-1$. Then the symplectic complement $T^\perp$ has dimension 1, so it is the span of a single nonzero $x \in \mathbb{F}_2^{2n}$. By Theorem 4.2,

$$\sum_{a\in T} q_\psi(a) = 2^{2n-1}\sum_{a\in T^\perp} p_\psi(a)^2 = \frac{1+\langle\psi|W_x|\psi\rangle^4}{2}.$$

Hence, the probability that there exists a T for which $\sum_{x\in T} q_\psi(x)$ exceeds $\frac{2}{3}$ is at most the probability that there exists a nonzero x for which $|\langle\psi|W_x|\psi\rangle| \geq \frac{1}{\sqrt[4]{3}}$. By Corollary 8.10, this probability is at most $2^{2n+1}\exp\left(-\frac{2^n}{36\sqrt{3}\pi^3}\right)$. □

We are now ready to analyze the soundness case for when $|\psi\rangle$ is Haar random.

Lemma 8.16. *Let $|\psi\rangle$ be an n-qubit Haar-random quantum state and fix $\delta > 0$. Taking $6n + \frac{9}{2}\log\frac{2}{\delta}$ samples from q_ψ suffices to sample $2n$ linearly independent elements of $\mathbb{F}_2^{2n}$ with probability at least $1-\delta$ over both the Haar measure and the sampling process.*

Proof. By Lemma 8.15, $\sum_{x\in T_i'} q_\psi(x) \leq \frac{2}{3}$ for all T' with $\dim(T') = 2n-1$, with overwhelmingly high probability over the Haar measure. Let us assume that this has happened. By the fact that $\sum_{x\in\mathbb{F}_2^{2n}} q_\psi(x) = 1$, we find that

$$\sum_{x\in\mathbb{F}_2^{2n}\setminus T'} q_\psi(x) \geq \frac{1}{3}.$$

Applying Lemma 3.10 with $c = \frac{1}{3}$, we can show that $6n + 3\log\frac{2}{\delta}$ samples suffice to generate $\mathbb{F}_2^{2n}$ with probability at least $1-\frac{\delta}{2}$. By the union bound, the total failure probability over both the Haar measure and the samples is at most

$$\frac{\delta}{2} + 2^{2n+1}\exp\left(-\frac{2^n}{36\sqrt{3}\pi^3}\right)$$

which in turn is at most δ, for reasonable choices of δ.[1] □

Theorem 8.17. *Algorithm 4 with parameters $k = n-1$, $\varepsilon = 1/4$, and failure probability $\delta/2$ succeeds at distinguishing whether $|\psi\rangle$ is Haar-random or a state with stabilizer dimension at least 1 with probability at least $1-\delta$. It uses $O(n+\log\frac{1}{\delta})$ copies of of the input state and $O(n^3 + n^2\log\frac{1}{\delta})$ time.*

Proof. In the case that $|\psi\rangle$ is has stabilizer dimension non-zero, q_ψ is supported on a subspace of dimension at most $2n-1$ by Corollary 4.5, so the algorithm always outputs 1. Therefore, we need only argue that in the case that $|\psi\rangle$ is Haar-random we see at least $2n$ linearly

[1] Of course, the union bound fails when δ is doubly exponentially small, as our bound for the error over the Haar measure is $2^{-2^{O(n)}}$. However, in this setting, it is information-theoretically impossible to distinguish a state from Haar-random.

independent elements of $\mathbb{F}_2^{2n}$. We note that the number of samples used by Algorithm 4 with parameters $k = n-1$, $\varepsilon = 1/4$, and failure probability $\delta/2$ is

$$\frac{32n + 8\log\frac{2}{\delta}}{\varepsilon} = 128n + 32\log\frac{2}{\delta}.$$

Since this is greater than the $6n + \frac{9}{2}\log\frac{1}{\delta}$ samples needed for Lemma 8.16, we will see $2n$ linearly independent elements of $\mathbb{F}_2^{2n}$ with probability at least $1-\delta$ over the Haar measure and the sampling process.

It is clear that the sample complexity is $O(m) = O(n + \log\frac{1}{\delta})$. To analyze the time complexity, we note that each sample takes $O(n)$ time, so sampling takes $O(mn)$ time. Gaussian elimination takes $O(mn^2) = O(n^3 + n^2\log\frac{1}{\delta})$ time and is the dominating term. □

In the case that $|\psi\rangle$ is the output of a t-doped Clifford circuit, q_ψ is supported on a subspace of dimension at most $n + 2t$ by Corollary 7.4, so the algorithm always outputs 1. This immediately implies a lower bound on the number of single-qubit non-Clifford gates needed to prepare an ensemble of n-qubit computationally pseudorandom quantum states.

Corollary 8.18. *Any family of t-doped Clifford circuits that produces an ensemble of n-qubit computationally pseudorandom quantum states must satisfy $t \geq n/2$.*

Proof. By Corollary 7.4, if $t < n/2$ then the output states will have stabilizer dimension at least 1. By Theorem 8.17, there exists an efficient algorithm that can distinguish these state from Haar random with non-negligible probability. □

Finally, note that this lower bound can be improved by a factor of 2 in the special case that all of the non-Clifford gates are diagonal (e.g., T-gates), because of the improved lower bound on stabilizer dimension in Lemma 7.3 for this case.

Corollary 8.19. *Any family of Clifford+T circuits that produces an ensemble of n-qubit computationally pseudorandom quantum states must use at least n T-gates.*

Part II

Statistical Learning Theory for Quantum States and Circuits

Chapter 9

Statistical Learning Theory Preliminaries

This chapter is an overview of basics concepts in Learning Theory. It borrows from [Lia23] and [GL22], which was joint work with Aravind Gollakota.

9.1 PAC learning

The goal of PAC learning is to learn a function relative to a certain distribution of inputs, rather than in an absolute sense. Let's say we want to learn an arbitrary f from some concept class $\mathcal{C}$. If a hypothesis function h matches the true function f on many of the high probability inputs, then we can say that we have *approximately learned* f. If we can do this with high probability for arbitrary f, then we *probably* approximately (PAC) learned $\mathcal{C}$.

Definition 9.1. *Let Ω be some domain of inputs and let $\mathcal{C}$ be a set of functions $f : \Omega \to [0, 1]$. We say that $\mathcal{C}$ is (ε, δ)-PAC-learnable if there exists a learner that, when given samples of the form $(x, f(x))$ for $x \sim \mathcal{D}$ for arbitrary f and unknown distribution $\mathcal{D}$, outputs with probability at least $1 - \delta$, over both the samples and the learning algorithm, a hypothesis h with error[1] satisfying*

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(f(x) - h(x))^2\right] \le \varepsilon.$$

The number of samples used is referred to as the sample complexity, and we refer to

[1] Since this problem is a regression problem rather than binary classification, we claim that this squared-loss is a more natural notion of error than the kind used in [Aar07], [CD20], and other older papers [AB00]. It also allows us to express PAC learning using only 2 error parameters, rather than 3. We can also still recover some form of the 3 parameter bound using $\mathbf{Pr}_{x\sim\mathcal{D}}[|f(x) - h(x)| > t] \le \frac{1}{t^2} \mathbb{E}_{x\sim\mathcal{D}}\left[(f(x) - h(x))^2\right]$ via Markov's inequality.

the learner being efficient if it can find such an h in time $\mathsf{poly}(n, \varepsilon^{-1}, \delta^{-1})$ *for arbitrary ε and δ.*

From here, one can define two types of learning, based on where h comes from. If h is allowed to be any function that meets the PAC constraints, we refer to this as IMPROPER LEARNING. If instead $h \in \mathcal{C}$, we get what is known as PROPER LEARNING, which will be the focus of the proceeding chapters. With proper learning, we can then begin to talk about the consistency problem formally.

Definition 9.2. *Let S refer to a set of labeled samples.* ConsistentSearch$(\mathcal{C}, s)$ *is the problem of, given an arbitrary S such that $|S| < s$, finding a function $h \in \mathcal{C}$ that is consistent with all of S (i.e., for all $(x, f(x)) \in S$, $f(x) = h(x)$) if such an h exists, otherwise reject.*

9.1.1 Generalization

Intuitively, given a set of samples the best one can really hope to do is find such an h that gets zero training error and hope that the true error for h is also low. This leads to the idea of *generalization*, which aims to show that doing well on a large enough set of training data (i.e., the consistency problem) allows one to give the PAC guarantee as well with high probability. In terms of computational efficiency, this effectively reduces the problem of proper learning to the consistency problem, or an approximation of the consistency problem. The most common approach to guarantee generalization is to bound the "expressiveness" of the concept class, such as with the VC-dimension [BEHW89]. Since VC-dimension is defined for $\{0, 1\}$ labels, we will now give a generalization of VC-dimension in the regression setting.

Definition 9.3. *Let Ω be some domain of inputs, $\eta > 0$, and let $\mathcal{C}$ be a set of functions $f : \Omega \to [0, 1]$. We say that a set of inputs $\{x_1, x_2, \cdots, x_m\} \subseteq \Omega$ is η-fat-shattered by $\mathcal{C}$ if there exists a set $y_1, y_2, \cdots, y_m \in [0, 1]$ such that for any vector $b = \{\pm 1\}^m$ there is an $f_C \in \mathcal{C}$ that satisfies $b_i \cdot (f_C(x_i) - y_i) \geq \eta$.*

Definition 9.4. *The η-fat-shattering dimension of a concept class $\mathcal{C}$ is the size of the largest set of inputs that is fat-shattered by $\mathcal{C}$. We denote this as $fat_{\mathcal{C}}(\eta)$.*

At a high-level, the η parameter provides a buffer such that each $f_C(x_i)$ is robustly bounded away from y_i by η in the appropriate direction. We now give a result saying that bounded fat-shattering dimension implies generalization from the training data to the actual learning task.

Theorem 9.5 ([AB00] Corollary 3.3)**.** *Let $\mathcal{C}$ be a concept class from Ω to $[0,1]$ and let $\mathcal{D}$ be some distribution on Ω. Let $\delta, \varepsilon, \alpha, \beta \in (0,1)$ be parameters such that $\beta > \alpha$. Furthermore, let $\{x_1, x_2, \cdots, x_m\}$ and $\{y_1, y_2, \cdots, y_m\}$ be a set of m samples drawn i.i.d. from some distribution $\mathcal{D}$ where $y_i = f(x_i)$ for some $f \in \mathcal{C}$. If $h \in \mathcal{C}$ satisfies $|h(x_i) - y_i| \leq \alpha$ for all $1 \leq i \leq m$ then*

$$m = \mathcal{O}\left(\frac{1}{\varepsilon}\left(fat_{\mathcal{C}}\left(\frac{\beta-\alpha}{8}\right)\log^2\left(\frac{fat_{\mathcal{C}}\left(\frac{\beta-\alpha}{8}\right)}{(\beta-\alpha)\varepsilon}\right) + \log\frac{1}{\delta}\right)\right)$$

number of samples suffices to achieve

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(h(x)-f(x))^2\right] \leq (1-\varepsilon)\beta^2 + \varepsilon$$

with probability at least $1-\delta$ over the samples.

The following folklore bound on fat-shattering dimension is very loose, but still sufficient for our purposes of complexity-theoretic hardness in Section 10.5.

Lemma 9.6 (Folklore)**.** *Given a concept class $\mathcal{C}$ such that $|\mathcal{C}|$ is finite, then for all $\eta > 0$, $fat_{\mathcal{C}}(\eta) \leq \log_2|\mathcal{C}|$.*

Proof. Assume for the sake of contradiction that $\mathcal{C}$ η-fat-shatters the set of points $\{x_1, x_2, \cdots, x_m\}$ for $m > \log_2|\mathcal{C}|$. Then $\mathcal{C}$ must be able to properly match each $b \in \{\pm 1\}^m$. There are $2^m > |\mathcal{C}|$ possible b vectors, and only one $f \in \mathcal{C}$ can be used per shattering attempt, since no f can ever satisfy two different b vectors. This is a contradiction since we don't have enough $f \in \mathcal{C}$ to go around to satisfy every b vector. □

9.1.2 Decision Problems

One can also define the decision version of the consistency problem, which is deciding if there even exists an $h \in \mathcal{C}$ that is consistent with all of S. We show that the existence of efficient learning algorithms can imply efficient one-sided error algorithms for the decision version of the consistency problem.

Definition 9.7. *Let* $\textsc{ConsistentDecide}(\mathcal{C}, s)$ *be decision version of the consistency problem for* $\mathcal{C}$ *using at most* s *samples.*

Proposition 9.8. *An efficient randomized* $\left(\varepsilon < \frac{\alpha^2}{s^2}, \delta < \frac{1}{2} + \frac{1}{2s}\right)$[2] *proper learning algorithm implies* $\textsc{ConsistentDecide}(\mathcal{C}, s) \in \mathsf{RP}$ *where* $\alpha = \inf_{x \in \Omega, f(x) \neq g(x)} |f(x) - g(x)|$ *is the minimum non-zero error any hypothesis function can make on a single input.*

Proof. For every set of samples S such that $|S| \leq s$, we can define the $\mathcal{D}_S$ to be the uniform distribution over all $x \in \chi$ such that $(x, f(x)) \in S$. By coupon collector, if we draw $\mathcal{O}(s \log s)$ many samples then with probability at least $1 - \frac{1}{s}$ we will have drawn every item from S. Now imagine that there exists some hypothesis $h \in \mathcal{C}$ that is not consistent with S. Then our error must be at least $\frac{\alpha^2}{s^2}$ by the definition of α.

Now assume we have some efficient randomized (ε, δ) proper learning algorithm for $\varepsilon < \frac{\alpha^2}{s^2}$ and $\delta < \frac{1}{2} + \frac{1}{2s}$. When running the learner on an arbitrary $\mathcal{D}_S$, it will see samples S with probability at least $1 - \frac{1}{s}$. To get error less than $\frac{\alpha^2}{s^2}$ the learner must then be able to solve the search version consistency problem with probability p such that $\frac{1}{s} + (1 - \frac{1}{s})p \geq 1 - \delta$. Solving for p we find $p \geq \frac{1}{2}$ on accepting instances.

This gives rise to the following algorithm in RP for solving $\textsc{ConsistentDecide}(\mathcal{C}, s)$. Given samples S with $|S| \leq s$, we can run our learning algorithm and pretend that S is what we sampled from $\mathcal{D}_S$ to get hypothesis h. If h is consistent with S then accept, otherwise

[2] We abuse notation to signify that ε is a value less than $\frac{\alpha^2}{s^2}$ and likewise for $\delta < \frac{1}{2s}$.

reject. On an accepting instance h will be consistent with probability at least $\frac{1}{2}$ while on rejecting instances it will never be consistent so the algorithm will always reject. □

Informally, if there exists enough structure on the concept class, it can be possible to go the other way and show that an efficient algorithm for $\textsc{ConsistentDecide}(\mathcal{C}, s)$ implies an efficient proper learner for $\mathcal{C}$. Namely, if a search-to-decision reduction exists for the consistency problem on $\mathcal{C}$ and $fat_{\mathcal{C}}$ is finite then we can also expect to show that an efficient algorithm for the decision problem would imply an efficient proper learner for $\mathcal{C}$. Of particular interest are NP-complete problems, which always admit search-to-decision reductions [Kat11]. We can now give a formal proof of this commonly used technique to show proper PAC learning if $\mathsf{RP} = \mathsf{NP}$.

Lemma 9.9. *Let $\mathcal{C}$ be a concept class and let*

$$m = \Theta\left(\frac{1}{\varepsilon}\left(fat_{\mathcal{C}}\left(\frac{\beta}{8}\right)\log^2\left(\frac{fat_{\mathcal{C}}(\frac{\beta}{8})}{\beta\varepsilon}\right) + \log\frac{1}{\delta}\right)\right)$$

be the parameter from Theorem 9.5 with $\alpha = 0$. If $\textsc{ConsistentDecide}(\mathcal{C}, s)$ for $s > m$ is NP-complete and $\mathsf{RP} = \mathsf{NP}$ then for $c = \mathcal{O}(\mathsf{poly}(n))$ there exists an efficient algorithm to $\left((1-\varepsilon)\beta^2 + \varepsilon, \delta + \frac{1}{2^c}\right)$ proper learn $\mathcal{C}$.

Proof. Because search-to-decision reductions exist for all NP-complete problems [Kat11], a zero-error efficient black-box algorithm for $\textsc{ConsistentDecide}(\mathcal{C}, s)$ can be used to efficiently solve $\textsc{ConsistentSearch}(\mathcal{C}, s)$ as well. Call this algorithm $\mathcal{Z}$. Let us run $\mathcal{Z}$ on a sample S such that $s > |S| \geq m$. Since the data is produced by something in $\mathcal{C}$, we are guaranteed to succeed in search. This means we now have an $h \in \mathcal{C}$ such that

$$|h(x_i) - f(x_i)| = 0 \quad \forall 1 \leq i \leq k$$

and so by Theorem 9.5

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(h(x) - f(x))^2\right] \leq (1-\varepsilon)\beta^2 + \varepsilon$$

with probability at most δ over the samples.

Our goal now will be to show that, even with random errors due to RP, we can make it so that the probability of differing from $\mathcal{Z}$ is arbitrarily small. To start, let γ be the number of calls to $\textsc{ConsistentDecide}(\mathcal{C}, s)$ used in the search-to-decision reduction for the construction of $\mathcal{Z}$. In order for the reduction to be efficient, $\gamma = \mathcal{O}(\mathsf{poly}(n))$. Since $\textsc{ConsistentDecide}(\mathcal{C}, s)$ is in NP and therefore RP, we have an efficient one-sided constant error algorithm $\mathcal{A}$ for $\textsc{ConsistentDecide}(\mathcal{C}, s)$. Using $O(c + \log \gamma) = \mathcal{O}(\mathsf{poly}(n))$ many calls to $\mathcal{A}$ and taking the majority, we can get error at most $\frac{1}{\gamma \cdot 2^c}$. Call this new algorithm $\mathcal{A}'$ and use it in place of the zero-error oracle for $\textsc{ConsistentDecide}(\mathcal{C}, s)$. By the union bound over all γ calls to $\mathcal{A}'$, the probability that any query to $\mathcal{A}'$ differs from the zero-error oracle is at most $\frac{1}{2^c}$. Thus the probability that the output of $\mathcal{A}'$ differs $\mathcal{Z}$ is also at most $\frac{1}{2^c}$.

By the union bound over both the samples and the error in $\mathcal{A}'$, the total error probability is at most $\delta + \frac{1}{2^c}$. □

9.2 PAC Learning Quantum States

A fundamental task in quantum computing is that of learning a description of an unknown quantum state ρ. Traditionally this is formalized as the problem of quantum state tomography, where we are granted the ability to form multiple copies of ρ and take arbitrary measurements, and must learn a state σ that is close to ρ in trace distance. In an influential work, Aaronson [Aar07] introduced the "Probably Approximately Correct" (PAC) framework from computational learning theory [Val84] as an alternative perspective on this problem. Here the key innovation is that instead of learning ρ in an absolute metric (such as trace distance), we only wish to learn it with respect to a pre-specified distribution on measurements. This requirement is considerably weaker than that of full tomography. At this point, it was already established that full tomography would take exponential time, but Aaronson showed that only a linear number of measurements were necessary for a PAC learner!

We begin by formally defining the problem of PAC-learning a quantum state.

Definition 9.10 (PAC-learnability of quantum states, [Aar07]). *Let $\mathcal{F}$ be a class of n-qubit quantum states. Let $\mathcal{D}$ be a distribution over $\mathcal{E}$, the set of 2-outcome measurements. We say $\mathcal{F}$ is (ε, δ) PAC-learnable with respect to $\mathcal{D}$ if there exists a learner that, given sample access to labeled examples $(E, \mathrm{Tr}[E\rho])$ for $E \sim \mathcal{D}$ and unknown $\rho \in \mathcal{F}$, is able to output a state σ satisfying*

$$\mathop{\mathbb{E}}_{E\sim\mathcal{D}}\left[(\mathrm{Tr}[E\sigma] - \mathrm{Tr}[E\rho])^2\right] \leq \varepsilon$$

with probability at least $1 - \delta$.

We note that this is a slight modification of the original definition in [Aar07], stated directly in terms of squared loss since this is the view that will be convenient for us. It is also very important to stress that in this framework, while the data *arises from* a quantum state, it is entirely classical in form and representation. That is, all of the quantum actions have already been taken, and we are at the stage of classical post-processing.

With PAC learners, one may speak of both computational efficiency (overall running time) and statistical or information-theoretic efficiency (sample complexity). The original result of Aaronson [Aar07] described a computationally inefficient algorithm for learning arbitrary states that nevertheless had $O(n)$ sample complexity. An *efficient* PAC learner is one that is computationally efficient, i.e., runs in polynomial time, and hence also draws at most polynomially many examples (each draw is considered as taking one unit of time).

Aaronson was able to prove the following generalization result for quantum states. We present a modified form of it in terms of squared loss.

Proposition 9.11 ([Aar07] Theorem 1.1[3]). *Let ρ be an n-qubit mixed state, $\mathcal{D}$ a distribution over two-outcome measurements of ρ and let $\{E_1, \cdots, E_m\} \subseteq \mathcal{E}$ be a training set consisting of*

[3] For an alternative proof to [Aar07], see [Aar22, Lecture 22].

m measurements drawn from $\mathcal{D}$. In addition, fix error parameters $\varepsilon, \alpha, \beta > 0$ with $\beta\varepsilon \geq 7\alpha$. If hypothesis quantum state σ satisfies

$$\left|\operatorname{Tr}[E_i\sigma] - \operatorname{Tr}[E_i\rho]\right| \leq \alpha$$

for all E_i then

$$\mathop{\mathbf{E}}_{E\sim\mathcal{D}}\left[(\operatorname{Tr}[E_i\sigma] - \operatorname{Tr}[E_i\rho])^2\right] \leq (1-\varepsilon)\beta^2 + \varepsilon$$

with probability at least $1-\delta$, provided that

$$m \geq O\left(\frac{1}{\beta^2\varepsilon^2}\left(\frac{n}{\beta^2\varepsilon^2}\log^2\frac{1}{\beta\varepsilon} + \log\frac{1}{\delta}\right)\right).$$

The $O(n)$ sample complexity result is a direct result of Proposition 9.11, since with infinite time, one can solve the exponentially sized SDP to find a hypothesis σ with zero training error (we note that such a σ must exist since ρ itself always suffices).

9.3 Warmup: PAC Learning Stabilizer States

[Roc18] was able to turn Proposition 9.11 into an time efficient PAC learner for stabilizer states with *Pauli measurements* by solving the consistency problem with zero error. We briefly go over this algorithm to build intuition and establish notation for proceeding chapters.

First we will define what a Pauli measurement is and what kind of information it gives.

Definition 9.12. *Let $\mathcal{C}$ be the class of all n-qubit stabilizer pure states. If $P \in \mathcal{P}_n^{\pm}$ is a Pauli operator with real phase, then the two-outcome measurement associated with P is $(I+P)/2$, and is referred to as a* PAULI MEASUREMENT.

Fact 9.13 ([Roc18] Lemma 1)**.** *Let $E^P = \frac{I^{\otimes n}+P}{2}$ be a Pauli measurement associated to a Pauli operator $P \in \mathcal{P}_n^{\pm}$ and $|\phi\rangle\langle\phi|$ be an n-qubit stabilizer state. Then $\operatorname{Tr}\left[E^P\,|\phi\rangle\langle\phi|\right]$ can only* *take on the values* $\left\{0, \frac{1}{2}, 1\right\}$*, and:*

$$\begin{cases} \mathrm{Tr}\left[E^P \left|\phi\right\rangle\!\left\langle\phi\right|\right] = 1 \textit{ iff } P \textit{ is a stabilizer of } \left|\phi\right\rangle\!\left\langle\phi\right|; \\ \mathrm{Tr}\left[E^P \left|\phi\right\rangle\!\left\langle\phi\right|\right] = 1/2 \textit{ iff neither } P \textit{ nor } -P \textit{ is a stabilizer of } \left|\phi\right\rangle\!\left\langle\phi\right|; \\ \mathrm{Tr}\left[E^P \left|\phi\right\rangle\!\left\langle\phi\right|\right] = 0 \textit{ iff } -P \textit{ is a stabilizer of } \left|\phi\right\rangle\!\left\langle\phi\right|. \end{cases}$$

Proof. This is a simple consequence of the definition of a stabilizer state, and the fact that the Weyl operators form an orthonormal basis under the Frobenius inner product. □

Fact 9.13 tells us that the labels of our training data are effectively indicators (with phase information!) for whether or not a particular member of $\mathcal{P}_n^{\pm}$ is in $\mathrm{Stab}(\left|\phi\right\rangle)$.

