JOHANNES KEPLER
UNIVERSITY LINZ

Rigorous and non-asymptotic theory support for near-term quantum computers

Assist.-Prof. Dr. Richard Kueng, MSc ETH

ORCID id: 0000-0002-8291-648X

*A thesis submitted to obtain the **Venia Docendi** in the area of*

Theoretical Computer Science

Johannes Kepler University, Linz, Austria

Abstract

Quantum computing has entered an interesting transient regime. Current quantum computers are becoming large and precise enough to outperform even the best conventional supercomputers at certain tasks. But, at the same time, they are still too small and too noisy to execute prototypical quantum algorithms. This era of near-term quantum computers comes with novel challenges, as well as opportunities. Well-established scientific approaches do not necessarily apply anymore. For instance, numerical simulations on conventional hardware have worked well for studying small quantum architectures of the past. But, these simulations are becoming too demanding for today's devices. Conversely, an asymptotic runtime analysis may reveal a quantum advantage in the limit of very large problem sizes. But near-term devices do not allow us to scale up to this far.

In this cumulative thesis, we showcase how to overcome near-term challenges and harness quantum advantages based on mathematically rigorous theory. Our method-oriented approach leads to assertions that are based on mathematical proofs (i.e. they are guaranteed to remain valid if we scale-up the number of qubits), but also non-asymptotic (i.e. we typically obtain actual numbers that are small enough to be meaningful for today's quantum computers). It combines techniques from theoretical computer science, mathematics, as well as (quantum) information theory. We showcase different aspects of this general approach, as well as its broad range of their applicability, by means of four exemplary research projects.

Acknowledgments

Much of the research that ultimately led to this thesis was conducted during my stay at Caltech. I am greatly indebted to my advisors John Preskill, Fernando G.S.L. Brandão and Joel A. Tropp for creating a truly unique scientific environment that fosters curiosity, exchange of ideas, collaboration and a pure joy for science without ulterior motives. These three have also been, and still are, amazing collaborators. The same is true for Hsin-Yuan (Robert) Huang, an exceptionally talented junior researcher, long-time collaborator and trusty friend. I am also grateful for the ties of friendship that formed with Philipp Walk, Victor Albert, Angelo Lucia, Burak Sahinoglu, Kamyar Azizzadenesheli, Nicholas Hunter-Jones, Philippe Faist, Eugene Tang, Nikola Kovachki, Tomas Jochym-O'Connor, Alex Dalzell, and many more, during my stay in Pasadena. I would also like to thank all my project partners around the globe for inspiring collaborations with great learning effect on my side.

Upon my return to Austria, I was fortunate enough to become affiliated with the Institute for Integrated Circuits (IIC) at the Johannes Kepler University, Linz. My colleagues there made me feel very welcome from the get go and helped me getting started in the local academic microcosm. This includes Robert Wille, Lukas Burgholzer, Stefan Hillmich, Thomas Grurl, Doris Nikolaus and Timm Ostermann. Head of the institute, Robert Wille, has also shielded me from much bureaucratic overhead which allowed me to truly focus on basic research. I also want to thank him for excellent advice and support throughout the past year and a half and am happy about several IIC-collaborations that pinpoint fruitful synergies between the empirical skills of the existing team and my more theoretical mindset.

Furthermore, I am grateful for the advice and encouragement I have received from my new, Linz-based, colleagues Armando Rastelli, Armin Biere (by now in Freiburg), Sepp Hochreiter, Oliver Bimber, Martina Seidl, Johannes Kofler, Wolfgang Schreiner, Alois Zoitl, Moritz Brehm, Ronny Ramlau, Johannes Brandstädter, and many more. This gratitude also extends to my new, Innsbruck-based, collaborators Barbara Kraus, Hannes Pichler and Peter Zoller, as well as my established collaborators David Gross, Jens Eisert and Felix Krahmer in Germany, but also Chris Ferrie in Australia.

I also appreciate the companionship that formed with my tenure-track colleague, Günter Klambauer, throughout the course of writing our habilitation theses