We now give a high-level summary of the algorithm in [Roc18]. First, given a sample $\left(E^P, \mathrm{Tr}[E^p \left|\phi\right\rangle\!\left\langle\phi\right|]\right)$, if $\mathrm{Tr}[E^p \left|\phi\right\rangle\!\left\langle\phi\right|] = \frac{1}{2}$ then we simply ignore the sample. Otherwise, put $\pm P$ in a list of members of $\mathrm{Stab}(\left|\phi\right\rangle)$ according to Fact 9.13. Once enough samples have been taken according to Proposition 9.11, use a modified form of Gaussian elimination to find independent generators of $\mathrm{Stab}(\left|\phi\right\rangle)$. Here, the modification is using Fact 2.27 to keep track of the phase bit properly. We note that this achieves zero training error due to the definition of a stabilizer group, so $\alpha = 0$ such that $\beta\varepsilon \geq 7\alpha$ is always true. We then apply Proposition 9.11 and appropriately chosen β and ε to (ε', δ') PAC learn $\left|\phi\right\rangle$ under $\mathcal{D}$.

Chapter 10

Clifford Circuits are Hard to PAC Learn

This chapter is based on [Lia23], which also appeared at QTML 2022. Some of the preliminary results were moved to Chapters 2 and 9.

10.1 Introduction

The goal of efficient learning of quantum states and the circuits that act on them, is to be able to predict the outcome of various measurements with some degree of accuracy. For example, given a quantum state ρ and a two-outcome measurement M can we predict the probability that the measurement accepts?

Naively, one can try and learn everything there is to know about the system via quantum process tomography [CN97, ABJ$^+$03], the generalization of state tomography to quantum processes. This also requires exponential time in the number of qubits. Much like states, it then becomes necessary to restrict what kind of objects we are trying to learn. For example, Low [Low09] showed how to learn an unknown Clifford circuit by understanding its action on Pauli matrices using Bell basis measurements. Lai and Cheng [LC22] built on these results in the case of actually recovering the circuit, as well removing the need for the inverse of the circuit.

A natural follow-up was whether or not Clifford circuits could be efficiently PAC learned in an analogous way to stabilizer states. Here, we are given inputs of the form $\left(\rho, \frac{I^{\otimes n}+P}{2}\right)$ for some stabilizer state ρ and Pauli matrix P, with labels $\mathrm{Tr}\left[\frac{I^{\otimes n}+P}{2} C\rho C^\dagger\right]$ corresponding to an unknown Clifford circuit C and asked to predict future labels. It is worth noting that we have slightly altered the definition of PAC learning a quantum circuit from that of [CD20] to a setting we find more comparable to Aaronson's original PAC learning result for quantum states [Aar07]. In the setting introduced by Caro and Datta, the measurements were limited to being rank 1 projectors with product structure, rather than the rank 2^{n-1} projectors we use in our proof.

When one attempts to create a PAC learning algorithm a natural first step is to try an elimination method, i.e., eliminating options that don't match the given training data and then outputting some option that does match the data well. Such algorithms are known as *proper* learning algorithms (see Section 9.1 for more details) and were the only kind of learning algorithms considered when the idea of PAC learning was first introduced by Valiant [Val84]. And while the learning theory community now considers things like improper learning algorithms, the original proper learning algorithms generally remain the most natural class of learning algorithms to consider first. We note for instance that [Roc18] is a proper learning algorithm, as well as learning algorithms for parities and other well known learning problems [Kli05]

To that extent, we show in this chapter that an efficient proper learner for Clifford circuits that achieves $1/\mathsf{poly}(n)$ error exists if and only if $\mathsf{RP} = \mathsf{NP}$, effectively ruling out "straightforward" learning algorithms for Clifford circuits. More generally, these results apply to any proper learner that achieves arbitrary error (ε, δ) with runtime $\mathsf{poly}(n, \varepsilon^{-1}, \delta^{-1})$, which is known as a *strong* learner. Furthermore, this is true even just for a learner of a subset of Clifford circuits called CNOT circuits. This subset essentially restricts to the set of Clifford circuits that map computational basis states to other computational basis states. We leave open the problem of showing $O(1)$ hardness for proper learners using complexity theoretic means, such as in [GR09].

One can also imagine that the learning algorithm has access to a quantum computer. Since there exists problems like factoring [Sho99] for which we have an efficient quantum algorithm but not an efficient classical algorithm, this learner may be able to efficiently learn more expressive concept classes. We also give results for this setting by relating NP to RQP, the quantum analogue of RP. We now informally state our main theorems regarding CNOT and Clifford circuits.

Theorem 10.1. *There exists an efficient randomized proper PAC learner for CNOT circuits if and only if $\mathsf{RP} = \mathsf{NP}$. Furthermore, an efficient quantum proper PAC learner for CNOT circuits exists if and only if $\mathsf{NP} \subseteq \mathsf{RQP}$.*

Corollary 10.2. *There exists an efficient randomized proper PAC learner for Clifford circuits if and only if $\mathsf{RP} = \mathsf{NP}$. Furthermore, an efficient quantum proper PAC learner for Clifford circuits exists if and only if $\mathsf{NP} \subseteq \mathsf{RQP}$.*

The proofs of these main results starts by realizing that finding a CNOT circuit with zero training error requires finding a full rank matrix in an affine subspace of matrices under matrix addition (so as to differentiate from a coset of a matrix group using matrix multiplication). This is known as the NonSingularity problem [BFS99] and is NP-complete. While this may seem like a backwards reduction, it turns out that the set of matrix affine subspaces used to show that NonSingularity can solve 3SAT are a subset of the ones needed to learn CNOT circuits with zero training error. Thus, there exist a set of samples such that a CNOT circuit with zero training error exists if and only if the SAT instance is satisfiable. Finding such a CNOT circuit is what is known as the search version of the *consistency problem* and in turn the decision version of the consistency problem is also NP-complete.

To show that an efficient proper learner for CNOT circuits implies $\mathsf{RP} = \mathsf{NP}$, we follow the same proof structure as similar results for NP-hardness of the consistency problem for 2-clause CNF, 3-DNF, or the intersection of two halfspaces [Blu15, BR92, Hag20]. First, let S be some sample from the decision version of the consistency problem for CNOT circuits. Using the uniform distribution over each element in S, we will sample every element of S

with high probability given enough queries. Since S contains at most a polynomial number of samples, we are able to show that an efficient learner with arbitrary $1/\mathsf{poly}(n)$ error would necessarily also solve the consistency problem with high enough probability to create a solution in RP.

Completing the proof in the other direction, if $\mathsf{RP} = \mathsf{NP}$ we utilize search-to-decision reductions for NP-complete problems to get an efficient algorithm for the search problem of minimizing training error. We can treat this search algorithm as our means of generating a hypothesis circuit C with low training error. By the generalization theorem provided by [CD20], assuming we have enough samples, this C will properly generalize and have low true error, thus completing the proof. The quantum forms of the proof essentially come for free by replacing RP with RQP everywhere and using learners capable of doing quantum computation.

10.1.1 Related Work

We emphasize that we are dealing with the problem of classically PAC learning a classical function (i.e., classical labels) derived from a quantum system. This is as opposed to quantum PAC learning of a classical function as in [Ad17, Ad18, AGY20, QAS21] where instead of a distribution over samples we receive access to copies of a quantum state. This state results in the same distribution classically when measured in the computational basis but can be measured in other basis to get different results. There is also the attempt to directly learn a quantum process with quantum labels, as in [CL21, Car21]. Here, they do not choose to measure the output state, and have samples of the form $(\rho, \mathcal{M}(\rho))$ for quantum process $\mathcal{M}$.

10.2 Preliminaries

We first define what it means to PAC learn a quantum circuit. We emphasize that this model is different from that of [CD20], in that the samples and labels are with respect to a 2-outcome measurement rather than a basis measurement. Instead, it more closely follows Definition 9.10.

Definition 10.3 (PAC-learnability of quantum circuit)**.** *Let $\mathcal{C}$ be a class of n-qubit quantum circuit, $\mathcal{F}$ the set of n-qubit quantum states, $\mathcal{E}$ be the set of 2-outcome measurements. Let $\mathcal{D}$ be a joint distribution over $\mathcal{F} \times \mathcal{E}$. We say $\mathcal{C}$ is (ε, δ) PAC-learnable with respect to $\mathcal{D}$ if there exists a learner that, given sample access to labeled examples $(\rho, E, \mathrm{Tr}[EC\rho C^\dagger])$ for $(\rho, E) \sim \mathcal{D}$ and unknown $C \in \mathcal{C}$, is able to output a hypothesis $U \in \mathcal{C}$ satisfying*

$$\mathop{\mathbb{E}}_{(\rho,E)\sim\mathcal{D}} \left[(\mathrm{Tr}[EU\rho U^\dagger] - \mathrm{Tr}[EC\rho C^\dagger])^2 \right] \le \varepsilon$$

with probability at least $1 - \delta$.

Much like with quantum states, efficient sample complexity can be achieved (see Appendix A for details), but what about time complexity? Because of the work of Rocchetto [Roc18], Clifford circuits with Pauli measurements (Definition 9.12) are a prime candidate for an efficiently PAC-learnable class of circuits. We give a very loose bound on the fat-shattering dimension of Clifford circuits that is sufficient for our purposes.

Lemma 10.4. *Let $\mathcal{C}$ be the set of Clifford circuits. For all $\eta > 0$, $fat_{\mathcal{C}}(\eta) \le O(n^2)$.*

Proof. By Proposition 2.29 there are at most $2^{O(n^2)}$ Clifford circuits. By Lemma 9.6, the fat-shattering dimension for all $\eta > 0$ is at most $\log_2\left(2^{O(n^2)}\right) = O(n^2)$. □

Because CNOT circuits are a subset of Clifford circuits, we can also upper-bound the fat-shattering dimension of CNOT circuits by $O(n^2)$.

10.2.1 Consistency Problem of Clifford Circuits

We now turn to the consistency problem. Noting that each Pauli matrix is Hermitian, a very natural way to measure a stabilizer state is in a product basis where we measure each qubit with respect to a Pauli.

Definition 10.5. *Let the problem of PAC learning Clifford circuits with respect to Pauli measurements be defined as follows. Let C be an unknown Clifford circuit and let $\mathcal{D}$ be an unknown joint distribution over both stabilizer states and Pauli measurements. Finally, let samples to C be given as*

$$\left(\rho, E, \mathrm{Tr}\left[EC\rho C^\dagger\right]\right)$$

where $\rho, E \sim \mathcal{D}$ are a stabilizer state and Pauli measurement jointly drawn from $\mathcal{D}$ and represented as classical bit strings using the stabilizer formalism. The goal is to then learn the measurements $\mathrm{Tr}\left[EC\rho C^\dagger\right]$ up to squared loss ε under the distribution $\mathcal{D}$.

We now revisit Fact 9.13 but account for an unknown Clifford circuit C that acts on our input state.

Corollary 10.6 ([Roc18] Lemma 1)**.** *Let $E^P = \frac{I^{\otimes n}+P}{2}$ be a Pauli measurement associated to a Pauli operator $P \in \mathcal{P}_n$ and $|\varphi\rangle\langle\varphi|$ be an n-qubit stabiliser state. Then $\mathrm{Tr}\left[E^P C|\varphi\rangle\langle\varphi| C^\dagger\right]$ can only take on the values $\left\{0, \frac{1}{2}, 1\right\}$, and:*

$$\begin{cases} \mathrm{Tr}\left[E^P C|\varphi\rangle\langle\varphi| C^\dagger\right] = 1 \textit{ iff } P \textit{ is a stabilizer of } C|\varphi\rangle\langle\varphi| C^\dagger; \\ \mathrm{Tr}\left[E^P C|\varphi\rangle\langle\varphi| C^\dagger\right] = 1/2 \textit{ iff neither } P \textit{ nor } -P \textit{ is a stabilizer of } C|\varphi\rangle\langle\varphi| C^\dagger; \\ \mathrm{Tr}\left[E^P C|\varphi\rangle\langle\varphi| C^\dagger\right] = 0 \textit{ iff } -P \textit{ is a stabilizer of } C|\varphi\rangle\langle\varphi| C^\dagger. \end{cases}$$

Proof. We simply use the fact that Clifford circuit C maps $|\varphi\rangle\langle\varphi|$ to some other stabilizer state $|\phi\rangle\langle\phi| = C|\varphi\rangle\langle\varphi| C^\dagger$ then apply Fact 9.13 on $|\phi\rangle\langle\phi|$. □

What information does a single sample $\left(|\varphi_i\rangle\langle\varphi_i|, E^{P_i}, \mathrm{Tr}\left[E^{P_i}C|\varphi_i\rangle\langle\varphi_i| C^\dagger\right]\right)$ tell us? Let $G_i = \mathrm{Stab}(|\varphi_i\rangle)$. From this, we can gather that if $\mathrm{Tr}\left[E^{P_i}C|\varphi_i\rangle\langle\varphi_i| C^\dagger\right] = 1$ then $C^\dagger P_i C \in$

G_i, and if $\mathrm{Tr}\big[E^{P_i}C\,|\varphi_i\rangle\!\langle\varphi_i|\,C^\dagger\big] = 0$ then $C^\dagger P_i C \in -G_i$ where $-G_i = \{-P_i : P_i \in G_i\}$. Finally, if $\mathrm{Tr}\big[E^{P_i}C\,|\varphi_i\rangle\!\langle\varphi_i|\,C^\dagger\big] = \frac{1}{2}$ then $C^\dagger P_i C$ is in the complement of $G_i \cup -G_i$.

If the measurement E^P appears multiple times across multiple samples, we can gather further information. For instance, have

$$S_P = \big\{\big(|\varphi_i\rangle\!\langle\varphi_i|\,, E^P, \mathrm{Tr}\big[E^P C\,|\varphi\rangle\!\langle\varphi|\,C^\dagger\big]\big)\big\}$$

be the set of all samples such that E^P is the measurement taken and let G_i be the stabilizer group of each stabilizer state $|\varphi_i\rangle\!\langle\varphi_i|$ that appears in S_P. Based on each label $\mathrm{Tr}\big[E^P C\,|\varphi_i\rangle\!\langle\varphi_i|\,C^\dagger\big]$, we know that $C^\dagger PC$ must lie in H_i, which is one of G_i, $-G_i$ or $G_i \cup -G_i$. We then deduce that $C^\dagger PC$ must lie in $\bigcap_i H_i$. To actually be a Clifford circuit, we must also add the constraint that $C^\dagger PC \neq I^{\otimes n}$, giving us

$$C^\dagger PC \in \left(\bigcap_i H_i\right) \setminus \{I^{\otimes n}\}.$$

The problem of finding a Clifford circuit with zero training error then reduces to the search problem of finding a set of $\alpha, \beta, \gamma, \theta, p, q$ from Eq. (2.1) representing a $C^\dagger$ that is consistent with all of these constraints while remaining symplectic according to Eq. (2.2). Let $\mathcal{C}$ be the set of Clifford circuits. We will call this problem $\textsc{CliffordSearch}(s) = \textsc{ConsistentSearch}(\mathcal{C}, s)$[1].

Due to Gottesman-Knill [Got98, AG04] showing that Clifford circuits are classically simulable, the act of verifying that we have a circuit that has zero training error is efficient, meaning that the decision version $\textsc{CliffordDecide}(s) = \textsc{ConsistentDecide}(\mathcal{C}, s)$ of the problem is in NP.

Proposition 10.7. *The decision problem,* $\textsc{CliffordDecide}(poly(n))$*, of deciding if there exists a Clifford circuit consistent with polynomially sized sample* S *is in* NP*.*

[1] We stress that via Definition 10.5, the input states are assumed to be stabilizer states.

Proof. Given the full encoding (see Section 2.6)

$$\begin{bmatrix} A & \Gamma \\ B & \Theta \\ p^T & q^T \end{bmatrix},$$

it easy to check that

$$\begin{bmatrix} A & \Gamma \\ B & \Theta \end{bmatrix}$$

forms a symplectic matrix by matrix multiplication over $\mathbb{F}_2$ with $\Lambda(n)$ (Definition 2.30). Checking that they are consistent with the samples in S can be done by iterating through S since the trace can be computed efficiently using Gottesman-Knill [AG04]. □

Knowing this, we find that CLIFFORDSEARCH(poly(n)) $\in$ FNP. This property extends to the analogous problems for CNOT circuits, CNOTSEARCH(poly(n)) and CNOTDECIDE(poly(n)), since one can also efficiently verify that $\Gamma = 0$ and $p = 0$.

10.3 Generating Samples with Certain Constraints

We will now show how we can use samples from PAC learning to generate certain kinds of constraints. It will suffice to only consider CNOT circuits with computational basis state measurements and measurements of the form $\{I, Z\}^{\otimes n}$. The net effect of this is that from a PAC learning standpoint, for unknown CNOT circuit C we only need to figure a set of $C^\dagger Z_i C$ that is consistent with the samples as described in Section 10.2. Since we will never be tested on a measurement with some component of X_i involved, this is equivalent to finding the θ_j and q_j values from Eq. (2.1) of $C^\dagger$. We will again choose to view the θ_j as the matrix Θ, such that Θ must be full rank due to Lemma 2.31.

We introduce the following notation to convert from generators of a stabilizer group to its corresponding stabilizer state.

Definition 10.8. *Given a set of abelian generators $\{P_i\}$, let*

$$\rho(P_1, P_2, \dots, P_n) = \frac{1}{2^n} \sum_{P \in \langle P_1, \cdots, P_n \rangle} P$$

be the stabilizer state that is formed from that stabilizer group.

The following observation will also notationally make the proceeding theorem states and proofs easier to follow.

Observation 10.9. *Any one-dimensional affine subspace $v + \langle w \rangle$ can be represented as $\{v, v + w\}$ and any set of two vectors/matrices $\{v, w\}$ represents the one-dimensional affine subspace $v + \langle v + w \rangle$. Thus we can freely move between the two representations.*

Lemma 10.10. *Let C be a CNOT circuit on n qubits and have $\{v, v + w\} \subset \mathbb{F}_2^n$ be a one-dimensional affine subspace of column vectors such that $v \neq w$ and $w, v \neq 0$. Given an arbitrary pauli P there exists a set of n samples that constrains $C^\dagger P C$ to only have consistent solutions lying in $\{Z^v, Z^{v+w}\}$. Furthermore these n samples can be efficiently generated.*

Proof. Let $(v, w, v_3, \cdots, v_n)$ be an arbitrary basis for $\mathbb{F}_2^n$ containing v and w. This can be found with $O(n)$ random vectors and the use of Gaussian elimination. Recalling Definition 2.13, let us start by creating the sample

$$\left(\left(\rho(Z^v, Z^w, Z^{v_3}, Z^{v_4}, \dots, Z^{v_n}), \frac{I^{\otimes n} + P}{2}\right), 1\right),$$

which limits $C^\dagger P C$ to be in $\{I, Z\}^{\otimes n}$ with positive phase. We can create the set of samples:

$$\left(\left(\rho(Z^v, Z^w, -Z^{v_3}, Z^{v_4}, \dots, Z^{v_n}), \frac{I^{\otimes n} + P}{2}\right), 1\right),$$
$$\left(\left(\rho(Z^v, Z^w, Z^{v_3}, -Z^{v_4}, \dots, Z^{v_n}), \frac{I^{\otimes n} + P}{2}\right), 1\right),$$
$$\vdots$$
$$\left(\left(\rho(Z^v, Z^w, Z^{v_3}, Z^{v_4}, \dots, -Z^{v_n}), \frac{I^{\otimes n} + P}{2}\right), 1\right).$$

By construction $C^\dagger PC$ cannot have any component of Z^{v_3} because of the first sample of this set, nor any Z^{v_i} for $i > 3$ due to the remaining samples. This leaves $C^\dagger PC$ to be one of Z^v, Z^w, or Z^{v+w} (since it cannot be identity). To remedy this, we can introduce the final sample:

$$\left(\left(\rho(-Z^v, Z^w, Z^{v_3}, Z^{v_4}, \dots Z^{v_n}, \frac{I^{\otimes n} + P}{2}\right), 0\right),$$

which then eliminates Z^w (and identity, due to the negative sign). The total number of samples is n and the whole process takes polynomial in n time to find the basis and create said samples. □

We can easily extend this to the 0-dimensional case by simply treating w as v_2, using an extra sample to remove the last dimension. More importantly, let's say we've constrained $C^\dagger Z^x C$ to lie in $\{Z^v, Z^{v+w}\}$. The effect of this on Θ is that if we sum the columns i where $x_i = 1$ then the sum must lie in $\{v, v + w\}$.

Corollary 10.11. *Let*

$$\{v, v+w\} = \left\{ \begin{bmatrix} | & | & | & | \\ v_1 & v_2 & \dots & v_k \\ | & | & | & | \end{bmatrix}, \begin{bmatrix} | & | & | & | \\ v_1 + w_1 & v_2 + w_2 & \dots & v_k + w_k \\ | & | & | & | \end{bmatrix} \right\}$$

be a one-dimensional affine subspace of $n \times k$ matrices over $\mathbb{F}_2$ such that for all i, $v_i \neq w_i$ and $v_i, w_i \neq 0$. Finally, let Θ' be an arbitrary $n \times k$ submatrix of Θ. Then there exists a set of $(2k-1)n$ samples that constrain Θ' to only have consistent solutions lying in $\{v, v+w\}$ for CNOT circuit C. Furthermore these samples can be efficiently generated.

Proof. WLOG, we will let the set of k different columns we choose for Θ' to be columns 1 through k. We will use induction on k to prove this corollary, with the base case covered by Lemma 10.10. Now let us assume that we have samples that constrain columns 2 through k to be either

$$\begin{bmatrix} | & | & | & | \\ v_2 & v_3 & \dots & v_k \\ | & | & | & | \end{bmatrix} \text{ or } \begin{bmatrix} | & | & | & | \\ v_2 + w_2 & v_2 + w_3 & \dots & v_k + w_k \\ | & | & | & | \end{bmatrix}.$$

The goal will be to generate constraints such that if column 2 is v_2 then column 1 must be v_1. Otherwise, if column 2 is $v_2 + w_2$ then column 1 is constrained to be $v_1 + w_1$.

To start us off, we can use Lemma 10.10 to constrain the sum of columns 1 and 2 to be either $v_1 + w_1$ or $v_1 + w_1 + v_2 + w_2$. If we focus on columns 1 and 2, the solutions to this constraint lie in an affine subspace defined by:

$$\begin{bmatrix} | & | \\ v_1 + v_2 + u & u \\ | & | \end{bmatrix} \text{ or } \begin{bmatrix} | & | \\ v_1 + w_1 + v_2 + w_2 + u & u \\ | & | \end{bmatrix}$$

for arbitrary vector u. We then apply Lemma 10.10 again to constrain column 1 to be either v_1 or $v_1 + w_1$. Thus the first two columns must either be

$$\begin{bmatrix} | & | \\ v_1 & v_2 \\ | & | \end{bmatrix} \text{ or } \begin{bmatrix} | & | \\ v_1 + w_1 & v_2 + w_2 \\ | & | \end{bmatrix}$$

Finally, to lie in the intersection from the inductive hypothesis, we note that if the second column is v_2 or $v_2 + w_2$ then columns 3 through k must be

$$\begin{bmatrix} | & | & | & | \\ v_3 & v_4 & \dots & v_k \\ | & | & | & | \end{bmatrix} \text{ or } \begin{bmatrix} | & | & | & | \\ v_3 + w_3 & v_4 + w_4 & \dots & v_k + w_k \\ | & | & | & | \end{bmatrix}.$$

respectively.

Collectively, we achieve our goal of constraining the entire solution to lie in $v + \langle w \rangle$. We used n samples at the first step and $2n$ for every inductive step after (one set of n samples for each call of Lemma 10.10), giving us a total number of samples of $2n(k-1) + n = 2kn - 2n + n = (2k-1)n$. Since each step was efficient, the whole process takes polynomial in n time to generate all of the samples. □

10.4 On the NP-completeness of NonSingularity

Definition 10.12. *Given $n \times n$ matrices $M_0, M_1, \dots, M_m$ over some field $\mathbb{F}$,* NonSingularity *is the problem of deciding if there exists $\alpha_1, \alpha_2, \dots, \alpha_m \in \mathbb{F}$ such that $M_0 + \sum_i \alpha_i M_i$ results in a non-singular matrix.*

Theorem 10.13 ([BFS99] Corollary 10)**.** NonSingularity *over* $\mathbb{F}_2$ *is* *NP-complete.*

The high level idea of the proof is to first reduce a 3SAT instance over variables $\{x_i\}$ to solving an arithmetic formula F. The formula is then turned into a weighted directed graph whose adjacency matrix $M(x)$ has a determinant that is equal to the formula F, where $M(x)$ has entries from $\mathbb{F}_2 \cup \{x_i\}$, and can thus be viewed as an affine subspace over $\mathbb{F}_2^{(|F|+2)\times(|F|+2)}$.

While we will not prove the correctness of this statement, we will want to ascertain exactly what kind of M_i are formed through the reduction. We now describe the construction of the graph (see Figs. 10.1 and 10.2 for relevant illustrations):

- For each atomic formula F', create vertices s and t.

  For each constant c create a unique node v_c with a path from s to v_c with weight c and a path from v_c to t with weight 1.

  For each variable x_i create a unique node v_{x_i} with a path from s to v_{x_i} with weight x_i and a path from v_{x_i} to t with weight 1.

- For multiplication of F_i and F_j, place the graphs of F_i and F_j in series.

- For addition of F_i and F_j, place the graphs of F_i and F_j in parallel.

- Once all of this is done, create a path of weight 1 from the global t vertex to the global s vertex.

- Create self loops at every vertex besides the global s vertex.

Let M be the resulting adjacency matrix of this graph. For every entry that is a constant, we can assign that to M_0. Then for each variable x_i, we can set M_i to be the

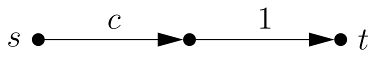
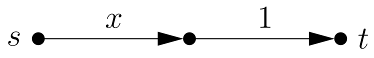
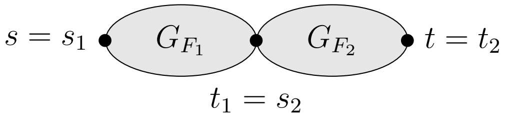
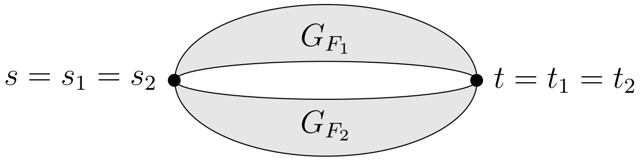
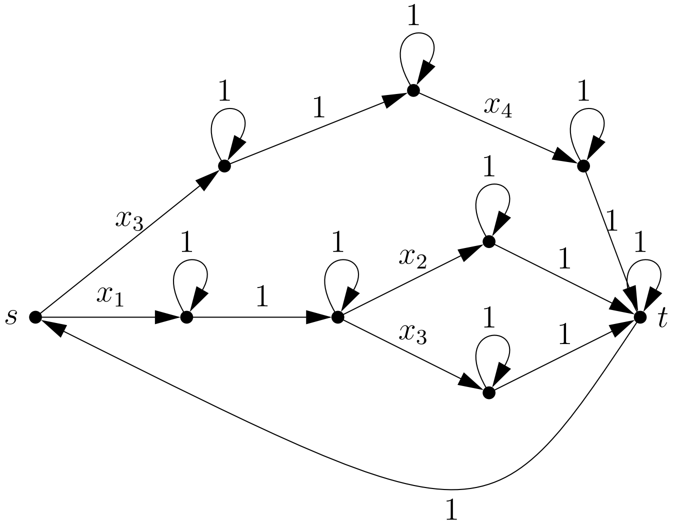
Formula F	The series-parallel s-t graph G_F with edge withs
Constant c	
Variable x	
$F = F_1 \cdot F_2$	 $s = s_1$ G_{F_1} G_{F_2} $t = t_2$ $t_1 = s_2$
$F = F_1 + F_2$	 G_{F_1} $s = s_1 = s_2$ $t = t_1 = t_2$ G_{F_2}

Figure 10.1: Inductive Construction from Formula to Graph

$$F : x_1(x_2 + x_3) + x_3 \cdot x_4$$



$$\begin{pmatrix}
0 & x_1 & 0 & 0 & 0 & x_3 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & x_2 & x_3 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & x_4 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}$$

Figure 10.2: Example of constructing the adjacency matrix with a specific determinant

matrix that is zero everywhere except where x_i appears in M. As an example, for a matrix

$$M(x) = \begin{pmatrix} x_1 & 1 \\ x_2 & x_1 \end{pmatrix}$$

we can describe it using $M(x) = M_0 + x_1 M_1 + x_2 M_2$ where

$$M_0 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \quad M_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad M_2 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$$

For a more succinct reduction later in Section 10.5, we want to isolate the kinds of matrix affine subspaces over $\mathbb{F}_2$ that are hard to solve (i.e., are used in the reduction from 3SAT). The following notation will be beneficial for that.

Definition 10.14. *For a $n \times n$ matrix M over a field $\mathbb{F}$, let $NZ(M) \subseteq \{1, 2, \cdots, n\}$ be the columns of M that are non-zero (i.e., are not the all zeros vector).*

Definition 10.15. *Let M and W be $n \times n$ matrices over a field $\mathbb{F}$ and let $k \coloneqq |NZ(W)|$ such that $NZ(W) = \{i_1, i_2, \cdots i_k\}$. If*

$$M = \begin{bmatrix} | & | & | & | \\ m_1 & m_2 & \dots & m_n \\ | & | & | & | \end{bmatrix}$$

then the restriction of the columns of M to the non-zero columns of W is defined as

$$R_W(M) = \begin{bmatrix} | & | & | & | \\ m_{i_1} & m_{i_2} & \dots & m_{i_k} \\ | & | & | & | \end{bmatrix}.$$

Likewise, for a set of matrices S, $R_W(S) = \{R_W(M) : M \in S\}$.

We now define the modified version of NonSingularity and show that it is still NP-complete.

Definition 10.16. *Let $M = M_0 + \langle M_1, M_2, \cdots M_m \rangle$ be an matrix affine subspace of $n \times n$ matrices over some field $\mathbb{F}$. In addition, require the M_i have the property that $NZ(M_i) \cap$*

$NZ(M_j) = \emptyset$ for all $i > 0$ and $j > 0$. Finally, for all $i > 0$ with $k_i := |NZ(M_i)|$, constrain M_0 and M_i such that the restriction of M to the non-zero columns of M_i,

$$R_{M_i}(M) = R_{M_i}(\{M_0, M_0 + M_i\}),$$

can be represented as:

$$\left\{ \begin{bmatrix} | & | & | & | \\ v_1 & v_2 & \dots & v_{k_i} \\ | & | & | & | \end{bmatrix}, \begin{bmatrix} | & | & | & | \\ v_1 + w_1 & v_2 + w_2 & \dots & v_{k_i} + w_{k_i} \\ | & | & | & | \end{bmatrix} \right\}$$

for some v and w such that $v_j \neq w_j$ and $v_j, w_j \neq 0$. The problem of deciding if there exists $\alpha_1, \alpha_2, \cdots, \alpha_m \in \mathbb{F}$ such that $M_0 + \sum_i \alpha_i M_i$ results in a non-singular matrix will be called the Modified-NonSingularity *problem.*

Corollary 10.17. *The* Modified-NonSingularity *problem is* **NP***-complete over $\mathbb{F}_2$.*

Proof. Rather than reduce NonSingularity to Modified-NonSingularity, we instead reduce 3SAT to Modified-NonSingularity directly by showing that the reduction from 3SAT to NonSingularity naturally leads to instances of Modified-NonSingularity.

Let G be the graph produced by the reduction from 3SAT with adjacency matrix $M(x)$. By the construction of $M(x)$ given by Fig. 10.1, we see that every instance of a variable will create its own unique subgraph such that the instance of each variable connects to a unique vertex. Because this unique vertex can never be used as an s or t vertex from Fig. 10.1, that vertex also necessarily has in-degree 1 so that no other edges connect to it. Because these are the only vertices that have incoming edges assigned with variable weight, this conversely means that the columns of $M(x)$ contain at most one variable.

Recall that for variable x_i with $i > 0$, the matrix M_i we form from the decomposition of $M(x)$ is the entries associated with x_i. Since each column only contains at most one x_i, then $NZ(M_i) \cap NZ(M_j)$ must be disjoint for $i > 0$ and $j > 0$. Due to the constraint being met, we know that the $R_{M_i}(M)$ is confined to a one-dimensional affine subspace. As such, it can be represented as:

$$\left\{ \begin{bmatrix} | & | & | & | \\ v_1 & v_2 & \dots & v_{k_i} \\ | & | & | & | \end{bmatrix}, \begin{bmatrix} | & | & | & | \\ v_1 + w_1 & v_2 + w_2 & \dots & v_{k_i} + w_{k_i} \\ | & | & | & | \end{bmatrix} \right\}$$

for some v and w, where the v is a sub-matrix of M_0 and w is the corresponding sub-matrix for M_i. Due to the definition of $NZ(M_i)$ it is clear that the w_j are non-zero. Furthermore, since the weight of an edge will never be $x_i + 1$ then an entry of w_j being one implies the corresponding entry of v_j is zero. This implies $v_j \neq w_j$ for all j. Finally, to ensure that $v_j \neq 0$, we note that each vertex besides s receives a self loop with weight 1. s instead receives an edge from t with weight 1. These self loops and the edge from t to s ensures that each column of M_0 has at least entry with 1 in it such that the $v_j \neq 0$.

We have now shown that every matrix affine subspace produced from the reduction in [BFS99] also meets the requirements for Modified-NonSingularity, thus showing that this problem is also NP-complete. □

10.5 PAC Learning CNOT Circuits and NP

Lemma 10.18. *The decision problem,* CNOTDecide$(2n^2)$, *of deciding whether or not there exists a CNOT circuit consistent with at most $2n^2$ samples is* NP*-complete.*

Proof. By Corollary 10.17 and the Cook-Levin theorem [2], if a problem can be used to solve Modified-NonSingularity then it is NP-hard. Looking at each individual M_i for $i > 0$ from Definition 10.16, they are all non-zero on disjoint columns. If we restrict to the non-zero columns of a particular M_i for $i > 0$, we find that $R_{M_i}(M)$ meet the requirements for Corollary 10.11. Since the M_i act on disjoint columns, if we apply Corollary 10.11 for each $R_{M_i}(M_i)$ then we have efficiently created a set of samples S_i that restricts those columns

[2] See [KT22] for details.

of Θ, the matrix form of the θ_j values in Eq. (2.1), to lie in $R_{M_i}(\{M_0, M_0 + M_i\})$. To fix the columns not touched by the M_i, it is not hard to show that the ideas of Lemma 10.10 can also create a 0-dimensional affine space over these columns, thereby setting the whole matrix to lie in $M_0 + \langle M_1, M_2, \cdots M_m \rangle$. Let such samples be called T. Altogether, by taking $S = T \cup (\bigcup_i S_i)$ we are able to use Lemma 10.10 and Corollary 10.11 to set Θ to lie in $M(x)$ from Modified-NonSingularity.

If $M(x)$ is an accepting instance of Modified-Singularity then there must be a full rank Θ that is consistent with S. Since Lemma 2.32 ensures us that CNOT circuits can instantiate any full rank Θ, there must also exist a CNOT circuit consistent with S. Alternatively, if $M(x)$ does not contain a non-singular matrix, then there does not exist a full rank theta that is consistent with S. By Lemma 2.31 there cannot exist a CNOT circuit that is consistent with the data. This gives us that $M(x)$ contains a non-singular matrix if and only if there is a CNOT circuit consistent S, which can be produced efficiently.

We now count the number of samples used. For each all $1 \leq i \leq m$, we use n samples to constrain the first column of $NZ(M_i)$. For the remaining columns, we either use $2n$ if that column is contained in some $NZ(M_j)$, otherwise, we use $n + 1$ samples from the generalization of Lemma 10.10. Since $2n >= n + 1$ for $n \geq 1$, we use at most $2n(n - m)$ samples for the remainder, giving us at most $2n(n - m) + nm = 2n^2 - mn \leq 2n^2$ samples. This shows CNOTDecide$(2n^2)$ is NP-hard. Combined with Proposition 10.7 we find that CNOTDecide$(2n^2)$ to be NP-complete. □

Since CNOTDecide$(2n^2) \subset$ CNOTDecide$(\mathsf{poly}(n))$ then CNOTDecide$(\mathsf{poly}(n))$ is also NP-hard.

Theorem 10.19 (Formal Statement of Theorem 10.1)**.** *There exists an efficient randomized (ε, δ) proper PAC learner for CNOT circuits with arbitrary ε and δ as arbitrary $\frac{1}{\mathsf{poly}(n)}$ values if and only if $\mathsf{RP} = \mathsf{NP}$. Furthermore, an efficient quantum (ε, δ) proper PAC learner* *for CNOT circuits with arbitrary ε and δ as arbitrary $\frac{1}{\mathsf{poly}(n)}$ values exists if and only if* $\mathsf{NP} \subseteq \mathsf{RQP}$.

Proof. We will start by proving the $\mathsf{NP} \subseteq \mathsf{RP}$ version for classical randomized learners. The quantum version will follow trivially by replacing the learner with a quantum algorithm and therefore RP with RQP. The only change then is that $\mathsf{NP} \subseteq \mathsf{RQP}$ does not necessary imply $\mathsf{NP} = \mathsf{RQP}$ like it does with RP.

Because CNOT circuits with classical inputs and measurements only has labels 0 and 1, the smallest non-zero error is 1. By Proposition 9.8 with $\alpha = 1$, an efficient

$$\left(\varepsilon < \frac{1}{4n^4}, \delta < \frac{1}{2} + \frac{1}{4n^2}\right)^{3}$$

randomized proper learner for CNOT circuits will imply $\textsc{CNOTDecide}(2n^2) \in \mathsf{RP}$. Since $\textsc{CNOTDecide}(2n^2)$ is NP-complete by Lemma 10.18, efficient randomized learners for CNOT circuits with arbitrary $\varepsilon = \frac{1}{\mathsf{poly}(n)}$ and $\delta = \frac{1}{\mathsf{poly}(n)}$ are only possible if $\mathsf{NP} \subseteq \mathsf{RP}$.

Conversely, by Lemmas 9.9 and 10.4 with $\beta = \frac{1}{2}$, $c = \mathsf{poly}(n)$, $\varepsilon = \frac{4\varepsilon'-1}{3}$, and $\delta = \delta' - 2^{-c}$ if $\mathsf{NP} \subseteq \mathsf{RP}$ then there must exist an efficient (ε', δ') proper learner for CNOT circuits as long as the number of samples m is polynomial in n. For arbitrary $\varepsilon' = \frac{1}{\mathsf{poly}(n)}$ and $\delta' = \frac{1}{\mathsf{poly}(n)}$ our required number of samples becomes

$$m = \Theta\left(\frac{1}{\varepsilon}\left(fat_{\mathcal{C}}\left(\frac{\beta}{8}\right)\log^2\left(\frac{fat_{\mathcal{C}}\left(\frac{\beta}{8}\right)}{\beta\varepsilon}\right) + \log\frac{1}{\delta}\right)\right) = O(\mathsf{poly}(n))$$

This is sufficiently small as desired, completing the proof. □

Corollary 10.20 (Formal Statement of Corollary 10.2)**.** *There exists an efficient randomized (ε, δ) proper PAC learner for Clifford circuits with arbitrary ε and δ as arbitrary $\frac{1}{\mathsf{poly}(n)}$ values if and only if* $\mathsf{RP} = \mathsf{NP}$. *Furthermore, an efficient quantum (ε, δ) proper PAC learner*

[3] We again abuse notation to signify that ε is a value less than $\frac{1}{4n^4}$ and likewise for $\delta < \frac{1}{2} + \frac{1}{4n^2}$.

for Clifford circuits with arbitrary ε and δ as arbitrary $\frac{1}{\mathsf{poly}(n)}$ values exists if and only if $\mathsf{NP} \subseteq \mathsf{RQP}$.

Proof. Since CNOT circuits are also a form of Clifford circuits, $\textsc{CNOTDecide}(2n^2) \subset \textsc{CliffordDecide}(2n^2)$, and so $\textsc{CliffordDecide}(2n^2)$ is also NP-hard. Combined with Proposition 10.7, we get that it is NP-complete. The proof of Theorem 10.19 continues but with $\alpha = \frac{1}{2}$ instead due to Corollary 10.6. This leads to slightly different constants, but the proof ideas all follow without major change. □

10.6 Special Cases with Efficient Proper Learners

Despite the results given, there still exist situations where it is possible to efficiently proper learn Clifford circuits and CNOT circuits. We give brief proof sketches of some of them here.

10.6.1 CNOT Circuits for a Distribution with Support over a Single Measurement

Let us try to learn CNOT circuits with regard to a distribution $\mathcal{D}$ such that there exists some pauli $P \in \{I, Z\}^{\otimes n}$ with $\mathbb{P}_{(\rho,E)\sim\mathcal{D}}[E = \frac{I^{\otimes n}+P}{2}] = 1$. Because we are dealing with CNOT circuit the labels will always be 0 and 1 so by Corollary 10.6 each label will tell us an affine subspace that $C^\dagger PC$ lies in. We can efficiently compute the intersection of this using Gaussian elimination with the generators to find a P' that is consistent will all of the labels. From there, let $P, Q_2, Q_3, \ldots, Q_n$ be a set of Paulis whose span is $\{I, Z\}^{\otimes n}$. Let $P', Q'_2, Q'_3, \ldots, Q'_n$ also be a set of Paulis whose span is $\{I, Z\}^{\otimes n}$. It is clear that if we define our CNOT circuit such that $C^\dagger PC = P'$ and $C^\dagger Q_i C = Q'_i$ then we have a valid CNOT circuit. Efficiently finding such $\{Q_i\}$ and $\{Q'_i\}$ only takes $O(n)$ expected samples of random Paulis in $\{I, Z\}^{\otimes n}$ and so can be done efficiently. Appealing to both Proposition 2.29 and Theorem 9.5 completes the proof.

10.6.2 Clifford circuits with the Uniform Distribution over Pauli Measurements

We note that if the distribution $\mathcal{D}$ entails the measurements being uniform over the Paulis then the problem is trivially easy to properly learn with $\varepsilon < \frac{1}{\exp(n)}$ and $\delta = 0$ by just outputting a random Clifford circuit. This is because the probability that a random Pauli is in a given state's stabilizer group is $\frac{2^n}{4^n}$ so we will almost always see the label $\frac{1}{2}$ regardless of the hypothesis circuit we choose.

10.6.3 CNOT Circuits with the Uniform Distribution over $\{I, Z\}^{\otimes n}$

Recall from Proposition 2.33, the only parts of the full encoding of a Clifford circuit that matter for a CNOT circuit are Θ and q (see Eq. (2.1) for a definition of these). We note that if we have enough independent samples that

$$\begin{bmatrix} \Theta \\ q^T \end{bmatrix}$$

is confined to a $O(\log n)$ dimensional affine subspace then we can simply iterate through all possible Θ and Q combinations to find one with a non-singular Θ in $\text{poly}(n)$ time.

Let M' be the true value of Θ. and let $M \neq M'$ be another arbitrary matrix. Likewise let x' be the true value of q^T and x some arbitrary vector. When we draw a sample, we will see both $(|v\rangle\langle v|, \frac{I^{\otimes n} + Z^w}{2})$ where $v, w \in \mathbb{F}_2^n$ are selected uniformly at random. The pair M and x will give the same label as the true label if and only if

$$\big(v^T(M + M') + x^T + (x')^T\big)w = 0 \mod 2.$$

Because w is uniform random, as long as $v^T(M + M') + x^T + (x')^T$ is not the zero vector over $\mathbb{F}_2$ then this will only be 0 at most half the time. If at least one of $M \neq M'$ or $x \neq x'$ is true, then $(M + M')^T v = x + x'$ will only be true with probability at most $\frac{1}{2}$ as well. So the probability that any arbitrary M and x have different labels is at least $\frac{1}{4}$ when they differ from the true values M' and x'. Thus with $O(n)$ expected samples we will have constrained

$$\big(v^T(M + M') + x^T + (x')^T\big)w = 0 \mod 2.$$

to be an affine subapce of $O(\log n)$. This means we can bruteforce search to find a full rank Θ and corresponding q_j values that is consistent with all samples. From there we again apply both Proposition 2.29 and Theorem 9.5 to generalize with zero training error as long as the number of samples is also at least the parameter m from the theorem statement.

10.6.4 Clifford Circuits for a Distribution with Support over a Single State

In the converse of an earlier situation, let us try to learn Clifford circuits with regard to a distribution $\mathcal{D}$ such that there exists some stabilizer state σ with $\mathbb{P}_{(\rho,E)\sim\mathcal{D}}[\rho = \sigma] = 1$. This situation effectively reduces to that of [Roc18]. If we run that algorithm we will find a state σ' that is consistent with all of the labels. Let $\{g_i\}$ be the generators of σ and $\{g_i'\}$ the generators of σ'. If we let $Cg_iC^\dagger = g_i'$ we define the first part of a Clifford circuit that maps σ to σ' as desired. We can then run the algorithm from [VDB21] to fill in the remainder of the Clifford circuit. Appealing to both Proposition 2.29 and Theorem 9.5 once again completes the proof.

10.7 Discussion and Open Problems

In this work, we prove a negative result in proper learning of one of the best candidates for efficient PAC learning of quantum circuits. However, it should be noted that in many cases there exist improper learners even in the case where proper learning is NP-hard, such as 2-clause CNF, 3-DNF, and intersection of half spaces[Blu15, BR92, Hag20]. This immediately leaves the problem of whether or not an improper learner exists for Clifford circuits. One way of showing hardness would be to leverage cryptographic hardness such as in [Kha93, AGS21]. Another approach would be to assume the hardness of random k-DNF, such as the work of Daniely [DSS16]. For upper bounds, the work of Caro and Datta [CD20] can also be used to get agnostic generalization results, providing a possible pathway to answering research questions in that direction.

Another thing worth considering is that the hardness results only apply for small errors (roughly $1/\mathsf{poly}(n)$). And while this is sufficient to give complexity-theoretic hardness for the kinds of PAC learners (i.e., strong proper learners) originally introduced by Valiant [Val84], it would be nice to get hardness results for larger errors as in the work by Guruswami and Raghavendra [GR09]. This work involved using PCP/hardness-of-approximation ideas to show that even constant training error was NP-hard.

We also note that a single output bit of a CNOT circuit is simply an XOR, which is easy to learn efficiently if there is no noise. However, because we are dealing with reversible computation each output bit has to be a linearly independent XOR such that each input bit is recoverable. Finding a CNOT circuit that matches a single XOR function f can be done by sampling expected $O(n)$ random XOR until we get n linearly independent XOR with one of them being f (see Section 10.6.1). Thus, the entire difficulty of proper learning CNOT circuits is this linear independence of the output bits. As such, even though $\mathsf{AC}^0 \subseteq \mathsf{TC}^0 \subseteq \mathsf{NC}^1 \subseteq \mathsf{L} \subseteq \oplus\mathsf{L}$ with the lower classes having improper hardness results based on cryptographic hardness [Kha93], one cannot directly give an improper learning result for CNOT circuits despite the fact that simulating CNOT circuits is complete for $\oplus\mathsf{L}$ [AG04].

As noted previously, our PAC learning framework is slightly different from that of [CD20], in that we use Pauli matrices, rather than rank 1 projectors as measurements. To the author's knowledge, there exists no proof showing that one framework is necessarily harder than the other. The author also do not see an obvious way of proving an analogous hardness theorem in the specific framework of [CD20] for Clifford or CNOT circuits.

Finally, with everything from the input states to the circuits involved being classical, it is entirely possible to prove the technical results about CNOT circuits only talking about bit strings and parity functions. Namely, one can replace the entire problem with samples of the form $(x, s, s^T C x)$ where $(x, s) \sim \mathcal{D}$ are in $\mathbb{F}_2^n$ respectively and C is a CNOT circuit. Since the stabilizer group of a computational basis state always lies in $\{\pm 1\} \times \{I, Z\}^{\otimes n}$, we can uniquely define it by the subgroup that has positive phase. This is equivalent to the orthogonal complement of x, which is the subspace $M_x = \{x \in \mathbb{F}_2^n : w \cdot x = 0\}$. From there, a sample of the form $(x, s, 0)$ simply says that $Cx \in M_s$, and one can get an analogous proof by copying the lemmas and theorems in Sections 10.3 and 10.5. However, this proof isn't anymore intuitive than the one given using stabilizer groups, and in fact is probably less intuitive to the average reader due to the lack of established formalism from stabilizers and paulis. It would be interesting if a more intuitive purely classical proof could be made to show hardness of learning CNOT circuits under this model.

Chapter 11

On the Hardness of PAC-learning Stabilizer States with Noise

This chapter is based on [GL22], which was joint work with Aravind Gollakota. Some of the preliminary results were moved to Chapters 2 and 9.

11.1 Introduction

While stabilizer states can be PAC learned (see Section 9.3), a major question left open by [Roc18] is: are stabilizer states also efficiently learnable in noisy settings? Motivated by this question, we introduce to the quantum setting a well-known tool for noise-resilient classical PAC learning, the *statistical query* (SQ) model, and define the problem of SQ-learning quantum states. In this model, rather than receiving labeled measurement-outcome examples of the form $(E, \mathrm{Tr}[E\rho])$, the learner is only allowed to make statistical queries to an oracle, and otherwise its goal remains the same. A statistical query is described by a function $\varphi : \mathcal{E} \times \{-1, 1\} \to [-1, 1]$ and a tolerance $\tau > 0$, and the oracle responds to the query with $\mathbb{E}[\varphi(E, Y)] \pm \tau$, where the expectation is taken over the random draw of $E \sim \mathcal{D}$ and Y, the random outcome of measuring ρ using E. One can think of this as modeling an experimental setup that is unable to report individual measurement outcomes, but is nevertheless able to estimate expectation values to any desired accuracy. Importantly, an algorithm that is able to work in this restricted setting automatically gains tolerance to several kinds of noise.

The SQ model was originally introduced by Kearns [Kea98] in the setting of Boolean function classes, and has since grown into a highly influential model (see [Fel16, Rey20] for surveys). The model is known to have the following properties:

- It is a natural restriction of the PAC model that nevertheless captures most known PAC algorithms for a wide range of common classes [HS07, Rey20].
- SQ algorithms are naturally resistant to mild forms of noise in the labels, such as "classification noise", where the label for each training example is flipped with some constant probability [Kea98].
- It is the most realistic learning model for which strong, unconditional lower bounds are known for many basic classes. Indeed, there is a considerable literature on this topic, with lower bounds usually proven using the so-called SQ-dimension and its generalizations [BFJ+94, Fel12, Rey20].
- SQ algorithms are naturally implementable in a way that satisfies differential privacy of the training data, and indeed are the main examples of realistic differentially private learning algorithms [BDMN05, DR14].

Given all of these properties, it is natural to wonder whether the SQ model has something to bring to quantum learnability, with a particular eye towards noise tolerance. In this work we show (among other results) that for stabilizer states this approach cannot work: SQ-learning stabilizer states is exponentially hard, and in general, learning stabilizer states with noise is as hard as the well-known Learning Parity with Noise (LPN) problem.

Theorem 11.1. *Let $\mathcal{D}$ denote the uniform distribution on Pauli measurements. Any SQ algorithm for learning n-qubit stabilizer states under $\mathcal{D}$ up to error $2^{-O(n)}$ (i.e., to significantly outperform the maximally mixed state) requires $2^{\Omega(n^2)}$ queries even when tolerance is $2^{-O(n^2)}$.*

Define a *parity measurement* to be a Pauli measurement of the form $E_x = \frac{I-P_x}{2}$ for some $x \in \mathbb{F}_2^n$, where $P_x = \sum_{y\in\mathbb{F}_2^n} \chi_x(y) |y\rangle\langle y|$ and $\chi_x(y) = (-1)^{x\cdot y}$. They are so named since for any computational basis state $|y\rangle\langle y|$, $\mathrm{Tr}[E_x\,|y\rangle\langle y|] = x\cdot y$. The following theorems hinge on the observation (stated as Proposition 11.21) that parities can be very naturally embedded within the problem of learning stabilizer states under distributions on parity measurements.

Theorem 11.2. *Let $\mathcal{D}'$ denote the uniform distribution on parity measurements. Any SQ algorithm for learning n-qubit stabilizer states under $\mathcal{D}'$ even up to constant error (say $1/3$) requires $2^{\Omega(n)}$ queries even when tolerance is $2^{-O(n)}$.*

Theorem 11.3. *Let $\mathcal{D}'$ be as above. Learning n-qubit stabilizer states under $\mathcal{D}'$ with classification noise at rate η is at least as hard as the classical problem of Learning Parity with Noise (LPN) at rate η.*

(These theorems are formally stated as Corollaries 11.18, 11.22 and 11.23 respectively.)

Our results position the problem of learning stabilizer states as a quantum analogue of the important classical problem of learning parities.[1] In both cases there are simple "algebraic" learning algorithms for the noiseless setting, and the problem seems to become intractable with even the simplest kinds of noise. The algorithm of [Roc18] thus joins a small class of PAC algorithms that do not fall into the SQ model, and hence do not admit any straightforward algorithms in noisy settings. In our view, this frames learning stabilizer states with noise as one of the more compelling problems on the frontier of learning quantum states with noise.

Another interpretation of our results is that they highlight limitations of the PAC framework of [Aar07]: insofar as this framework reduces the problem of learning quantum states to an essentially classical problem, it also inherits longstanding problems from classical learning theory. In particular, Theorem 11.3 arises from a purely classical sub-problem of stabilizer states and Pauli measurements, much like we see in Chapter 10.

[1] This recalls the way in which stabilizer codes are a quantum analogue of classical parity check codes.

We also hope that our introduction of the SQ model to quantum state learning will be of independent interest and help spur new ideas in this area.

We now detail the rest of our contributions and lay out the organization of this chapter:

- In Section 11.2, we formally define the problem of SQ-learning quantum states and extend the notion of the SQ-dimension to this setting, building on recent work that formally analyzed the SQ-dimension as applicable to the p-concept setting [GGJ+20].

- In Section 11.3, we show that SQ algorithms for learning quantum states are indeed resistant to mild forms of noise, including classical classification noise as well as quantum channels with bounded noise (such as depolarizing noise).

- In Section 11.4, we give exponential SQ lower bounds on learning stabilizer states. Under the uniform distribution on Pauli measurements, we show (Corollary 11.18) that it requires exponentially many queries in order to improve on the maximally mixed state's performance. Under a different natural distribution on Pauli measurements, namely the uniform distribution over parity measurements, we show (Corollaries 11.22, 11.23) that learning stabilizer states with noise is as hard as learning parities with noise.

- In Section 11.5, by way of positive results, we give SQ algorithms for the simple setting of learning product states. We describe an SQ algorithm for learning product states under Haar-random single-qubit measurements, and show that it allows one to perform tomography on the individual qubits.

- In Section 11.6, we relate SQ learning to a form of differential privacy for quantum state learners. This form of differential privacy has recently been studied by [QAS21].

11.1.1 Related Work

We emphasize that (just like Chapter 10) we operate in the alternative PAC framework introduced by Aaronson [Aar07] (see Chapter 9) and not the tomographical framework like in Section 3.4 and Chapter 7. In recent years, this framework has been extended to the online setting [ACH+19] as well as verified in experimental setups [RAS+19]. To our knowledge, the only known computationally efficient PAC learners for supervised learning of a commonly-considered class of states are the algorithm of Rocchetto [Roc18] for learning stabilizer states, as well as that of Yoganathan [Yog19] for other classes of states whose generating circuits can be efficiently classically simulated and inverted, including low Schmidt rank states. While the focus of this chapter is on stabilizer states, we remark that Yoganathan's algorithm for low Schmidt rank states also involves solving a system of polynomial equations in the examples, and hence would also not admit any straightforward SQ implementation. Cheng et al. [CHY15] frame the problem of PAC-learning unknown quantum measurements under a distribution of states as a dual problem to PAC-learning an unknown state, and are able to recover Aaronson's main sample complexity bound using a classical proof.

Recent work by Arunachalam et al. [QAS21] extends work by Bun et al. [BLM20] to the quantum setting, and relates differentially private (DP) learning of quantum states to one-way communication, online learning, and other models. We show in Section 11.6 that our notion of SQ learnability implies their notion of DP learnability, and hence by their results also implies finite sequential fat-shattering dimension, online learnability, and "quantum stability."

We re-emphasize that the problem of PAC-learning quantum states is very different from the problem of PAC-learning Boolean functions using quantum representations of data, as considered in a recent active line of work [Ad18, Ad17]. In particular, the model of SQ-learning that we introduce is unrelated to a recent notion of SQ-learning of *Boolean* functions using quantum representations [AGY20]. When one is given quantum samples of Boolean or integer-valued functions, there have been important results on learning in the presence of noise, showing that both Learning Parity with Noise (LPN) [CSS15] and Learning With Errors (LWE) [GKZ19] are tractable in this setting.

11.2 Preliminaries

Notation and terminology. For ease of notation, we will often refer to the identity on $2^n \times 2^n$ matrices as simply I rather than $I^{\otimes n}$. It will be clear based on context when this is happening.

For sake of establishing a nice inner product (much like viewing parities over $\{\pm 1\}$ rather than $\{0, 1\}$), we will slightly modify our definition of PAC learning quantum states such that the outcomes are $\{\pm 1\}$. Let $\mathcal{E}$ denote the space of two-outcome n-qubit measurements E (corresponding to the POVM $\{E, I - E\}$), which accept a state ρ with probability $\mathrm{Tr}[E\rho]$. If we view the measurement outcomes as $\{-1, 1\}$-valued, the outcome of measuring ρ using E is a random variable Y that is 1 with probability $\mathrm{Tr}[E\rho]$ and -1 otherwise. Define $f_\rho : \mathcal{E} \to [-1, 1]$ to be the conditional mean function

$$f_\rho(E) = \mathbb{E}[Y|E] = 1 \cdot \mathrm{Tr}[E\rho] + (-1) \cdot (1 - \mathrm{Tr}[E\rho]) = 2\mathrm{Tr}[E\rho] - 1.$$

We will often identify a state ρ with its behavior with respect to two-outcome measurements, namely with the function f_ρ, and use the notation $Y \sim f_\rho(E)$ to mean that $Y \in \{-1, 1\}$ is the random measurement outcome satisfying $\mathbb{E}[Y|E] = f_\rho(E)$. In learning theoretic terms, this means f_ρ describes a *probabilistic concept*, or p-concept, on the space $\mathcal{E}$. A p-concept on a domain $\mathcal{X}$ is a classification rule that assigns random $\{-1, 1\}$-valued labels to each point in $\mathcal{X}$ according to a fixed conditional mean function; we always identify the p-concept with its conditional mean function. Given a set $\mathcal{F}$ of quantum states, we use $\mathcal{F}$ to also mean the class of associated p-concepts, with the meaning clear from context.

Given a distribution $\mathcal{D}$ over $\mathcal{E}$, we will often regard functions $f_\rho, f_\sigma : \mathcal{E} \to \mathbb{R}$ as members of the L^2 space $L^2(\mathcal{D}, \mathcal{E})$, with the inner product given by $\langle f_\rho, f_\sigma \rangle_{\mathcal{D}} = \mathbb{E}_{E\sim\mathcal{D}}[f_\rho(E) f_\sigma(E)]$, and the norm given by $\|f_\rho\|_{\mathcal{D}} = \sqrt{\langle f_\rho, f_\rho \rangle_{\mathcal{D}}} = \sqrt{\mathbb{E}_{E\sim\mathcal{D}}[(f_\rho(E))^2]}$.

11.2.1 Learning Models

Recall the problem of PAC-learning a quantum state from Section 9.2. We will now modify it to account for probabilistic concepts. The essence of the change is that the labels are now $\{\pm 1\}$ and drawn according the probability distribution defined by expectation of the observable, rather than directly being given the expectation as a value in $[-1, 1]$.

Definition 11.4 (PAC-learnability of quantum states, [Aar07])**.** *Let $\mathcal{F}$ be a class of n-qubit quantum states. Let $\mathcal{D}$ be a distribution over $\mathcal{E}$. We say $\mathcal{F}$ is PAC-learnable up to squared loss ε with respect to $\mathcal{D}$ if there exists a learner that, given sample access to labeled examples (E, Y) for $E \sim \mathcal{D}, Y \sim f_\rho(E)$ for an unknown $\rho \in \mathcal{F}$, is able to output a state σ satisfying*

$$\mathop{\mathbb{E}}_{E\sim\mathcal{D}} \left[(f_\sigma(E) - f_\rho(E))^2\right] \leq \varepsilon.$$

The number of examples used by the learner is called its sample complexity.

Again, one may speak of both computational efficiency (overall running time) and statistical or information-theoretic efficiency (sample complexity). To reiterate, an *efficient* PAC learner is one that is computationally efficient, i.e., runs in polynomial time, and hence also draws at most polynomially many examples (each draw is considered as taking one unit of time). In addition to Proposition 9.11, [Aar07] also described a computationally inefficient algorithm for probabilistic concepts via a generalization theorem.

We now introduce the following natural extension of these definitions to the SQ setting. In both cases, we operate in the so-called distribution-specific setting, where the learner is assumed to have knowledge of the distribution $\mathcal{D}$.

Definition 11.5 (SQ-learnability of quantum states)**.** *Let $\mathcal{F}$ be a class of n-qubit quantum states. Let $\mathcal{D}$ be a distribution over $\mathcal{E}$. An SQ oracle for an unknown state $\rho \in \mathcal{F}$ is an* *oracle that accepts a query and a tolerance, (φ, τ), where $\varphi : \mathcal{E} \times \{-1,1\} \to [-1,1]$ and $\tau > 0$, and responds with y such that*

$$\left| y - \mathop{\mathbb{E}}_{E \sim \mathcal{D}, Y \sim f_\rho(E)} [\varphi(E,Y)] \right| \leq \tau.$$

We say $\mathcal{F}$ is SQ-learnable up to squared loss ε if there is a learner that, given only queries to the SQ oracle for an unknown $\rho \in \mathcal{F}$, is able to output a state σ satisfying

$$\mathop{\mathbb{E}}_{E \sim \mathcal{D}} \left[(f_\sigma(E) - f_\rho(E))^2 \right] \leq \varepsilon.$$

The number of queries used by the learner is called its query complexity.

An SQ learner is considered *efficient* if it uses polynomially many queries and its queries all have tolerance $\tau \geq 1/\mathsf{poly}(n)$.

11.2.2 SQ Lower Bounds for Probabilistic Concepts

One of the chief features of the classical SQ model is the possibility of proving unconditional lower bounds on learning a class $\mathcal{C}$ in terms of its so-called statistical dimension. The quantum setting that we work in, where we identify a state ρ with the p-concept f_ρ, becomes a special case of the SQ model for learning p-concepts. Building on recent work [GGJ+20] that formally proved SQ lower bounds for p-concepts, we extend this framework to the quantum setting. Let $\mathcal{X}$ denote an arbitrary domain (for us, $\mathcal{X}$ will be $\mathcal{E}$, while in the classical setting, $\mathcal{X}$ is usually $\mathbb{R}^n$).

Definition 11.6 (Statistical dimension)**.** *Let $\mathcal{D}$ be a distribution on $\mathcal{X}$, and let $\mathcal{C}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}$. The* average (un-normalized) correlation *of $\mathcal{C}$ is defined to be $\rho_\mathcal{D}(\mathcal{C}) = \frac{1}{|\mathcal{C}|^2} \sum_{c,c' \in \mathcal{C}} |\langle c, c' \rangle_\mathcal{D}|$. The* statistical dimension on average *at threshold γ, $\mathrm{SDA}_c\, alD(\mathcal{C}, \gamma)$, is the largest d such that for all $\mathcal{C}' \subseteq \mathcal{C}$ with $|\mathcal{C}'| \geq |\mathcal{C}|/d$, $\rho_\mathcal{D}(\mathcal{C}') \leq \gamma$.*

Theorem 11.7 ([GGJ+20], Cor. 4.6)**.** *Let $\mathcal{D}$ be a distribution on $\mathcal{X}$, and let $\mathcal{C}$ be a p-concept class on $\mathcal{X}$. Say our queries are of tolerance τ, the final desired squared loss is ε, and that* *the functions in $\mathcal{C}$ satisfy $\|c\|_{\mathcal{D}} \geq \beta$ for all $c \in \mathcal{C}$. For technical reasons, we require $\tau \leq \varepsilon$, $\varepsilon^2 \leq \beta/3$. Then learning $\mathcal{C}$ up to squared loss ε (we may pick ε as large as $\sqrt{\beta/3}$) requires at least $\mathrm{SDA}_{\mathcal{D}}(\mathcal{C}, \tau^2)$ queries of tolerance τ.*

We remark that the way to interpret such a lower bound is as follows: if the SQ learner's queries have tolerance at least τ, then at least $\mathrm{SDA}_{\mathcal{D}}(\mathcal{C}, \tau^2)$ queries are required. That is, one must *either* use small tolerance *or* many queries.

The following lemma will be convenient in order to bound the SDA when we have bounds on pairwise correlations.

Lemma 11.8 ([GGJ+20], Lemma 2.6)**.** *Let $\mathcal{D}$ be a distribution on $\mathcal{X}$, and let $\mathcal{C}$ be a p-concept class on $\mathcal{X}$ such that for all $c, c' \in \mathcal{C}$ with $c \neq c'$, $|\langle c, c' \rangle_{\mathcal{D}}| \leq \gamma$, and for all $c \in \mathcal{C}$, $\|c\|_{\mathcal{D}}^2 \leq \kappa$. Then for any $\gamma' > 0$, $\mathrm{SDA}(\mathcal{C}, \gamma + \gamma') \geq |\mathcal{C}| \frac{\gamma'}{\kappa - \gamma}$.*

11.2.3 The Problem of Learning Parities

One of the most basic problems in classical learning theory is that of learning the concept class of parity functions. To review [O'D14], let the domain be $\mathbb{F}_2^n$, and for any subset $s \in \mathbb{F}_2^n$, define $\chi_s(x) = (-1)^{x \cdot s}$ to be the parity on s. Since the output is $\{\pm 1\}$, we can recover the traditional form of $\{0, 1\}$ parity by $\frac{1 - \chi_s(x)}{2} = s \cdot x$ such that $1 = (-1)^0$ maps to 0 and $-1 = (-1)^1$ maps to 1. Let $\mathcal{D}$ be any distribution on $\mathbb{F}_2^n$. We say a learner is able to learn parities under $\mathcal{D}$ if given access to labeled examples $(x, s \cdot x)$ where $x \sim \mathcal{D}$ and $s \in \mathbb{F}_2^n$ is unknown (or, in the SQ setting, given access to the corresponding SQ oracle), and for any error parameter ε, it is able to output a function h such that $\mathbb{P}_{x \sim \mathcal{D}}[h(x) \neq \chi_s(x)] \leq \varepsilon$.

The problem of learning parities displays a striking phase transition in going from the noiseless to the noisy setting. Given noiseless labeled examples, the problem of recovering the right parity is simply a question of solving linear equations over $\mathbb{F}_2$, and can be done using Gaussian elimination by a PAC learner using only $\Theta(n)$ examples. With just a little noise, however, the problem seems to become intractable. Perhaps the simplest noise model one can consider is the classification noise model, where every example has its label flipped with some constant probability η (known as the noise rate). Learning parities under classification noise is the basis of the famous *Learning Parity with Noise (LPN)* problem. Formally, the search version of LPN with noise rate η is precisely the problem of learning parities under the uniform distribution on $\mathbb{F}_2^n$ and with classification noise at rate η. Usually one also has the additional knowledge that the true target χ_s (the "secret") is picked uniformly at random from the set of all parities. This problem is widely conjectured to be hard, including for quantum algorithms, and is even used as a basis for cryptography (see [Pie12] for a survey). The best-known algorithms in the PAC setting runs in slightly subexponential time [BKW03, Lyu05].

Since SQ learners are naturally tolerant of classification noise, one would expect that there are no SQ learners for parities under the uniform distribution, and indeed, this is one of the foundational results in the SQ literature.

Theorem 11.9 ([Kea98])**.** *Any SQ learner requires $2^{\Omega(n)}$ queries (even using tolerance $2^{-O(n)}$) to learn parities under the uniform distribution on $\mathbb{F}_2^n$ even up to constant error (say $1/3$).*

Thus we see that simple Gaussian elimination is an example of an efficient PAC learner that is not SQ. This establishes a characteristic limitation of SQ algorithms: while they include a wide range of common algorithms, they do not include algorithms that depend entirely on "algebraic" structure.

It is worth emphasizing that this discussion has considered learning parities with a *classical* representation of the data. When given a *quantum* representation of the data, as in the quantum "example state" $|\psi\rangle = 2^{-n/2}\sum_{x\in\mathbb{F}_2^n}|x, s\cdot x\rangle$ (taking the distribution over the domain to be uniform), the task becomes easy even with noise [CSS15]. This is because we can now use Hadamard gates to implement a Boolean Fourier transform à la the famous Bernstein–Vazirani algorithm [BV97].

11.3 Noise-Tolerant SQ Learning

One of the prime features of classical SQ learning is its inherent noise tolerance. From an intuitive standpoint, certain common stochastic noise models are systematic enough that their effects *in expectation* can be predicted in advance, and hence either be corrected for or bounded. Slightly more precisely, the query expectations of a noisy state are often related in simple ways to the query expectations on a noiseless state, so that the latter can be recovered from the former. We mainly consider three such noise models here: (a) classical classification noise and malicious noise, (b) quantum depolarizing noise, and (c) more general quantum channels with bounded noise.

11.3.1 Classification and Malicious Noise

Classification noise [AL88] and malicious noise [Val85, KL93] are two classical Boolean noise models that SQ algorithms are able to handle. In the classification noise model, every example's label is flipped with probability η (known as the noise rate). The malicious noise model is a stronger form of noise where for any given example, with probability $1 - \eta$, the label is reported correctly, but with probability η both the point and its label may be arbitrary (and adversarially selected based on the learner's behavior so far). We note that these models are well-defined even in the p-concept setting and hence for quantum states, and simply introduce further randomness into the label. The following results were originally stated for Boolean functions but readily extend to p-concepts.

Theorem 11.10 ([Kea98])**.** *Let $\mathcal{C}$ be a p-concept class learnable under distribution $\mathcal{D}$ in the SQ model up to error ε using q queries of tolerance τ. Then for any constant $0 < \eta < 1/2$, even with respect to an SQ oracle with classification noise at rate η (i.e., one that computes* *expectations with classification noise), $\mathcal{C}$ is learnable up to ε using $O(q)$ queries of tolerance $O(\tau(1-2\eta))$. If the learner is given noisy training examples as opposed to access to a noisy SQ oracle, then $\tilde{O}(\frac{q}{\mathsf{poly}(\tau(1-2\eta))})$ noisy examples suffice.*

Theorem 11.11 ([AD98]). *Let $\mathcal{C}$ be a p-concept class learnable under distribution $\mathcal{D}$ in the SQ model up to error ε using q queries of tolerance τ. An SQ oracle with malicious noise at rate η is one that computes query expectations with respect to a distribution $(1-\eta)f(\mathcal{D})+\eta Q$, where $f(\mathcal{D})$ denotes the true labeled distribution (x, y) for $x \sim \mathcal{D}, y \sim f(x)$ (f being the unknown target p-concept), and Q is an arbitrary and adversarially selected distribution on $\mathcal{X} \times \{-1, 1\}$. If $\eta = \tilde{O}(\varepsilon)$ and $\eta < \tau$, then even with respect to an SQ oracle with malicious noise at rate η, $\mathcal{C}$ is learnable up to ε using $O(q)$ queries of tolerance $\tau - \eta$. If the learner is given noisy training examples as opposed to access to a noisy SQ oracle, then $\mathcal{C}$ is learnable (with constant probability) using $\tilde{O}(\frac{q}{\mathsf{poly}(\tau-\eta)})$ noisy examples suffice. (More efficient implementations are also available in some special cases).*

The proofs of both theorems are similar: one first relates the noisy query expectations to the true expectations, and then argues that when using a suitably small tolerance (or sufficiently many examples) the effects of the noise can be corrected for (within information theoretic limits).

11.3.2 Depolarizing Noise

Depolarizing noise acts on quantum states by shifting them closer to the maximally mixed state. One can consider a setting where it acts on an entire n-qubit state at once, as well as one where it acts independently on each individual qubit. We will consider the former.

Definition 11.12 (Depolarizing noise). *Let ρ be an arbitrary n-qubit state. Then depolarizing noise at rate η ($0 < \eta < 1$) acts on this state by transforming it into $\Lambda_\eta(\rho) = (1-\eta)\rho + \eta(I/2^n)$.*

Theorem 11.13. *Let $0 < \eta < 1$ be any constant, and let Λ_η denote the depolarizing channel at noise rate η. Let $\mathcal{C}$ be a class of n-qubit quantum states and $\mathcal{D}$ be a distribution on $\mathcal{E}$, the space of two-outcome measurements on such states. Let L be an SQ learner capable of learning $\mathcal{C}$ under $\mathcal{D}$ using q queries of tolerance τ. There there exists a learner L' such that for any $\rho \in \mathcal{C}$, L' is capable of learning ρ under $\mathcal{D}$ using q queries of tolerance $\tau(1-\eta)$ given only SQ access to $\Lambda_\eta(\rho)$ as well as sampling access to $\mathcal{D}$.*

Proof. For simplicity, we will assume that we know the noise rate η exactly. (So long as we have an upper bound on η, then by a standard "grid search" argument due to [Kea98], we can estimate η sufficiently closely simply by trying out many different values. Briefly: if say we try out $\eta = 0, \delta, 2\delta, \ldots, 1$ ($1/\delta$ values in all), then one of these will be within $\delta/2$ of the true η. The algorithm when run with this guess for η will produce a good hypothesis. By taking $\delta = O(\tau(1-\eta)^2)$ and testing all $1/\delta$ hypotheses produced by our guesses for η on a sufficiently large validation set, we can ensure the best one will perform and generalize well.)

Let $\rho \in \mathcal{C}$ be the unknown target. Observe that for any $E \in \mathcal{E}$, by linearity,

$$f_{\Lambda_\eta(\rho)}(E) = 2\mathrm{Tr}[E \cdot \Lambda_\eta(\rho)] - 1 = (1-\eta)f_\rho(E) + \eta f_{I/2^n}(E).$$

Let $\varphi : \mathcal{E} \times \{-1,1\} \to [-1,1]$ be any query made by L. Let $\varphi[\rho]$ denote the query expectation of φ under ρ, given by $\mathbb{E}_{x\sim\mathcal{D}}\,\mathbb{E}_{y\sim f_\rho(x)}[\varphi(x,y)]$. Similarly let the noisy analogue be $\varphi[\Lambda_\eta(\rho)]$. Again just by linearity,

$$\varphi[\Lambda_\eta(\rho)] = (1-\eta)\varphi[\rho] + \eta\varphi[I/2^n].$$

The latter quantity is independent of ρ and can be estimated to arbitrary accuracy by sampling from $\mathcal{D}$, allowing us to estimate $\varphi[\rho]$ as $\frac{\varphi[\Lambda_\eta(\rho)] - \eta\varphi[I/2^n]}{1-\eta}$. So long as η is bounded away from 1, we can use a query of tolerance $\tau(1-\eta)$ to estimate $\varphi[\mathcal{D}_\eta(\rho)]$ (as well as $1/\mathsf{poly}(\tau(1-\eta))$ unlabeled examples from $\mathcal{D}$ to compute $\varphi[I/2^n]$), and thereby estimate $\varphi[\rho]$ to within τ. Thus we can simulate L even with depolarizing noise. □

It is worth stressing that we are able to handle any constant noise rate $\eta \in (0, 1)$, and the price we pay is requiring the tolerance to scale as $\tau(1 - \eta)$.

11.3.3 Quantum Channels with Bounded Noise

We can also consider more general kinds of quantum channels with bounded noise. As long as the queries are bounded, small amounts of noise cannot alter query expectations too much, and so can be "absorbed" into the tolerance. This is similar to classical malicious noise: since classical malicious noise at rate η only can only change query expectations by η (recall that the queries are bounded by 1), a noisy query of tolerance $\tau - \eta$ is able to simulate a noiseless query of tolerance τ. Unlike with depolarizing noise, this means we cannot handle arbitrary η; this is an artifact of the fact that more general kinds of noise do not permit the kind of systematic correction we were able to perform for depolarizing noise.

For concreteness here we consider a noisy quantum channel Λ such that $\|\Lambda - 1_n\|_\diamond \leq \eta$, where 1_n is the identity map on n-qubit states and the norm is the diamond norm. We do not define this norm here, but its chief property for our purposes is that for any n-qubit state ρ and 2-outcome measurement E, $|\mathrm{Tr}[E(\rho - \Lambda(\rho))]| \leq \eta$. Similar theorems can be proven with respect to other distance measures such as fidelity.

Theorem 11.14. *Let Λ be a quantum channel such that $\|\Lambda - 1_n\|_\diamond \leq \eta$, as above. Let $\mathcal{C}$ be a class of n-qubit quantum states learnable under distribution $\mathcal{D}$ using q queries of tolerance $\tau > 2\eta$. Then $\mathcal{C}$ is still learnable under noise Λ (i.e., when our queries are answered not with respect to ρ but $\Lambda(\rho)$) using q noisy queries of tolerance $\tau - 2\eta$.*

Proof. As noted, for any state ρ and measurement E, $|\mathrm{Tr}[E(\rho - \Lambda(\rho))]| \leq \eta$. Consider any query $\varphi : \mathcal{E} \times \{-1, 1\} \to [-1, 1]$. If $\varphi[\rho]$ denotes the query expectation on a noiseless state and $\varphi[\Lambda(\rho)]$ denotes the noisy one, then a straightforward calculation shows that

$$\begin{aligned}|\varphi[\rho] - \varphi[\Lambda(\rho)]| &= \left| \mathop{\mathbb{E}}_{E\sim\mathcal{D}} \left[(\varphi(E,1) - \varphi(E,-1))\mathrm{Tr}[E(\rho - \Lambda(\rho))]\right] \right| \\ &\leq 2 \mathop{\mathbb{E}}_{E\sim\mathcal{D}} |\mathrm{Tr}[E(\rho - \Lambda(\rho))]| \\ &\leq 2\eta.\end{aligned}$$

Thus just by the triangle inequality, if we calculate $\varphi[\Lambda(\rho)]$ within tolerance $\tau - 2\eta$, then we also get $\varphi[\rho]$ within τ.

□

11.3.4 General Noise for Distribution-Free Learning

So far, we've only considered distribution-specific learning, where the learner is only required to succeed with respect to a pre-specified distribution $\mathcal{D}$. In the distribution-free case, where the learner is required to succeed no matter what $\mathcal{D}$is, we now give a simple proof that any SQ algorithm for a concept class can also handle any kind of quantum noise on the state, as long as the noise is known. This is unsurprising, and at a high level, the approach simply boils down to off-loading the noise from the state to the measurement. Learning a noisy set of measurements is thus handled by distribution-free learning algorithm.

Given a quantum operation Λ, its adjoint $\Lambda^\dagger$ is such that $\forall\rho$, $\mathrm{Tr}[E\cdot\Lambda(\rho)] = \mathrm{Tr}[\Lambda^\dagger(E)\cdot\rho]$ and always exists (see [RLCK19] for details on how to prove this folklore result). Let $\mathcal{D}$ be the distribution we are trying to learn concept class $\mathcal{C}$ using statistical queries and let Λ be the noise applied to the quantum state. We can then define $\mathcal{D}^\dagger$ to be the distribution $\Lambda^\dagger(E)$ where E is drawn from $\mathcal{D}$ and by definition the traces (and thus the statistical queries) are the same when applied to ρ and $\Lambda(\rho)$ respectively. Also by definition, a distribution-free learner for $\mathcal{C}$ would also be able to learn with distribution $\mathcal{D}^\dagger$.

11.4 Lower Bounds on Learning Stabilizer States with Noise

In this section, we focus on the question of: Can stabilizer states be PAC learned in the presence of noise? We first use the ideas of statistical dimension to show that, in relevant error regimes, beating the trivial hypothesis (i.e., the maximally mixed state) is very difficult over the uniform distribution of Pauli measurements. We then look at a more contrived distribution, and reduce it to the problem of parities. We we know parities to be impossible to SQ learn efficiently. More generally, we show that the problem of PAC learning stabilizers states with noise, under this distribution, is equivalent to the problem of LPN.

11.4.1 Difficulty of Beating the Maximally Mixed State on Uniform Pauli Measurements

We will first examine the natural distribution $\mathcal{D}$ given by the uniform distribution over Pauli measurements (Definition 9.12). In doing so, we will show that performing better than the trivial algorithm of always outputting the maximally mixed state $I/2^n$ is difficult.

Recall Fact 9.13. For stabilizer state $|\phi\rangle$, simple algebraic manipulations tells us that $f_\rho(E^P) = \text{Tr}[P\,|\phi\rangle\langle\phi|]$ can only take on the values $\{1, 0, -1\}$ with $\text{Tr}[P\,|\phi\rangle\langle\phi|]$ being 1 or -1 if and only if P or $-P$ is in the stabilizer group of $|\phi\rangle\langle\phi|$ respectively.

For compactness of notation, for a quantum state $|\psi\rangle$, let $\psi = |\psi\rangle\langle\psi|$. We now want to show that for stabilizer states $|\phi\rangle$ and $|\varphi\rangle$, the inner product between f_ϕ and f_φ isn't too big in absolute value. First, we will need need to show that two different stabilizer states cannot share more than half of their stabilizers[2].

Proposition 11.15. *Given two n-qubit stabilizer states $|\phi\rangle\langle\phi| \neq |\varphi\rangle\langle\varphi|$ with stabilizer groups $S = \mathit{Stab}(|\phi\rangle)$ and $S' = \mathit{Stab}(|\varphi\rangle)$ respectively, then $|S \cap S'| \leq 2^{n-1}$.*

[2] Note the difference from the unsigned stabilizer group. $|0^n\rangle$ and $|1^n\rangle$ share the same unsigned stabilizer group, but only half of their phases are aligned the same way.

Proof. Because $|\phi\rangle\langle\phi| \neq |\varphi\rangle\langle\varphi|$ then $S \neq S'$. We also know that $S \cap S'$ is an abelian group without $-I$, so $|S \cap S'| < 2^n$. Since $2^n/|S \cap S'|$ is the dimension of the space stabilized by this group [Ham89], it must be an integer. Due to the prime factorization of 2^n, $|S \cap S'| = 2^m$ for some integer $0 \leq m < n$, of which the largest possible m is $n-1$. □

With this result, we can compute bounds on $|\langle f_\phi, f_\varphi \rangle_{\mathcal{D}}|$ for stabilizer states $|\phi\rangle$ and $|\varphi\rangle$. By the nature of the uniform distribution over finite domains, this reduces to a counting problem that we solved with Proposition 11.15.

Lemma 11.16. *Let $\mathcal{C}$ be the concept class of n-qubit stabilizer pure states, and let $\mathcal{D}$ denote the uniform distribution on n-qubit Pauli measurements. Then for any stabilizer states ρ, ρ' with $\rho \neq \rho'$, $|\langle f_\rho, f_\rho \rangle_{\mathcal{D}}| = \|f_\rho\|_{\mathcal{D}}^2 = \frac{1}{2^n}$, and $|\langle f_\rho, f_{\rho'} \rangle_{\mathcal{D}}| \leq \frac{1}{2^{n+1}}$. Furthermore, this inequality is tight.*

Proof. Let $|\phi\rangle\langle\phi|, |\varphi\rangle\langle\varphi| \in \mathcal{C}$. Let $S = \mathrm{Stab}(|\psi\rangle)$ and $S' = \mathrm{Stab}(|\varphi\rangle)$, and also let $-S = \{-P : P \in S\}$ and $-S' = \{-P : P \in S'\}$. The correlation of the two p-concepts is:

$$\begin{aligned}
|\langle f_\rho, f_{\rho'} \rangle_{\mathcal{D}}| &= \frac{1}{|\mathcal{P}_n|} \left| \sum_{P \in \mathcal{P}_n} \mathrm{Tr}[P\, |\phi\rangle\langle\phi|] \cdot \mathrm{Tr}[P\, |\varphi\rangle\langle\varphi|] \right| \\
&= \frac{1}{2 \cdot 4^n} \left(|S \cap S'| + |-S \cap -S'| - |S \cap -S'| - |-S \cap S'| \right) \\
&= \frac{1}{4^n} \left(|S \cap S'| - |S \cap (-S')| \right)
\end{aligned}$$

If $|\phi\rangle\langle\phi| = |\varphi\rangle\langle\varphi|$ then $S = S'$ such that $|S \cap S'| = 2^n$ and $|S \cap -S'| = 0$. Thus for all $|\phi\rangle\langle\phi|$

$$|\langle f_\phi, f_\phi \rangle_{\mathcal{D}}| = \frac{2^n}{4^n} = \frac{1}{2^n}$$

If $|\phi\rangle\langle\phi| \neq |\varphi\rangle\langle\varphi|$ then by Proposition 11.15 $|S \cap S'| \leq 2^{n-1}$. $|S \cap -S'| \geq 0$ trivially, so we get an upper bound of

$$|\langle f_\phi, f_\varphi \rangle_{\mathcal{D}}| \leq \frac{1}{4^n}(2^{n-1}) = \frac{1}{2^{n+1}}.$$

We can show that this inequality is tight because the state $|\phi\rangle\langle\phi| = |0\rangle\langle 0|^{\otimes n}$ and $|\varphi\rangle\langle\varphi| = |0\rangle\langle 0|^{\otimes n-1} \otimes |+\rangle\langle +|$ saturate this inequality. The generators for the stabilizer group of $|\phi\rangle\langle\phi|$ are (omitting tensor products): $ZIII\cdots I$, $IZIII\cdots I$, ... , and $IIII\cdots IZ$. The generator of $|\varphi\rangle\langle\varphi|$ are the same, except the last generator is replaced with $IIII\cdots IX$. We see that $|S \cap S'| = 2^{n-1}$ while $|S \cap -S'| = 0$.

□

With this result, we can use Lemma 11.8 to compute the SDA and by extension prove a lower bound on the number of statistical queries needed to learn this concept class under this distribution.

Theorem 11.17. *Let $\mathcal{D}$ be the uniform distribution over n-qubit Pauli measurements and let $\mathcal{C}$ be the concept class of all n-qubit stabilizer pure states. Then* $\mathrm{SDA}(\mathcal{C}, \frac{1}{2^n}) = 2^{\Theta(n^2)}$.

Proof. By Proposition 2.20, $|\mathcal{C}| = 2^{\Theta(n^2)}$. Using Lemma 11.8 with $\kappa = \frac{1}{2^n}$ and $\gamma = \frac{1}{2^{n+1}}$ as calculated from Lemma 11.16, we find that

$$\mathrm{SDA}(\mathcal{C}, \gamma' + \frac{1}{2^{n+1}}) = \mathrm{SDA}(\mathcal{C}, \gamma' + \gamma) \geq |\mathcal{C}| \frac{\gamma'}{\beta - \gamma} = 2^{\Theta(n^2)} \gamma' 2^{n+1} = 2^{\Theta(n^2)} \gamma'$$

Setting $\gamma' = \frac{1}{2^{n+1}}$ gives the result. □

Corollary 11.18. *Any SQ algorithm needs at least $2^{\Omega(n^2)}$ statistical queries of tolerance $\tau = 2^{-O(n)}$ to learn $\mathcal{C}$ up to error $2^{-O(n)}$ over $\mathcal{D}$.*

Proof. Simply apply Theorem 11.7, with $\beta = 2^{-n}$. □

Since the norms of our p-concepts are exponentially small (i.e., $2^{-n/2}$), we only get hardness for error on the order of $2^{-O(n)}$. But as we now show, the p-concept norm corresponds almost exactly to the squared loss achieved by the maximally mixed state. Our results show that doing significantly better than the maximally mixed state requires $2^{\Omega(n^2)}$ statistical queries even when the tolerance is exponentially small.

Proposition 11.19. *Let $\mathcal{D}$ be the uniform distribution over n-qubit Pauli measurements, $\mathcal{E}_{Pauli}$. Let ρ be any state, and let $I/2^n$ be the maximally mixed state. Then*

$$\|f_\rho\|_{\mathcal{D}}^2 = \|f_\rho - f_{I/2^n}\|_{\mathcal{D}}^2 + \frac{1}{4^n}.$$

Proof. In essence, this is simply because the p-concept $f_{I/2^n}$ is almost always zero. Specifically, for all $E \in \mathcal{E}_{\text{Pauli}} \setminus \{0, I\}$, $f_{I/2^n}(E) = 2\text{Tr}[E/2^n] - 1 = 0$, since $\text{Tr}[E] = 2^{n-1}$ for all such E. As for $E \in \{0, I\}$, we note that $f_\rho(E) = f_{I/2^n}(E)$. Thus

$$\begin{aligned}
\|f_\rho\|_{\mathcal{D}}^2 &= \frac{1}{|\mathcal{E}_{\text{Pauli}}|} \sum_{E \in \mathcal{E}_{\text{Pauli}}} f_\rho(E)^2 \\
&= \frac{1}{|\mathcal{E}_{\text{Pauli}}|} \left(\sum_{E \in \mathcal{E}_{\text{Pauli}} \setminus \{0,I\}} f_\rho(E)^2 + \sum_{E \in \{0,I\}} f_\rho(E)^2 \right) \\
&= \frac{1}{|\mathcal{E}_{\text{Pauli}}|} \left(\sum_{E \in \mathcal{E}_{\text{Pauli}} \setminus \{0,I\}} (f_\rho(E) - f_{I/2^n}(E))^2 + \sum_{E \in \{0,I\}} f_\rho(E)^2 \right) \\
&= \frac{1}{|\mathcal{E}_{\text{Pauli}}|} \left(\sum_{E \in \mathcal{E}_{\text{Pauli}}} (f_\rho(E) - f_{I/2^n}(E))^2 + \sum_{E \in \{0,I\}} f_\rho(E)^2 \right) \\
&= \|f_\rho - f_{I/2^n}\|_{\mathcal{D}}^2 + \frac{2}{|\mathcal{E}_{\text{Pauli}}|} \\
&= \|f_\rho - f_{I/2^n}\|_{\mathcal{D}}^2 + \frac{1}{4^n}.
\end{aligned}$$

□

11.4.2 Lower Bounds via a Direct Reduction from Learning Parities

To get around this norm issue, we look at a subset of stabilizer states such that we can produce p-concepts with norm 1. Recall that the Pauli measurements are the set of all projectors onto the eigenvalue-1 space of some Pauli matrix P, i.e., $\{\frac{P+I}{2} : P \in \mathcal{P}_n\}$. We define a subset of the Pauli measurements called the parity measurements, and show the hardness of SQ-learning stabilizer states under the uniform distribution on such measurements. This is via a simple equivalence, holding essentially by construction, with the problem of learning parities under the uniform distribution. As a further consequence, we obtain that learning stabilizer states with noise is at least as hard as Learning Parity with Noise (LPN). This holds for general PAC-learning, even outside the SQ model. We remark that this hardness result emerges from a purely *classical* sub-problem of learning stabilizer states.

Definition 11.20 (Parity measurements)**.** *For all $x \in \mathbb{F}_2^n$, let $P_x = \sum_{y \in \mathbb{F}_2^n} \chi_x(y) \left|y\right\rangle\!\left\langle y\right|$. Since the set of P_x is equivalent to $\{I, Z\}^{\otimes n}$, the corresponding measurement $E_x = \frac{I - P_x}{2}$ is by definition a Pauli measurement. We will refer to such measurements as* PARITY MEASUREMENTS*.*

Proposition 11.21. *For every distribution $\mathcal{D}$ on $\mathbb{F}_2^n$ there is a corresponding distribution $\mathcal{D}'$ on parity measurements such that learning computational basis states under $\mathcal{D}'$ is equivalent to learning parities under $\mathcal{D}$. Furthermore, this equivalence holds even with classification noise: for any η, learning computational basis states under $\mathcal{D}'$ with noise rate η is equivalent to learning parities under $\mathcal{D}$ with noise rate η.*

In particular, learning stabilizer states under $\mathcal{D}'$ is at least as hard as learning parities under $\mathcal{D}$.

Proof. If the unknown state ρ is a computational basis state $\left|y\right\rangle\!\left\langle y\right|$, then the value

$$\mathrm{Tr}[E_x \left|y\right\rangle\!\left\langle y\right|] = \frac{1 - \chi_x(y)}{2} = x \cdot y.$$

In the PAC setting, this would be equivalent to getting the sample $(E_x, x \cdot y)$. Accordingly, let us define $\mathcal{D}'$ simply as the distribution over E_x for $x \sim \mathcal{D}$. It is clear that these are different representations of the same problem, such that a learning algorithm for one implies a learning algorithm for the other. We note that this relationship holds even in the presence of classification noise. Finally, note that computational basis states are a subset of the stabilizer states, so any learner for stabilizer states implies a learner for the computational basis states as well. This implies that learning stabilizer states on $\mathcal{D}'$ is at least as hard as learning parities on $\mathcal{D}$, even in the presence of classification noise. □

Corollary 11.22. *SQ-learning stabilizer states under the uniform distribution over parity measurements requires $2^{\Omega(n)}$ queries even with constant error (say 1/3).*

Proof. By Proposition 11.21, SQ-learning stabilizer states under the uniform distribution on E_x parity measurements is at least as hard as learning parities over the uniform distribution. Applying Theorem 11.9, we get the exponential lower bound. □

Corollary 11.23. *Learning stabilizer states under the uniform distribution over parity measurements with classification noise rate η is at least as hard as LPN with noise rate η.*

Proof. Proposition 11.21 directly implies that learning computational basis states under the uniform distribution on parity measurements and with classification noise is equivalent to LPN. □

11.5 An SQ Learner for Product States

Turning to positive results, we now give SQ algorithms for some simple concept classes, namely the computational basis states and, more generally, products of n single-qubit states. Such states have very well known folklore learning algorithms, since the number of parameters is merely polynomial in the number of qubits. This demonstrates that there are indeed learning algorithms for quantum states that can be used in the SQ framework, unlike stabilizer states. We hope that this helps motivate research into noise-tolerant learning algorithms, in spite of the lower-bound results given in Section 11.4.

The distribution on measurements that we will consider will correspond to a natural scheme for these classes: pick a qubit at random and measure it using a Haar-random unitary. Concretely, let $\mathcal{D}'$ be the distribution of single qubit measurements formed from the projection onto Haar-random single qubit state (i.e., $U\left|0\right\rangle\!\left\langle 0\right| U^{\dagger}$ where U is a Haar random unitary), and let $\mathcal{D}$ be the distribution on n-qubit measurements that corresponds to picking a qubit at random and measuring it using a measurement drawn from $\mathcal{D}'$. That is, $\mathcal{D} = \frac{1}{n}\sum_{i=1}^{n} I^{\otimes i-1} \otimes \mathcal{D}' \otimes I^{\otimes n-i}$. Let $\mathcal{C}$ be the concept class of product states $\rho = \otimes_{i=1}^{n}\rho_i$. Of course, this class includes the computational basis states. The main result of this section will be a simple $O(n)$-query SQ algorithm for learning $\mathcal{C}$ under the distribution $\mathcal{D}$.

We remark that our algorithm's guarantee actually trivially extends to learning arbitrary (not just product) states under the above distribution $\mathcal{D}$ of single-qubit Haar-random measurements. This is simply because such measurements only ever inspect each qubit individually, so that a product state $\otimes_i\rho_i$ is indistinguishable—*under* $\mathcal{D}$—from a more general mixed state ρ whose reduced density matrix on qubit i is ρ_i for every i.[3] Yet since this distribution on measurements is fundamentally not very interesting for anything other than product states, we state the results in this section only for product states.

The following technical lemma will be the backbone of our results.

Lemma 11.24. *For any single qubit pure state $\left|\psi\right\rangle\!\left\langle\psi\right| = \frac{I+P}{2}$ and mixed state ρ:*

$$\mathop{\mathbb{E}}_{E\sim\mathcal{D}'}\left[\mathsf{sgn}\left(\mathrm{Tr}\left[E\left|\psi\right\rangle\!\left\langle\psi\right|\right] - \frac{1}{2}\right)\left(\mathrm{Tr}\left[E\rho\right] - \frac{1}{2}\right)\right] = \frac{1}{4}\mathrm{Tr}[P\rho].$$

Proof. We will decompose $\rho = \lambda\left|\phi\right\rangle\!\left\langle\phi\right| + (1-\lambda)\left|\phi^{\perp}\right\rangle\!\left\langle\phi^{\perp}\right|$ such that $\left|\phi\right\rangle = \cos\theta'\left|\psi\right\rangle + \sin\theta'\left|\psi^{\perp}\right\rangle$ and $\left|\phi^{\perp}\right\rangle = e^{i\phi'}(\sin\theta'\left|\psi\right\rangle - \cos\theta'\left|\psi^{\perp}\right\rangle)$. The following identity will be useful at the end:

[3] We stress that in the PAC formalism, the goal is not necessarily to learn the exact state, but simply to find one that behaves similarly under the specified input distribution of measurements. Thus for measurements of the kind drawn from $\mathcal{D}$, learning the product of reduced density states is sufficient.

$$\begin{aligned}
\mathrm{Tr}[P\rho] &= 2\mathrm{Tr}[|\psi\rangle\!\langle\psi|\,\rho] - 1 \\
&= 2\big[\lambda \mathrm{Tr}[|\psi\rangle\!\langle\psi|\,|\phi\rangle\!\langle\phi|] + (1-\lambda)\mathrm{Tr}[|\psi\rangle\!\langle\psi|\,|\phi^{\perp}\rangle\!\langle\phi^{\perp}|]\big] - 1 \\
&= 2\cos^2\theta'\lambda + 2(1-\lambda)\sin^2\theta' - (\sin^2\theta' + \cos^2\theta') \\
&= (2\lambda-1)\cos^2\theta' - (2\lambda-1)\sin^2\theta' \\
&= (2\lambda-1)\cos 2\theta'
\end{aligned}$$

Let U be the unitary such that $U|0\rangle = |\psi\rangle$ and $U|1\rangle = |\psi^{\perp}\rangle$. Due to symmetry, we can parameterize a Haar-random single qubit state using spherical coordinates as $E = \frac{1}{2}U(I + \cos\phi\sin\theta X + \sin\phi\sin\theta Y + \cos\theta Z)U^{\dagger}$ for the Pauli matrices X, Y, and Z.

$$\begin{aligned}
\mathrm{Tr}\left[E\,|\psi\rangle\!\langle\psi|\right] &= \mathrm{Tr}\left[\frac{1}{2}U(I + \cos\phi\sin\theta X + \sin\phi\sin\theta Y + \cos\theta Z)U^{\dagger}\,|\psi\rangle\!\langle\psi|\right] \\
&= \mathrm{Tr}\left[\frac{1}{2}(I + \cos\phi\sin\theta X + \sin\phi\sin\theta Y + \cos\theta Z)\,|0\rangle\!\langle 0|\right] \\
&= \frac{1+\cos\theta}{2}
\end{aligned}$$

We can also do the same thing for ρ:

$$\begin{aligned}
&\mathrm{Tr}\left[E\rho\right] \\
&= \lambda\mathrm{Tr}[E\,|\phi\rangle\!\langle\phi|] + (1-\lambda)\mathrm{Tr}[E\,|\phi^{\perp}\rangle\!\langle\phi^{\perp}|] \\
&= \lambda\frac{1+\cos\theta\cos 2\theta' + \cos\phi\sin\theta\sin 2\theta'}{2} + (1-\lambda)\frac{1-\cos\theta\cos 2\theta' - \cos\phi\sin\theta\sin 2\theta'}{2} \\
&= \frac{1+(2\lambda-1)(\cos\theta\cos 2\theta' + \cos\phi\sin\theta\sin 2\theta')}{2}
\end{aligned}$$

This allows us to perform a spherical integral over θ and ϕ to get the expectation:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{E\sim\mathcal{D}'}\left[\mathsf{sgn}\left(\mathrm{Tr}\left[E\,|\psi\rangle\langle\psi|\right]-\frac{1}{2}\right)\left(\mathrm{Tr}\left[E\,|\phi\rangle\langle\phi|\right]-\frac{1}{2}\right)\right]\\
&=\frac{1}{4\pi}\int_0^{2\pi}d\phi\int_0^{\pi}d\theta\sin\theta\mathsf{sgn}\left(\cos\theta\right)\left((2\lambda-1)\frac{\cos\theta\cos2\theta'+\cos\phi\sin\theta\sin2\theta'}{2}\right)\\
&=\frac{2\lambda-1}{8\pi}\int_0^{2\pi}d\phi\left[\int_0^{\pi/2}-\int_{\pi/2}^{\pi}\right]d\theta\sin\theta\left(\cos\theta\cos2\theta'+\cos\phi\sin\theta\sin2\theta'\right)\\
&=(2\lambda-1)\frac{\pi\cos2\theta'+\pi\cos2\theta'}{8\pi}\\
&=(2\lambda-1)\frac{\cos2\theta'}{4}\\
&=\frac{1}{4}\mathrm{Tr}[P\rho] \qquad\square
\end{aligned}$$

Our algorithm for learning product states will be work by learning each qubit in the Pauli basis. This gives an estimate of each qubit's (reduced density matrix) location on the Bloch sphere. We then relate this location to squared loss. This results in a $3n$-query algorithm, corresponding to the $3n$ parameters that it takes to define a product state.

We first require a well-known fact about the trace distance (see Definition 2.5 for a definition) between single qubit states relative to their euclidean distance on the Bloch sphere.

Proposition 11.25 (folklore). *Given two single qubit states ρ and σ, the trace distance $\mathrm{TD}(\rho,\sigma)$ is half the Euclidean distance between their points on the Bloch sphere.*

The following lemma will then be necessary to relate trace distance of the states to the squared loss in learning under this distribution.

Lemma 11.26. *For n-qubit product states $\rho=\bigotimes_i\rho_i$ and $\sigma=\bigotimes_i\sigma_i$, let $f_\rho(E)=2\mathrm{Tr}[E\rho]-1$ and $f_\sigma(E)=2\mathrm{Tr}[E\sigma]-1$. Let $\mathcal{D}$ be the distribution over measurements defined earlier. Then*

$$\mathop{\mathbb{E}}_{E\sim\mathcal{D}}[(f_\rho(E)-f_\sigma(E))^2]=\frac{4}{3n}\sum_{i=1}^{n}\mathrm{TD}(\rho_i,\sigma_i)^2$$

Proof. Let $\xi = \rho - \sigma$. Then by linearity

$$f_\rho(E) - f_\sigma(E) = 2(\mathrm{Tr}[E\rho] - \mathrm{Tr}[E\sigma]) = 2\mathrm{Tr}[E\xi].$$

We will define $\xi_i = \mathrm{Tr}_i(\xi) = \rho_i - \sigma_i$ to be the reduced density matrix on the i^{th} qubit of ξ. Noting that each ξ_i is traceless, then by diagonalizing we can write $\xi_i = \lambda_i \left|\lambda_i\right\rangle\!\left\langle\lambda_i\right| - \lambda_i \left|\lambda_i^\perp\right\rangle\!\left\langle\lambda_i^\perp\right|$ for $\lambda_i \in [0,1]$ such that $\lambda_i = \mathrm{TD}(\rho_i, \sigma_i)$ is the trace distance of the reduced density matrix.

Like in Lemma 11.24, we can parameterize a single-qubit Haar-random projection as $E = \frac{1}{2}U(I + \cos\phi\sin\theta X + \sin\phi\sin\theta Y + \cos\theta Z)U^\dagger$, where $U\left|0\right\rangle = U\left|\lambda_i\right\rangle$ and $U\left|1\right\rangle = U\left|\lambda_i^\perp\right\rangle$. This implies that $U\xi_i U^\dagger = \lambda_i Z$.

$$\begin{aligned}
\mathrm{Tr}[E\xi_i] &= \mathrm{Tr}\left[\frac{1}{2}U\big(I + \cos\phi\sin\theta X + \sin\phi\sin\theta Y + \cos\theta Z\big)U^\dagger\rho\right] \\
&= \mathrm{Tr}\left[\frac{1}{2}\big(I + \cos\phi\sin\theta X + \sin\phi\sin\theta Y + \cos\theta Z\big)\cdot\lambda_i Z\right] \\
&= \lambda_i\cos\theta
\end{aligned}$$

Using this, we now compute the squared-loss as follow.

$$\begin{aligned}
\mathop{\mathbb{E}}_{E\sim\mathcal{D}}[(f_\rho(E) - f_\sigma(E))^2] &= \frac{1}{n}\sum_{i=1}^{n}\mathop{\mathbb{E}}_{E'\sim\mathcal{D}'}[(f_{\rho_i}(E') - f_{\sigma_i}(E')^2] \\
&= \frac{4}{n}\sum_{i=1}^{n}\mathop{\mathbb{E}}_{E'\sim\mathcal{D}'}\left[\mathrm{Tr}^2\left[E'\xi_i\right]\right] \\
&= \frac{4}{n}\sum_{i=1}^{n}\frac{1}{2}\int_0^\pi \sin\theta\cdot\lambda_i^2\cos^2\theta \\
&= \frac{4}{3n}\sum_{i=1}^{n}\lambda_i^2 \\
&= \frac{4}{3n}\sum_{i=1}^{n}\mathrm{TD}(\rho_i,\sigma_i)^2
\end{aligned}$$

□

We now show how to use Lemma 11.24 to learn each qubit of the product state, allowing us to then apply Lemma 11.26 to get our learning result.

Theorem 11.27. *Let $\mathcal{D}$ be the distribution on measurements and let $\mathcal{C}$ be the concept class of product states defined earlier. There is an SQ learner that is able to learn $\mathcal{C}$ under $\mathcal{D}$ up to squared loss ε using $3n$ queries of tolerance $\sqrt{\varepsilon}/n$.*

Proof. Let the unknown $\rho \in \mathcal{C}$ be given by $\rho = \bigotimes_i \rho_i$. If we define $P_1 = X$, $P_2 = Y$, and $P_3 = Z$, then our queries will be

$$\varphi_{i,j}(E, Y) = \mathsf{sgn}\left(\frac{1}{2^{n-1}}\mathrm{Tr}\left[E \cdot \left(I^{\otimes i-1} \otimes \frac{I + P_j}{2} \otimes I^{\otimes n-i}\right)\right] - \frac{1}{2}\right) \cdot Y$$

The query $\varphi_{i,j}$ will correspond to taking the projection of the i^{th} qubit along the Pauli P_j, as we now show:

$$\begin{aligned}
\mathop{\mathbb{E}}_{E\sim\mathcal{D}, Y\sim f_\rho(E)}[\varphi_{i,j}(E, Y)] &= \mathop{\mathbb{E}}_{E\sim\mathcal{D}}\left[\varphi_{i,j}(E,1)\mathrm{Tr}\left[E\rho\right] + \varphi_{i,j}(E,-1)\Big(1 - \mathrm{Tr}\left[E\rho\right]\Big)\right] \\
&= \mathop{\mathbb{E}}_{E\sim\mathcal{D}}\left[\varphi_{i,j}(E,1)\Big(2\mathrm{Tr}\left[E\rho\right] - 1\Big)\right] \\
&= \frac{1}{n}\mathop{\mathbb{E}}_{E'\sim\mathcal{D}'}\left[\mathsf{sgn}\left(\mathrm{Tr}\left[E' \left|0\right\rangle\!\left\langle 0\right|\right] - \frac{1}{2}\right)\Big(2\mathrm{Tr}\left[E'\rho_i\right] - 1\Big)\right] \\
&= \frac{1}{2n}\mathrm{Tr}[P_j\rho_i].
\end{aligned}$$

Here the third equality exploits the definition of $\mathcal{D}$ as $\frac{1}{n}\sum_{i=1}^{n} I^{\otimes i-1} \otimes \mathcal{D}' \otimes I^{\otimes n-i}$ (only the i^{th} term yields a nonzero expectation), and the fourth equality is Lemma 11.24.

Any specific qubit ρ_i can be written in Bloch sphere coordinates as $\frac{I+x_iX+y_iY+z_iZ}{2}$. We can estimate $x_i = \frac{1}{2}\mathrm{Tr}[P_1\rho_i]$ up to error $\sqrt{\varepsilon}$ using a single query of tolerance $\sqrt{\varepsilon}/n$. The same holds true for y_i and z_i. If we use this to construct our estimate

$$\widehat{\rho}_i = \frac{I + \widehat{x}_iX + \widehat{y}_iY + \widehat{z}_iZ}{2}$$

then by Proposition 11.25 we get

$$\mathrm{TD}(\rho_i, \widehat{\rho}_i)^2 = \frac{1}{4}\left[(x_i - \widehat{x}_i)^2 + (y_i - \widehat{y}_i)^2 + (z_i - \widehat{z}_i)^2\right] \leq 3\varepsilon/4.$$

Finally, using Lemma 11.26:

$$\mathop{\mathbb{E}}_{E\sim\mathcal{D}}[(f_\rho(E) - f_\sigma(E))^2] \leq \frac{4}{3n}\sum_{i=1}^{n} 3\varepsilon/4 = \varepsilon.$$

We note that if the estimated point is outside of the Bloch sphere, we can simply normalize the point to the surface of the Bloch sphere and this will never increase the error. To quickly sketch the proof of this, take the plane formed by the center of the sphere, the estimated point $\widehat{p}$ that is outside of the sphere, and the real point p which is both within the Bloch sphere and within a sphere ε radius located at $\widehat{p}$. The normalized point $\widehat{p}'$ is always located on the line from the $\widehat{p}$ to the origin, and one can make a separating plane that bisects the line segment between $\widehat{p}$ and $\widehat{p}'$ that denotes whether one is closer to $\widehat{p}$ or $\widehat{p}'$. Since the Bloch sphere will always be on the side closer to $\widehat{p}'$ and the real point p is in the Bloch sphere, p will always be closer to $\widehat{p}'$ than $\widehat{p}$. □

We can simplify this algorithm if we know in advance that ρ is a computational basis state. In that case, we know that each qubit ρ_i is either $(I+Z)/2$ or $(I-Z)/2$, and so we only need to make n queries $\varphi_{i,3}$, one for each i. Moreover, we only need to identify the coordinate z_i to within an accuracy of 1 in order to distinguish the $z_i = 1$ and $z_i = -1$ cases, so that our tolerance need only scale as $O(1/n)$ in order to learn ρ perfectly (i.e., with $\varepsilon = 0$).

11.6 Connections to Differential Privacy

A PAC learning algorithm L can be viewed as a randomized algorithm that takes as input a training dataset (i.e., a set of labeled examples (x, y) sampled from a distribution) and outputs a hypothesis that with high probability has low error over the distribution. That is, if S is a training dataset, then $L(S)$ describes a probability distribution over hypotheses (where the randomness arises from the internal randomness of the learner). Intuitively, differential privacy requires L to satisfy a kind of stability: on any two inputs S and S' that are close, the distributions $L(S)$ and $L(S')$ must be close as well.

Definition 11.28 (Differential privacy, [DR14])**.** *Call two datasets $S = \{(x_i, y_i)\}_{i=1}^m$ and $S' = \{(x_i', y_i')\}_{i=1}^m$ neighbors if they only differ in one entry. A learner L (understood in the sense just discussed) is said to be α-differentially private (or α-DP for short) if for any S and S' that are neighbors, the distributions $L(S)$ and $L(S')$ are close in the sense that for any hypothesis h, $\mathbb{P}[L(S) = h] \leq e^\alpha\, \mathbb{P}[L(S') = h]$.*

A well-known property of SQ algorithms is that they can readily be made differentially private [BDMN05, DR14]. Since differential privacy is a notion that is well-defined only in the PAC setting where the input is a set of training examples (as opposed to access to an SQ oracle), such a statement is necessarily of the form "any SQ learner yields a PAC learner that satisfies differential privacy."

Theorem 11.29 (see e.g., [Bal15])**.** *Let $\mathcal{C}$ be a concept class learnable up to error ε by an SQ learner L using q queries of tolerance τ. Then it is also learnable up to error ε in the PAC setting by an α-DP learner L' with sample complexity $\tilde{O}(\frac{q}{\alpha\tau} + \frac{q}{\tau^2})$ (with constant probability).*

The proof is standard and proceeds by simulating each of L's queries using empirical estimates over a sample of size roughly $1/\tau^2$ and then using the Laplace mechanism to add some further noise.

One can extend this notion to the quantum setting. One natural and direct way of doing so is simply by replacing the classical dataset of labeled pairs (x_i, y_i) by one of measurement-outcome pairs (E_i, Y_i); the rest remains exactly analogous. Theorem 11.29 then carries over verbatim to our notion of quantum SQ learnability. This form of quantum differential privacy was recently studied by Arunachalam et al. [QAS21], who were able to relate it to online learning, one-way communication complexity, and shadow tomography of quantum states, extending ideas of [BLM20]. Since our notion of quantum SQ learnability implies quantum DP learnability, it also fits into their framework. In particular, by the chain of implications established in that work, efficient quantum SQ learnability of a class of states implies DP PAC learnability, which implies finite sequential fat-shattering (sfat) dimension, which in turn implies online learnability, gentle shadow tomography, and "quantum stability." In fact, in the classical setting, some of the main examples of realistic DP learners are SQ (even though technically the inclusion is known to be strict) [BDMN05, KLN+11], and one might expect the same to hold in the quantum setting as well.

We remark that a somewhat different kind of quantum differential privacy, where privacy is with respect to copies of the unknown state, may also be defined as follows. View a quantum state learner L as an algorithm that takes in multiple copies $\rho^{\otimes m}$ of some unknown state ρ, is allowed to sample and perform random measurements from a distribution $\mathcal{D}$, and outputs another state σ that is required to be close to ρ with respect to $\mathcal{D}$ with high probability. If the random measurements are viewed as the internal randomness of the learner, then this is similar to the view we took of a classical learner earlier. We can now define a notion of differential privacy for quantum state learners by requiring that $L(\rho^{\otimes m})$ and $L(\rho^{\otimes m-1} \otimes \rho')$ (where $\rho \neq \rho'$, so that $\rho^{\otimes m}$ and $\rho^{\otimes m-1} \otimes \rho'$ are neighbors) are α-close as distributions over states (in the natural way). This can also be seen as a stylized kind of tolerance to noise or corruptions. The following analogue of Theorem 11.29 can then be proven using almost exactly the same proof; essentially, we are only replacing classical examples with copies of quantum states.

Theorem 11.30. *Let $\mathcal{C}$ be a class of quantum states learnable up to error ε by an SQ learner L using q queries of tolerance τ. Then it is also learnable up to error ε in the PAC setting by an α-DP learner L' (in the specific sense just described) with copy complexity $\tilde{O}(\frac{q}{\alpha\tau} + \frac{q}{\tau^2})$ (with constant probability).*

Note that these notions are different from those of [AR19], which defined differential privacy for quantum measurements. Here two n-qubit states are considered neighbors if it is possible to reach one from the other by a quantum operation (sometimes called a superoperator) on a single qubit. In particular, two product states $\rho = \otimes_i \rho_i$ and $\sigma = \otimes_i \sigma_i$ are neighbors if $\rho_i = \sigma_i$ for all i but one.

Definition 11.31 (Quantum differential privacy for measurements, [AR19])**.** *A measurement M is said to be α-DP if for any n-qubit neighbor states ρ, σ, and any outcome y, $\mathbb{P}[M(\rho) = y] \leq e^{\alpha}\, \mathbb{P}[M(\sigma) = y]$.*

The authors show that this definition can be related to the notion of a "gentle quantum measurement," and this connection can be carefully exploited to perform shadow tomography [Aar19]. However, this kind of quantum DP is not applicable in a natural way to a PAC or SQ learner, since such a learner is an algorithm rather than just a single measurement.

11.7 Discussion and Open Problems

Statistical vs. query complexity. Conceptually, the contrast between our SQ model and the original PAC model of [Aar07] is interesting. Apart from the definition of an elegant model, Aaronson's main insight was to characterize learnability in a purely *statistical* sense, showing bounds on sample complexity via an analysis of the so-called fat-shattering dimension of quantum states. In learning theoretic terms, this took advantage of a separation of concerns that the PAC model encourages: (a) empirical performance, i.e., a learner achieving low error with respect to the training data, and (b) generalization, i.e., this performance actually generalizing to the true distribution. The SQ model, however, does not naturally accommodate such a separation. SQ algorithms are instead primarily characterized by the number of queries required; generalization is "in-built." The closest analogue to a notion of sample complexity is the role played by the tolerance, and the closest thing to studying generalization on its own might have been to show a phase transition in what different regimes of the tolerance are able to accomplish. The formal statements of our SQ lower bounds do have such a flavor: "*either* use small tolerance *or* many queries."

Suitable classes and distributions for PAC-learning. It is notable that the algorithms of [Roc18] for learning stabilizer states and [Yog19] for low Schmidt rank states are essentially the only known positive results in the framework of [Aar07]. Both these algorithms are "algebraic" and involve solving a system of polynomial equations, something that SQ cannot handle. A longstanding question in this area is: what other interesting classes can be learned, and under what distributions on measurements? And can they also be learned in the SQ setting?

A major issue in picking suitable distributions on measurements is that under many natural distributions, the maximally mixed state actually performs quite well, so that the problem of learning becomes essentially superfluous. Even in this work, we obtained lower bounds for learning stabilizer states under the uniform distribution on Pauli measurements only for learning up to exponentially small squared loss. This was because the norms of the p-concepts are themselves exponentially small, or in other words the maximally mixed state already achieves exponentially small loss. We were able to get around this and obtain a $\Omega(2^n)$ lower bound via a direct reduction from learning parities (by considering parity measurements). Can we do better than just 2^n? Is there a $\omega(2^n)$-sized (e.g., 4^n or 2^{n^2}) subset of stabilizer states such that there exists a distribution over Pauli measurements inducing norms that are only polynomially small yet have an exponentially small average correlation? That is, is there a $\omega(2^n)$-sized set of stabilizer states and accompanying distribution over Pauli measurements such that the maximally mixed state does not do well?

Other forms of noise. Can we extend the noise tolerance of SQ algorithms to more forms of noise, or improve the parameters of the noise tolerated? One such interesting form of noise would be depolarizing noise that acts on individual qubits (as opposed to acting directly on the whole state).

Probabilistic Concepts for Stabilizer States in the PAC framework In this work, we assumed both probabilistic concepts as well as access to statistical queries only. We note that despite the decision to use the more difficult probabilistic concepts, Proposition 11.21 also gives hardness even in the case where exact measurement values are given. Likewise, the PAC learning algorithm in [Roc18] requires exact trace values, rather than probabilistic concepts. That leads to perhaps the most important open question of: can stabilizer states be PAC learned (even without noise) in the probabilistic concept model?

Noise-tolerant learning beyond SQ. The best-known PAC algorithms for learning parities with noise are due to [BKW03] and [Lyu05] and runs in slightly subexponential time. Interestingly, this means it beats the exponential SQ lower bound and is hence essentially the only known example of a noise-tolerant PAC algorithm that is not SQ (although it cannot handle noise arbitrarily close to the information-theoretic limit). Can we similarly hope for a noise-tolerant but non-SQ learner for stabilizer states that runs in subexponential time?

Flat distributions and unitary designs [HCP22] defined a so-called *locally flat* distribution, which is any distribution that is invariant under the single-qubit Clifford gates. The high level idea is that due to the fact that the Clifford unitaries form a unitary 3-design [Web16]. Thus the expectation of a locally flat distribution and the distribution over product states where each qubit is Haar random are the same for the purposes of their algorithm. Can a similar thing be done with the algorithm in Section 11.5, where rather than measure a uniformly drawn qubit with a Haar random state, the qubit is measured using a locally flat distribution?

Appendix

Appendix A

Sample Efficient PAC Learning of General Quantum Circuits

This chapter is based on the appendix of [Lia23]. We generalize the main results of [CD20], which is itself analogous to [GJ95], to allow for projective measurements beyond rank 1, such as in the settings used in Chapter 10. While this was not necessary for any results, we hope to give this proof as a reference in a way that is also more black-box accessible for future work.

A.1 Quantum Circuits as Polynomials

The end goal will be to show that the outputs of our concept class can be described as a set of polynomials with bounded degree. Combined with an upper bound on the number of polynomials in that set, we can later arrive at an upper bound on the pseudo-dimension, which itself is an upper-bound on fat-shattering dimension.

We now show a more terse version of Lemma 3 from [CD20].

Lemma A.1. *Consider a quantum circuit C with a fixed circuit structure (i.e., the location of the 2-qudit gate are in fixed positions, though the gates themselves can be arbitrary) comprised of at most Γ 2-qudit gates. Such circuit can be described using variables $c_1, c_2, \cdots, c_k \in \mathbb{R}$ such that $k = 2\gamma d^4$. Then for every pair of quantum states $|\psi\rangle$ and $|\phi\rangle$ there exists a polynomial $p_{(|\psi\rangle,|\phi\rangle)}(c_1, c_2, \cdots, c_k) = \mathrm{Tr}\big[C\, |\psi\rangle\!\langle\psi|\, C^\dagger\, |\phi\rangle\!\langle\phi|\big]$ with degree at most 2Γ.*

Proof. Every 2-qudit gate U can be naïvely expressed as the d^4 complex values that make up the $d^2 \times d^2$ unitary. By splitting up the complex values into a real and imaginary part, we get $2d^4$ real values to describe each 2-qudit gate. If $|\psi\rangle = \sum_{s\in\{0,1\}^n} \alpha_s |s\rangle$ then applying a 2-qudit unitary to $|\psi\rangle$ leaves us with the amplitudes of this new state being a polynomial of degree 1 in terms of the entries of U. Note that the α_i, along with the circuit structure, are what determine the coefficients of this polynomial. By applying all γ 2-qudit gates that comprise C, the amplitudes of $C|\psi\rangle$ can be described as a polynomial of degree Γ in $2\gamma d^4$ variables. Finally, since we can write $|\phi\rangle = \sum_{s\in\{0,1\}^n} \beta_s |s\rangle$, then the inner product $\langle\phi|C|\psi\rangle$ is some weighted linear combination of the amplitudes of $C|\psi\rangle$, which is a again polynomial with degree at most Γ. To get

$$\mathrm{Tr}\big[C|\psi\rangle\langle\psi| C^\dagger |\phi\rangle\langle\phi|\big] = |\langle\phi|C|\psi\rangle|^2$$

we note that the degree at most doubles when we multiply a polynomial by itself. This leaves us with $p_{(|\psi\rangle,|\phi\rangle)}(c_1, c_2, \cdots, c_k)$ as polynomial of degree at most 2Γ and $m = 2\gamma d^4$. □

Corollary A.2. *Consider a quantum circuit C with a fixed circuit structure (i.e., the location of the 2-qudit gate are in fixed positions, though the gates themselves can be arbitrary) comprised of at most γ 2-qudit gates. Such circuit can be described using variables $c_1, c_2, \cdots, c_k \in \mathbb{R}$ such that $k = 2\gamma d^4$. Then for every pair of quantum state $|\psi\rangle$ and projector Π there exists a polynomial $p_{(|\psi\rangle,\Pi)}(c_1, c_2, \cdots, c_k) := \mathrm{Tr}\big[C|\psi\rangle\langle\psi| C^\dagger\Pi\big]$ with degree at most 2γ.*

Proof. We note that $\Pi = \sum_i |\phi_i\rangle\langle\phi_i|$. By linearity of the trace

$$\begin{aligned}
p_{(|\psi\rangle,\Pi)}(c_1, c_2, \cdots, c_k) &= \mathrm{Tr}\big[C|\psi\rangle\langle\psi| C^\dagger\Pi\big] \\
&= \mathrm{Tr}\left[C|\psi\rangle\langle\psi| C^\dagger \sum_i |\phi_i\rangle\langle\phi_i|\right] \\
&= \sum_i \mathrm{Tr}\big[C|\psi\rangle\langle\psi| C^\dagger |\phi_i\rangle\langle\phi_i|\big] \\
&= \sum_i p_{(|\psi\rangle,|\phi_i\rangle)}(c_1, c_2, \cdots, c_k m).
\end{aligned}$$

By Lemma A.1, this is the sum of real polynomials in $2\Gamma d^4$ variables with degree at most 2γ. Since the sum does not increase the degree, we are done. □

Since we fixed the circuit structure, we will want to know how many circuit structures there are, because this directly bounds the number of polynomials we need to consider. The following result was the main ingredient in the proof of Lemma 2 from [CD20].

Lemma A.3 ([CD20] Lemma 2)**.** *There are at most $\frac{\gamma!\delta^{\gamma-\delta}}{(\gamma-\delta)!}(n!)^{\delta}$ ways to structure 2-qudit circuits with size γ and depth δ.*

A.2 Pseudo-Dimension of Concept Classes Described via Polynomials

The following is a generalization of [GJ95], which used the degree of polynomials to bound the pseudo-dimension of concept classes that could be defined using polynomials in the parameter space of the concepts.

Definition A.4. *The pseudo-dimension of a concept class $\mathcal{C}$ is the limit of the fat-shattering dimension parameter η as η goes to zero. Formally, the pseudo-dimension is $\lim_{\eta\to 0^+} fat_{\mathcal{C}}(\eta)$.*

Because fat-shattering dimension increases as η decreases, fat-shattering dimension is always upper-bounded by the pseudo-dimension for all values of $\eta > 0$.

Definition A.5. *Let $\{p_1, p_2, \cdots, p_m\} \subseteq \mathbb{R}^k \to \mathbb{R}$ be a set of m polynomials on k variables. For $\eta > 0$, the η-sign assignment of $\{p_1, p_2, \cdots, p_m\}$ on the input $(x_1, x_2, \cdots, x_k) \in \mathbb{R}^k$ is the vector $b \in \{-1, 0, 1\}^m$ such that*

$$b_i = \begin{cases} 1 & p_i(x_1, x_2, \cdots, x_k) \geq \eta \\ -1 & p_i(x_1, x_2, \cdots, x_k) \leq -\eta \\ 0 & \textit{otherwise} \end{cases}.$$

Lemma A.6 ([GJ95] Corollary 2.1)**.** *Let $\{p_1, p_2, \cdots, p_m\} \subseteq \mathbb{R}^k \to \mathbb{R}$ be a set of real polynomials in k variables with $m \geq k$, each of degree at most $d \geq 1$. Then the number of unique*

η-sign assignments that $\{p_1, p_2, \cdots, p_m\}$ can create over all inputs in $\mathbb{R}^k$ is at most $\left(\frac{8edm}{k}\right)^k$ in the limit as $\eta \to 0^+$.

Lemma A.6 is a stronger notion than pseudo-dimension, since it upper bounds the number of sign assignments over arbitrarily large sets of inputs. Since pseudo-dimension requires that $\mathcal{C}$ can achieve all (i.e., an exponential number of) sign assignments on some large set of samples, we can show that the pseudo-dimension cannot be too large. We formalize that notion here. Note that the polynomials in question in the following proof are over the parameters of the concept class, not the inputs. The intuition is that if the output of the concept is some bounded-degree polynomial in the parameter space, there cannot be too many sign assignments.

Theorem A.7 (Generalization of [GJ95] Theorem 2.2). *Let $\mathcal{C} \subseteq \Omega \to [0,1]$ be a concept class such that every element of $\mathcal{C}$ can be described via k different real variables $c_1, c_2, \cdots c_k \in \mathbb{R}$, as well as an index $l \in [s]$ for $s \geq 0$. Furthermore, for every $f_{c_1,c_2,\cdots c_k,l} \in \mathcal{C}$ and $x \in \Omega$, let $f_{c_1,c_2,\cdots c_k,l}(x) = p_{x,l}(c_1, c_2, \cdots c_k)$ where $p_{x,l}$ is one of s polynomials each with degree at most d for $d \geq 1$. Then the pseudo-dimension of $\mathcal{C}$ is at most $2k\log_2(8eds)$.*

Proof. Let $(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m) \subseteq \mathbb{R}$ be the largest set of points pseudo-shattered by $\mathcal{C}$. If $ms < k$, then there is no issue because the largest shattered set is smaller than k, which is smaller than $2k\log_2(8eds)$. Now assume that $ms \geq k$. By Definition A.4, there must exist some points $y_1, y_2, \cdots, y_m \in \mathbb{R}$ and some (potentially arbitrarily small) value $\eta > 0$ such that for all $b_i \in \{\pm 1\}^m$, there is a $f_{c_1,c_2,\cdots c_k} \in \mathcal{C}$ with $b_i \cdot (f_{c_1,c_2,\cdots c_k}(x_i) - y_i) \geq \eta$. Because $f_{c_1,c_2,\cdots c_k,l}(x_i) = p_{x_i,l}(c_1, c_2, \cdots c_k)$, we can define the new set of polynomials $p'_{i,l} = p_{x_i,l} - y_i$. This means that $\bigcup_{i,l}\{p'_{i,l}\}$ is a set of ms polynomials that must be able to create at all 2^m different sign assignments that define b. However, we know from Lemma A.6 that the number of different sign assignments is at most $\left(\frac{8edms}{k}\right)^k$ as long as $ms \geq k$, which we have assumed to be true. Therefore, $2^m \leq \left(\frac{8edms}{k}\right)^k$. Taking the logarithm of both sides,

$$m \leq k \log_2\left(\frac{8edms}{k}\right) = k \log_2(8eds) + k \log_2 \frac{m}{k}.$$

We divide the situation into two cases based on which of these two logarithms is bigger: $8eds \geq \frac{m}{k}$ and $8eds < \frac{m}{k}$. The first case is easy to analyze, since if $8eds \geq \frac{m}{k}$, then we directly get $m \leq 2k \log_2(8eds)$ via substitution on the right-hand side. The other case leads to $m < 2k \log_2 \frac{m}{k}$, also via substitution on the right-hand side. Solving this with the Lambert W-function tells us that if $k > 0$ then $m < ke^{-W_{-1}\left(-\frac{\ln 2}{2}\right)} = 4k$. Because $d \geq 1$ and $s \geq 1$ then $\log_2(8eds) \geq \log_2(4e) > 2$, so $m < 4k < 2k \log_2(8eds)$ for this other case as well. □

A.3 Pseudo-dimension for Quantum Circuits

Proposition A.8 (Stirling's approximation)**.**

$$\ln(n!) = n \ln n - n + O(\ln n) = O(n \ln n)$$

Theorem A.9 (Generalization of [CD20] Theorem 3)**.** *The pseudo-dimension of quantum circuits on n qudits comprised of at most γ 2-qudit gates with depth δ is upper bounded by $O(d^4 \delta \gamma^2 \log \gamma)$.*

Proof. We want to apply Theorem A.7 to the concept class of quantum circuits. We know from Corollary A.2 that for fixed circuit structure with γ gates and depth δ that it can be described as a polynomial with degree at most 2γ in the $2\gamma d^4$ real variables that describe the entries of the circuit. Furthermore, Lemma A.3 tells us that there is at most $\frac{\gamma \delta^{\gamma-\delta}}{(\gamma-\delta)!}(n!)^\delta$ different circuit structures. We then apply Theorem A.7 with $k = 2\gamma d^4$, $d = 2\gamma$ and $s = \frac{\gamma! \delta^{\gamma-\delta}}{(\gamma-\delta)!}(n!)^\delta$. As a result we get that the pseudo-dimension is at most

$$2k \log_2(8eds) = 4\gamma d^4 \log_2\left(8e(2\gamma)\frac{\gamma! \delta^{\gamma-\delta}}{(\gamma-\delta)!}(n!)^\delta\right) \tag{A.1}$$

We will now focus on giving an upper bound for the logarithmic term by showing that

$$\log_2\left(8e(2\gamma)\frac{\gamma!\delta^{\gamma-\delta}}{(\gamma-\delta)!}(n!)^{\delta}\right) = O(\delta\gamma\log\gamma).$$

Splitting up the logarithm into sums and applying Stirling's approximation to each factorial, we arrive at

$$\begin{aligned}\log_2\left(8e(2\gamma)\frac{\gamma!\delta^{\gamma-\delta}}{(\gamma-\delta)!}(n!)^{\delta}\right) &= 4+\log_2 e+\log_2\gamma+(\gamma-\delta)\log_2\gamma \\ &\quad + O(\gamma\ln\gamma)+\delta\cdot O(n\ln n)+O((\delta-\gamma)\ln(\gamma-\delta)) \\ &= O(\gamma\log\gamma+\delta n\log n+\delta\log\gamma)\end{aligned}$$

Due to the definition of circuit structure, we know that $\delta \leq \gamma$. WLOG, we can also assume that every qubit has been acted upon by at least one gate (even if it's just the identity gate) such that $n \leq \gamma$. Together, we arrive that the logarithmic term is at most $O(\delta\gamma\log\gamma)$.

Since we have achieved our goal of upper-bounding the logarithmic term, Eq. (A.1) shows that the pseudo-dimension is at most $O(d^4\delta\gamma^2\log\gamma)$. □

We now state the generalization of the main result in [CD20] to projective measurements of arbitrary rank.

Corollary A.10 (Generalization of [CD20] Corollary 3)**.** *Let X be the set of quantum states on n qudits, and let Y be the set of all projectors on n qudits. Let U_* be a quantum circuit of 2-qudit quantum gates with size Γ and depth Δ. Let $\mathcal{D}$ be a probability distribution on $X \times Y$ unknown to the learner. Let*

$$S = \left\{\left(\left(x^{(i)}, y^{(i)}\right), \mathrm{Tr}\left[y^{(i)}U_* x^{(i)}U_*^{\dagger}\right]\right)\right\}_{i=1}^{m}$$

be corresponding training data where each $\left(x^{(i)}, y^{(i)}\right)$ is drawn i.i.d according to $\mathcal{D}$. Let $\delta, \varepsilon, \alpha, \beta \in (0,1)$ where $\beta > \alpha$. Then, training data of size

$$m = O\left(\frac{1}{\varepsilon}\left(\Delta d^4\Gamma^2\log\Delta\log^2\left(\frac{\Delta d^4\Gamma^2\log(\Gamma)}{(\beta-\alpha)\varepsilon}\right)+\log\frac{1}{\delta}\right)\right)$$

suffice to guarantee that, with probability $\geq 1-\delta$ *with regard to choice of the training data, any quantum circuit* U *of size* Γ *and depth* Δ *that satisfies*

$$\left|\mathrm{Tr}\left[y^{(i)}U_* x^{(i)}U_*^\dagger\right] - \mathrm{Tr}\left[y^{(i)}Ux^{(i)}U^\dagger\right]\right| \leq \alpha \quad \forall 1 \leq i \leq m$$

also satisfies

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\left(\mathrm{Tr}\left[y^{(i)}U_* x^{(i)}U_*^\dagger\right] - \mathrm{Tr}\left[y^{(i)}Ux^{(i)}U^\dagger\right]\right)^2\right] \leq (1-\varepsilon)\beta^2 + \varepsilon$$

Proof. We combine Theorem A.9 with Theorem 9.5, along with the fact that for all $\eta > 0$ the η-fat-shattering dimension is upper-bounded by pseudo-dimension. □

As shown by Theorem 4 of [CD20], a similar thing can be done with n-qudit quantum processes by simply changing the d^4 to d^8 in Lemma A.1 and Corollary A.2. This is because a quantum process is still a linear operation, but contains d^8 many entries now in parameter space. This propagates to Theorem A.9 and Corollary A.10 by again replacing every appearance of d^4 with d^8.

Bibliography

[Aar07] Scott Aaronson. The learnability of quantum states. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 463(2088):3089–3114, 2007. doi:10.1098/rspa.2007.0113. 25, 26, 27, 90, 138, 143, 144, 148, 173, 175, 177, 200, 201

[Aar19] Scott Aaronson. Shadow tomography of quantum states. *SIAM Journal on Computing*, 49(5):STOC18–368, 2019. doi:10.1145/3188745.3188802. 105, 200

[Aar22] Scott Aaronson. Introduction to quantum information science ii lecture notes, May 2022. URL: https://www.scottaaronson.com/qisii.pdf. 20, 123, 144

[AB00] Martin Anthony and Peter L. Bartlett. Function learning from interpolation. *Combinatorics, Probability and Computing*, 9(3):213–225, 2000. doi:10.1017/S0963548300004247. 138, 140

[AB08] A. Angelow and M. C. Batoni. About heisenberg uncertainty relation (by e.schrodinger), 2008. arXiv:quant-ph/9903100. 40

[ABDY22] Srinivasan Arunachalam, Sergey Bravyi, Arkopal Dutt, and Theodore J. Yoder. Optimal algorithms for learning quantum phase states, 2022. doi:10.48550/arxiv.2208.07851. 104

[ABF$^+$22] Scott Aaronson, Adam Bouland, Bill Fefferman, Soumik Ghosh, Umesh Vazirani, Chenyi Zhang, and Zixin Zhou. Quantum Pseudoentanglement, 2022. arXiv:2211.00747. 87, 125, 127

[ABJ$^+$03] J. B. Altepeter, D. Branning, E. Jeffrey, T. C. Wei, P. G. Kwiat, R. T. Thew, J. L. O'Brien, M. A. Nielsen, and A. G. White. Ancilla-assisted quantum process tomography. *Phys. Rev. Lett.*, 90:193601, May 2003. doi:10.1103/PhysRevLett.90.193601. 147

[ABNOGm22] Srinivasan Arunachalam, Sergey Bravyi, Chinmay Nirkhe, and Bryan O'-Gorman. The Parametrized Complexity of Quantum Verification. In *17th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2022)*, volume 232 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 3:1–3:18, 2022. doi:10.4230/LIPIcs.TQC.2022.3. 102

[ABSS97] Sanjeev Arora, László Babai, Jacques Stern, and Z Sweedyk. The Hardness of Approximate Optima in Lattices, Codes, and Systems of Linear Equations. *Journal of Computer and System Sciences*, 54(2):317–331, 1997. doi:10.1006/jcss.1997.1472. 98

[ACH$^+$19] Scott Aaronson, Xinyi Chen, Elad Hazan, Satyen Kale, and Ashwin Nayak. Online learning of quantum states. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124019, 2019. doi:10.1088/1742-5468/ab3988. 175

[AD98] Javed A Aslam and Scott E Decatur. Specification and simulation of statistical query algorithms for efficiency and noise tolerance. *Journal of Computer and System Sciences*, 56(2):191–208, 1998. doi:10.1145/225298.225351. 27, 182

[Ad17] Srinivasan Arunachalam and Ronald de Wolf. Guest column: A survey of quantum learning theory. *ACM SIGACT News*, 48(2):41–67, 2017. doi:10.1145/3106700.3106710. 150, 175

[Ad18] Srinivasan Arunachalam and Ronald de Wolf. Optimal quantum sample complexity of learning algorithms. *The Journal of Machine Learning Research*, 19(1):2879–2878, 2018. doi:10.48550/arXiv.1607.00932. 150, 175

[ADR82] Alain Aspect, Jean Dalibard, and Gérard Roger. Experimental test of bell's inequalities using time-varying analyzers. *Phys. Rev. Lett.*, 49:1804–1807, Dec 1982. doi:10.1103/PhysRevLett.49.1804. 17

[AG04] Scott Aaronson and Daniel Gottesman. Improved simulation of stabilizer circuits. *Phys. Rev. A*, 70:052328, Nov 2004. doi:10.1103/PhysRevA.70.052328. 19, 28, 39, 40, 41, 43, 102, 104, 108, 153, 154, 169

[AG08] Scott Aaronson and Daniel Gottesman. Identifying Stabilizer States, 2008. https://pirsa.org/08080052. 123

[AG23] Scott Aaronson and Sabee Grewal. Efficient learning of non-interacting fermion distributions, 2023. arXiv:2102.10458. 104, 118

[AGS21] Srinivasan Arunachalam, Alex Bredariol Grilo, and Aarthi Sundaram. Quantum hardness of learning shallow classical circuits. *SIAM Journal on Computing*, 50(3):972–1013, 2021. doi:10.1137/20M1344202. 168

[AGY20] Srinivasan Arunachalam, Alex B Grilo, and Henry Yuen. Quantum statistical query learning. *arXiv preprint arXiv:2002.08240*, 2020. arXiv:2002.08240. 150, 175

[AL88] Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988. doi:10.1023/a:1022873112823. 181

[AQY22] Prabhanjan Ananth, Luowen Qian, and Henry Yuen. Cryptography from Pseudorandom Quantum States. In *Advances in Cryptology – CRYPTO 2022*, pages 208–236. Springer Nature Switzerland, 2022. doi:10.1007/978-3-031-15802-5_8. 124

[AR19] Scott Aaronson and Guy N Rothblum. Gentle measurement of quantum states and differential privacy. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 322–333, 2019. doi:10.1145/3313276.3316378. 105, 200

[Bal15] Maria-Florina Balcan. Differential privacy and statistical query learning, 2015. URL: https://www.cs.cmu.edu/~avrim/ML07/lect1207.pdf. 198

[BBB$^+$23] Sara Bartolucci, Patrick Birchall, Hector Bombin, Hugo Cable, Chris Dawson, Mercedes Gimeno-Segovia, Eric Johnston, Konrad Kieling, Naomi Nickerson, Mihir Pant, et al. Fusion-based quantum computation. *Nature Communications*, 14(1):912, 2023. doi:10.1038/s41467-023-36493-1. 101

[BBC$^+$19] Sergey Bravyi, Dan Browne, Padraic Calpin, Earl Campbell, David Gosset, and Mark Howard. Simulation of quantum circuits by low-rank stabilizer decompositions. *Quantum*, 3:181, September 2019. doi:10.22331/q-2019-09-02-181. 21, 45, 46, 80, 89, 101, 104

[BCG13] K. Banaszek, M. Cramer, and D. Gross. Focus on quantum tomography. *New Journal of Physics*, 15(12):125020, 2013. doi:10.1088/1367-2630/15/12/125020. 100

[BCHK20] Michael Beverland, Earl Campbell, Mark Howard, and Vadym Kliuchnikov. Lower bounds on the non-Clifford resources for quantum computations. *Quantum Science and Technology*, 5(3):035009, 2020. doi:10.1088/2058-9565/ab8963. 46

[BCQ23] Zvika Brakerski, Ran Canetti, and Luowen Qian. On the Computational Hardness Needed for Quantum Cryptography. In *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*, volume 251 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 24:1–24:21, 2023. doi:10.4230/LIPIcs.ITCS.2023.24. 124

[BDMN05] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: The SuLQ framework. In *Proceedings of the Twenty-Fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 128–138, 2005. doi:10.1145/1065167.1065184. 172, 198, 199

[BEHW89] Anselm Blumer, A. Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *J. ACM*, 36(4):929–965, oct 1989. doi:10.1145/76359.76371. 139

[BFJ$^+$94] Avrim Blum, Merrick Furst, Jeffrey Jackson, Michael Kearns, Yishay Mansour, and Steven Rudich. Weakly learning DNF and characterizing statistical query learning using Fourier analysis. In *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing*, pages 253–262, 1994. doi:10.1145/195058.195147. 172

[BFS99] Jonathan F Buss, Gudmund S Frandsen, and Jeffrey O Shallit. The computational complexity of some problems of linear algebra. *Journal of Computer and System Sciences*, 58(3):572–596, 1999. doi:https://doi.org/10.1006/jcss.1998.1608. 28, 149, 158, 163

[BFV20] Adam Bouland, Bill Fefferman, and Umesh Vazirani. Computational Pseudorandomness, the Wormhole Growth Paradox, and Constraints on the AdS/CFT Duality (Abstract). In *11th Innovations in Theoretical Computer Science Conference (ITCS 2020)*, volume 151 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 63:1–63:2, 2020. doi:10.4230/LIPIcs.ITCS.2020.63. 124

[BG16] Sergey Bravyi and David Gosset. Improved Classical Simulation of Quantum Circuits Dominated by Clifford Gates. *Phys. Rev. Lett.*, 116:250501, 2016. doi:10.1103/PhysRevLett.116.250501. 104

[BHMT02] Gilles Brassard, Peter Høyer, Michele Mosca, and Alain Tapp. Quantum Amplitude Amplification and Estimation, 2002. doi:10.1090/conm/305/05215. 82

[BK05] Sergey Bravyi and Alexei Kitaev. Universal quantum computation with ideal Clifford gates and noisy ancillas. *Phys. Rev. A*, 71:022316, 2005. doi:10.1103/PhysRevA.71.022316. 89, 101

[BKW03] Avrim Blum, Adam Kalai, and Hal Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *J. ACM*, 50(4):506–519, jul 2003. doi:10.1145/792538.792543. 180, 202

[BLM20] M. Bun, R. Livni, and S. Moran. An equivalence between private classification and online prediction. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 389–402, Los Alamitos, CA, USA, November 2020. IEEE Computer Society. doi:10.1109/focs46700.2020.00044. 175, 199

[BLR93] Manuel Blum, Michael Luby, and Ronitt Rubinfeld. Self-Testing/Correcting with Applications to Numerical Problems. *J. Comput. Syst. Sci.*, 47(3):549–595, 1993. doi:10.1016/0022-0000(93)90044-W. 75

[Blu15] Avrim Blum. Computational Hardness of Learning, 2015. Lecture notes for CS 10-806 Foundations of Machine Learning and Data Science. URL: http://www.cs.cmu.edu/~avrim/ML07/lect1007.pdf. 149, 168

[BM99] Dagmar Bruß and Chiara Macchiavello. Optimal state estimation for d-dimensional quantum systems. *Physics Letters A*, 253(5-6):249–251, 1999. doi:10.1016/S0375-9601(99)00099-7. 100, 102

[BO21] Costin Bădescu and Ryan O'Donnell. Improved Quantum Data Analysis. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2021, page 1398–1411. Association for Computing Machinery, 2021. doi:10.1145/3406325.3451109. 105

[BPA22] Ludwig Burger, Niklas Pollard, and Jonathan Allen. Sleuths of 'spooky' quantum science win nobel physics prize. *Reuters*, 2022. URL: https://www.reuters.com/world/aspect-clauser-zeilinger-win-2022-nobel-prize-physics-2022-10-04/. 17

[BR92] Avrim L. Blum and Ronald L. Rivest. Training a 3-node neural network is np-complete. *Neural Networks*, 5(1):117–127, 1992. doi:https://doi.org/10.1016/S0893-6080(05)80010-3. 149, 168

[Bra22] Zvika Brakerski. Black-Hole Radiation Decoding is Quantum Cryptography, 2022. arXiv:2211.05491. 124

[BS19] Zvika Brakerski and Omri Shmueli. (pseudo) random quantum states with binary phase. In Dennis Hofheinz and Alon Rosen, editors, *Theory of Cryp-*

tography, pages 229–250, Cham, 2019. Springer International Publishing. doi:10.1007/978-3-030-36030-6_10. 127

[BSS16] Sergey Bravyi, Graeme Smith, and John A. Smolin. Trading Classical and Quantum Computational Resources. *Physical Review X*, 6(2):021043, 2016. doi:10.1103/PhysRevX.6.021043. 89

[BV97] Ethan Bernstein and Umesh Vazirani. Quantum complexity theory. *SIAM Journal on Computing*, 26(5):1411–1473, 1997. doi:10.1137/S0097539796300921. 181

[Car21] Matthias C. Caro. Binary classification with classical instances and quantum labels. *Quantum Machine Intelligence*, 3(1), may 2021. doi:10.1007/s42484-021-00043-z. 150

[CD20] Matthias C. Caro and Ishaun Datta. Pseudo-dimension of quantum circuits. *Quantum Machine Intelligence*, 2(2), Nov 2020. doi:10.1007/s42484-020-00027-5. 27, 28, 138, 148, 150, 151, 168, 169, 205, 207, 209, 210, 211

[CHL$^+$22] Sitan Chen, Brice Huang, Jerry Li, Allen Liu, and Mark Sellke. Tight Bounds for State Tomography with Incoherent Measurements, 2022. arXiv:2206.05265. 100

[CHY15] Hao-Chung Cheng, Min-Hsiu Hsieh, and Ping-Cheng Yeh. The learnability of unknown quantum measurements, 2015. arXiv:1501.00559, doi:10.26421/QIC16.7-8-4. 175

[CL21] Kai-Min Chung and Han-Hsuan Lin. Sample Efficient Algorithms for Learning Quantum Channels in PAC Model and the Approximate State Discrimination Problem. In Min-Hsiu Hsieh, editor, *16th Conference on the Theory of*

Quantum Computation, Communication and Cryptography (TQC 2021), volume 197 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 3:1–3:22, Dagstuhl, Germany, 2021. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.TQC.2021.3. 150

[CN97] Isaac L. Chuang and M. A. Nielsen. Prescription for experimental determination of the dynamics of a quantum black box. *Journal of Modern Optics*, 44(11-12):2455–2467, 1997. doi:10.1080/09500349708231894. 147

[CPF+10] 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. *Nature Communications*, 1(1):1–7, 2010. doi:10.1038/ncomms1147. 104

[CS96] A. R. Calderbank and Peter W. Shor. Good quantum error-correcting codes exist. *Phys. Rev. A*, 54:1098–1105, 1996. doi:10.1103/PhysRevA.54.1098. 19

[CSS15] Andrew W Cross, Graeme Smith, and John A Smolin. Quantum learning robust against noise. *Physical Review A*, 92(1):012327, 2015. doi:10.1103/physreva.92.012327. 176, 180

[DKS98] Irit Dinur, Guy Kindler, and Shmuel Safra. Approximating-CVP to within almost-polynomial factors is NP-hard. In *Proceedings 39th Annual Symposium on Foundations of Computer Science*, FOCS '98, pages 99–109. IEEE, 1998. doi:10.1109/SFCS.1998.743433. 98

[DPS03] G. Mauro D'Ariano, Matteo G.A. Paris, and Massimiliano F. Sacchi. Quantum tomography. *Advances in Imaging and Electron Physics*, 128:206–

309, 2003. doi:10.1002/(SICI)1521-3978(200005)48:5/7<579::AID-PROP579>3.0.CO;2-P. 100

[DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014. doi:10.1561/0400000042. 27, 172, 198

[DSS16] Amit Daniely and Shai Shalev-Shwartz. Complexity theoretic limitations on learning dnf's. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 815–830, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL: https://proceedings.mlr.press/v49/daniely16.html. 168

[EPR35] A. Einstein, B. Podolsky, and N. Rosen. Can quantum-mechanical description of physical reality be considered complete? *Phys. Rev.*, 47:777–780, May 1935. doi:10.1103/PhysRev.47.777. 17

[Fan57] Ugo Fano. Description of States in Quantum Mechanics by Density Matrix and Operator Techniques. *Reviews of Modern Physics*, 29(1):74, 1957. doi:10.1103/RevModPhys.29.74. 100

[FBaK21] Daniel Stilck França, Fernando G.S L. Brandão, and Richard Kueng. Fast and Robust Quantum State Tomography from Few Basis Measurements. In *16th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2021)*, volume 197 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 7:1–7:13, 2021. doi:10.4230/LIPIcs.TQC.2021.7. 118

[FC72] Stuart J. Freedman and John F. Clauser. Experimental test of local hidden-variable theories. *Phys. Rev. Lett.*, 28:938–941, Apr 1972. doi:10.1103/PhysRevLett.28.938. 17

[Fel12] Vitaly Feldman. A complete characterization of statistical query learning with applications to evolvability. *Journal of Computer and System Sciences*, 78(5):1444–1459, 2012. doi:10.1109/focs.2009.35. 172

[Fel16] Vitaly Feldman. *Statistical Query Learning*, pages 2090–2095. Springer New York, New York, NY, 2016. doi:10.1007/978-1-4939-2864-4_401. 171

[FLY22] Zhiyuan Fan, Jiatu Li, and Tianqi Yang. The exact complexity of pseudorandom functions and the black-box natural proof barrier for bootstrapping results in computational complexity. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2022, page 962–975, New York, NY, USA, 2022. Association for Computing Machinery. doi:10.1145/3519935.3520010. 127

[Ger13] Manuel Gerken. Measure concentration: Levy's Lemma, 2013. 129

[GGJ+20] Surbhi Goel, Aravind Gollakota, Zhihan Jin, Sushrut Karmalkar, and Adam Klivans. Superpolynomial lower bounds for learning one-layer neural networks using gradient descent. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3587–3596. PMLR, 13–18 Jul 2020. URL: https://proceedings.mlr.press/v119/goel20a.html. 27, 174, 178, 179

[GGR98] Oded Goldreich, Shari Goldwasser, and Dana Ron. Property Testing and Its Connection to Learning and Approximation. *Journal of the ACM (JACM)*, 45(4):653–750, 1998. doi:10.1145/285055.285060. 73

[GIKL23a] Sabee Grewal, Vishnu Iyer, William Kretschmer, and Daniel Liang. Efficient learning of quantum states prepared with few non-clifford gates, 2023. arXiv:2305.13409. 48, 60, 100

[GIKL23b] Sabee Grewal, Vishnu Iyer, William Kretschmer, and Daniel Liang. Improved stabilizer estimation via bell difference sampling, 2023. arXiv:2304.13915. 23, 30, 48, 60, 73, 88, 124

[GIKL23c] Sabee Grewal, Vishnu Iyer, William Kretschmer, and Daniel Liang. Low-Stabilizer-Complexity Quantum States Are Not Pseudorandom. In Yael Tauman Kalai, editor, *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*, volume 251 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 64:1–64:20, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.ITCS.2023.64. 23, 30, 48, 52, 60, 73, 74, 124, 127

[GJ95] Paul W. Goldberg and Mark R. Jerrum. Bounding the vapnik-chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning*, 18:131–148, 1995. doi:10.1007/BF00993408. 205, 207, 208

[GKZ19] Alex B Grilo, Iordanis Kerenidis, and Timo Zijlstra. Learning-with-errors problem is easy with quantum samples. *Physical Review A*, 99(3):032314, 2019. doi:10.1103/physreva.99.032314. 176

[GL22] Aravind Gollakota and Daniel Liang. On the Hardness of PAC-learning Stabilizer States with Noise. *Quantum*, 6:640, February 2022. doi:10.22331/q-2022-02-02-640. 26, 27, 138, 171

[GNW21] David Gross, Sepehr Nezami, and Michael Walter. Schur–Weyl duality for the Clifford group with applications: Property testing, a robust Hudson theorem, and de Finetti representations. *Communications in Mathematical Physics*, 385(3):1325–1393, 2021. doi:10.1007/s00220-021-04118-7. 8, 9, 16, 22, 23, 48, 52, 53, 70, 74, 76, 79, 83

[Got97] Daniel Gottesman. Stabilizer Codes and Quantum Error Correction, 1997. doi:10.48550/arxiv.quant-ph/9705052. 30

[Got98] Daniel Gottesman. The heisenberg representation of quantum computers, 1998. doi:10.48550/ARXIV.QUANT-PH/9807006. 19, 153

[GR09] Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. *SIAM Journal on Computing*, 39(2):742–765, 2009. arXiv:https://doi.org/10.1137/070685798, doi:10.1137/070685798. 148, 169

[Gro06] David Gross. Hudson's theorem for finite-dimensional quantum systems. *Journal of Mathematical Physics*, 47(12):122107, 2006. doi:10.1063/1.2393152. 88

[Hag20] Nika Haghtalab. Lecture 9: Hardness of Learning, 2020. Lecture notes for CS6781 - Theoretical Foundations of Machine Learning. URL: https://www.cs.cornell.edu/courses/cs6781/2020sp/lectures/09-hardness1.pdf. 149, 168

[Ham89] M. Hamermesh. *Group Theory and Its Application to Physical Problems*. Addison Wesley Series in Physics. Dover Publications, 1989. doi:10.1119/1.1941790. 187

[HCP22] Hsin-Yuan Huang, Sitan Chen, and John Preskill. Learning to predict arbitrary quantum processes, 2022. doi:10.48550/ARXIV.2210.14894. 22, 202

[HHJ+17] Jeongwan Haah, Aram W. Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu. Sample-optimal tomography of quantum states. *IEEE Transactions on Information Theory*, pages 1–1, 2017. doi:10.1109/tit.2017.2719044. 20, 100

[HIN+22] Marcel Hinsche, Marios Ioannou, Alexander Nietner, Jonas Haferkamp, Yihui Quek, Dominik Hangleiter, Jean-Pierre Seifert, Jens Eisert, and Ryan Sweke. A single t-gate makes distribution learning hard, 2022. doi:10.48550/arxiv.2207.03140. 105, 126

[HK23] Tobias Haug and M. S. Kim. Scalable Measures of Magic Resource for Quantum Computers. *PRX Quantum*, 4(1):010301, 2023. doi:10.1103/PRXQuantum.4.010301. 22

[HKP20] 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. doi:10.1038/s41567-020-0932-7. 24, 88, 94, 95, 103, 105

[HMMH+20] Jonas Haferkamp, Felipe Montealegre-Mora, Markus Heinrich, Jens Eisert, David Gross, and Ingo Roth. Quantum homeopathy works: Efficient unitary designs with a system-size independent number of non-Clifford gates, 2020. doi:10.48550/arxiv.2002.09524. 101

[HMY22] Minki Hhan, Tomoyuki Morimae, and Takashi Yamakawa. From the Hardness of Detecting Superpositions to Cryptography: Quantum Public Key

Encryption and Commitments, 2022. arXiv:2210.05978. 124

[HS07] Lisa Hellerstein and Rocco A Servedio. On pac learning algorithms for rich boolean function classes. *Theoretical Computer Science*, 384(1):66–76, 2007. doi:10.1007/11750321_42. 172

[IKOS08] Yuval Ishai, Eyal Kushilevitz, Rafail Ostrovsky, and Amit Sahai. Cryptography with constant computational overhead. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, STOC '08, page 433–442, New York, NY, USA, 2008. Association for Computing Machinery. doi:10.1145/1374376.1374438. 127

[JLS18] Zhengfeng Ji, Yi-Kai Liu, and Fang Song. Pseudorandom quantum states. In *Advances in Cryptology – CRYPTO 2018: 38th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 19–23, 2018, Proceedings, Part III*, page 126–152, Berlin, Heidelberg, 2018. Springer-Verlag. doi:10.1007/978-3-319-96878-0_5. 9, 24, 103, 124, 133

[Kat11] Jonathan Katz. Notes on Complexity Theory Lecture 3, 2011. Lecture notes for CS 652 — Complexity Theory. URL: https://www.cs.umd.edu/~jkatz/complexity/f11/lecture3.pdf. 142

[Kea98] Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998. doi:10.1145/293347.293351. 26, 27, 171, 172, 180, 181, 183

[KG15] Richard Kueng and David Gross. Qubit stabilizer states are complex projective 3-designs, 2015. doi:10.48550/ARXIV.1510.02767. 98

[Kha93] Michael Kharitonov. Cryptographic hardness of distribution-specific learning. In *Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory* *of Computing*, STOC '93, page 372–381, New York, NY, USA, 1993. Association for Computing Machinery. doi:10.1145/167088.167197. 168, 169

[KL93] Michael Kearns and Ming Li. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22(4):807–837, 1993. doi:10.1137/0222052. 181

[Kli05] Adam Klivans. The PAC Learning Model, 2005. Lecture notes for CS 395T Computational Learning Theory. URL: https://www.cs.utexas.edu/~klivans/f06lec2.pdf. 148

[KLN+11] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011. doi:10.1137/090756090. 199

[KLR+08] E. Knill, D. Leibfried, R. Reichle, J. Britton, R. B. Blakestad, J. D. Jost, C. Langer, R. Ozeri, S. Seidelin, and D. J. Wineland. Randomized benchmarking of quantum gates. *Physical Review A*, 77(1), 2008. doi:10.1103/physreva.77.012307. 19

[Kni04] E. Knill. Fault-Tolerant Postselected Quantum Computation: Schemes, 2004. arXiv:quant-ph/0402171. 101

[Koc22] Lucas Kocia. Improved Strong Simulation of Universal Quantum Circuits, 2022. arXiv:2012.11739. 89

[KQST22] William Kretschmer, Luowen Qian, Makrand Sinha, and Avishay Tal. Quantum Cryptography in Algorithmica, 2022. arXiv:2212.00879. 125

[Kre21] William Kretschmer. Quantum Pseudorandomness and Classical Complexity. In *16th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2021)*, volume 197 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 2:1–2:20, 2021. doi:10.4230/LIPIcs.TQC.2021.2. 125

[KRT17] 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. doi:10.1016/j.acha.2015.07.007. 100

[KS14] Robert Koenig and John A. Smolin. How to efficiently select an arbitrary clifford group element. *Journal of Mathematical Physics*, 55(12):122202, Dec 2014. doi:10.1063/1.4903507. 42

[KSS92] Michael J. Kearns, Robert E. Schapire, and Linda M. Sellie. Toward efficient agnostic learning. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 341–352, 1992. doi:10.1007/BF00993468. 89

[KT22] J. Kleinberg and E. Tardos. *Algorithm Design*. Pearson Education, 2022. URL: https://books.google.com/books?id=GORecgAACAAJ. 163

[LC22] Ching-Yi Lai and Hao-Chung Cheng. Learning quantum circuits of some t gates. *IEEE Transactions on Information Theory*, pages 1–1, 2022. doi:10.1109/TIT.2022.3151760. 22, 24, 27, 104, 105, 147

[Lia23] Daniel Liang. Clifford Circuits can be Properly PAC Learned if and only if $\mathsf{RP} = \mathsf{NP}$. *Quantum*, 7:1036, June 2023. doi:10.22331/q-2023-06-07-1036. 28, 30, 138, 147, 205

[LOLH22] Lorenzo Leone, Salvatore F. E. Oliviero, Seth Lloyd, and Alioscia Hamma. Learning efficient decoders for quasi-chaotic quantum scramblers, 2022. arXiv:2212.11338. 102, 105

[Low09] Richard A. Low. Learning and testing algorithms for the clifford group. *Phys. Rev. A*, 80:052314, Nov 2009. doi:10.1103/PhysRevA.80.052314. 27, 42, 147

[Lyu05] Vadim Lyubashevsky. The parity problem in the presence of noise, decoding random linear codes, and the subset sum problem. In Chandra Chekuri, Klaus Jansen, José D. P. Rolim, and Luca Trevisan, editors, *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pages 378–389, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. doi:10.1007/11538462_32. 180, 202

[Mer07] N. David Mermin. *The probability of success in Simon's problem*, page 187–188. Cambridge University Press, 2007. doi:10.1017/CBO9780511813870.015. 59

[Mon17] Ashley Montanaro. Learning stabilizer states by Bell sampling, 2017. arXiv:1707.04012. 8, 20, 22, 24, 48, 49, 52, 53, 57, 58, 104, 122

[MY22a] Tomoyuki Morimae and Takashi Yamakawa. One-Wayness in Quantum Cryptography, 2022. arXiv:2210.03394. 124

[MY22b] Tomoyuki Morimae and Takashi Yamakawa. Quantum Commitments And Signatures Without One-Way Functions. In *Advances in Cryptology – CRYPTO 2022: 42nd Annual International Cryptology Conference*, page 269–295, 2022. doi:10.1007/978-3-031-15802-5_10. 124

[MZ94] Stéphane Mallat and Zhifeng Zhang. Matching Pursuit with Time-Frequency Dictionaries. *IEEE Transactions on Signal Processing*, 41:3397–3415, 1994. doi:10.1109/78.258082. 89, 99

[NC02] Michael A. Nielsen and Isaac Chuang. Quantum Computation and Quantum Information, 2002. doi:10.1017/CBO9780511976667. 30, 32, 33

[O'D14] Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. doi:10.1017/CBO9781139814782. 49, 75, 179

[OW16] Ryan O'Donnell and John Wright. Efficient quantum tomography. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '16, page 899–912, New York, NY, USA, 2016. Association for Computing Machinery. doi:10.1145/2897518.2897544. 20, 100

[OW17] Ryan O'Donnell and John Wright. Efficient quantum tomography ii. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, page 962–974, New York, NY, USA, 2017. Association for Computing Machinery. doi:10.1145/3055399.3055454. 20

[Pie12] Krzysztof Pietrzak. Cryptography from learning parity with noise. In *International Conference on Current Trends in Theory and Practice of Computer Science*, pages 99–114. Springer, 2012. doi:10.1007/978-3-642-27660-6_9. 180

[PRR06] Michal Parnas, Dana Ron, and Ronitt Rubinfeld. Tolerant property testing and distance approximation. *Journal of Computer and System Sciences*, 72(6):1012–1042, 2006. doi:10.1016/j.jcss.2006.03.002. 73

[QAS21] Yihui Quek, Srinivasan Arunachalam, and John A Smolin. Private learning implies quantum stability. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S.

Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 20503–20515. Curran Associates, Inc., 2021. URL: https://proceedings.neurips.cc/paper/2021/file/abdbeb4d8dbe30df8430a8394b7218ef-Paper.pdf. 150, 174, 175, 198

[QPG21] Hammam Qassim, Hakop Pashayan, and David Gosset. Improved upper bounds on the stabilizer rank of magic states. *Quantum*, 5:606, December 2021. doi:10.22331/q-2021-12-20-606. 104

[RAS$^+$19] Andrea Rocchetto, Scott Aaronson, Simone Severini, Gonzalo Carvacho, Davide Poderini, Iris Agresti, Marco Bentivegna, and Fabio Sciarrino. Experimental learning of quantum states. *Science advances*, 5(3):eaau1946, 2019. doi:10.1126/sciadv.aau1946. 175

[RB00] Robert Raussendorf and Hans J. Briegel. Quantum computing via measurements only, 2000. doi:10.48550/arxiv.quant-ph/0010033. 19

[Rey20] Lev Reyzin. Statistical queries and statistical algorithms: Foundations and applications. *arXiv preprint arXiv:2004.00557*, 2020. arXiv:2004.00557. 171, 172

[RLCK19] Patrick Rall, Daniel Liang, Jeremy Cook, and William Kretschmer. Simulation of qubit quantum circuits via Pauli propagation. *Phys. Rev. A*, 99:062337, 2019. doi:10.1103/PhysRevA.99.062337. 104, 185

[Roc18] Andrea Rocchetto. Stabiliser states are efficiently PAC-learnable. *Quantum Information & Computation*, 18(7-8):541–552, 2018. doi:10.26421/qic18.7-8-1. 25, 27, 28, 145, 146, 148, 151, 152, 168, 171, 173, 175, 201, 202

[RS96] Ronitt Rubinfeld and Madhu Sudan. Robust Characterizations of Polynomials with Applications to Program Testing. *SIAM Journal on Computing*, 25(2):252–271, 1996. doi:10.1137/S0097539793255151. 73

[SBM06] Vivek V. Shende, Stephen S. Bullock, and Igor L. Markov. Synthesis of quantum-logic circuits. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 25(6):1000–1010, 2006. doi:10.1109/TCAD.2005.855930. 123

[Sch30] Erwin Schrödinger. The Uncertainty Relations in Quantum Mechanics. Zum Heisenbergschen Unschärfeprinzip. *Sitzungsberichte der Preussischen Akademie der Wissenschaften, Physikalisch-mathematische Klasse*, 14:296–303, 1930. 40

[Shi03] Yaoyun Shi. Both toffoli and controlled-not need little help to do universal quantum computing. *Quantum Info. Comput.*, 3(1):84–92, jan 2003. doi:10.26421/qic3.1-7. 101

[Sho95] Peter W. Shor. Scheme for reducing decoherence in quantum computer memory. *Phys. Rev. A*, 52:R2493–R2496, 1995. doi:10.1103/PhysRevA.52.R2493. 19

[Sho99] Peter W. Shor. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Review*, 41(2):303–332, 1999. doi:10.1137/S0036144598347011. 148

[SXL$^+$17] Chao Song, Kai Xu, Wuxin Liu, Chui-Ping Yang, Shi-Biao Zheng, Hui Deng, Qiwei Xie, Keqiang Huang, Qiujiang Guo, Libo Zhang, et al. 10-qubit entanglement and parallel logic operations with a superconducting circuit.

Physical Review Letters, 119(18):180511, 2017. doi:10.1103/PhysRevLett.119.180511. 100

[TTT06] Etsuji Tomita, Akira Tanaka, and Haruhisa Takahashi. The worst-case time complexity for generating all maximal cliques and computational experiments. *Theoretical Computer Science*, 363(1):28–42, 2006. doi:10.1016/j.tcs.2006.06.015. 24, 95

[Val84] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984. doi:10.1145/1968.1972. 10, 25, 89, 143, 148, 169

[Val85] LG Valiant. Learning disjunction of conjunctions. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence-Volume 1*, pages 560–566, 1985. URL: https://dl.acm.org/doi/abs/10.5555/1625135.1625242. 181

[VDB21] Ewout Van Den Berg. A simple method for sampling random clifford operators. In *2021 IEEE International Conference on Quantum Computing and Engineering (QCE)*, pages 54–59, 2021. doi:10.1109/QCE52317.2021.00021. 168

[vEB81] Peter van Emde Boas. Another NP-complete partition problem and the complexity of computing short vectors in lattices. *TR*, 1981. URL: https://books.google.com/books?id=tCQiHQAACAAJ. 98

[Ved06] Vlatko Vedral. *Introduction to Quantum Information Science*. Oxford University Press, 09 2006. doi:10.1093/acprof:oso/9780199215706.001.0001. 17

[Web16] Zak Webb. The clifford group forms a unitary 3-design. *Quantum Information and Computation*, 16(15&16):1379–1400, nov 2016. doi:10.26421/qic16.15-16-8. 98, 202

[WJS+98] Gregor Weihs, Thomas Jennewein, Christoph Simon, Harald Weinfurter, and Anton Zeilinger. Violation of bell's inequality under strict einstein locality conditions. *Phys. Rev. Lett.*, 81:5039–5043, Dec 1998. doi:10.1103/PhysRevLett.81.5039. 17

[Yog19] Mithuna Yoganathan. A condition under which classical simulability implies efficient state learnability, 2019. arXiv:1907.08163. 175, 201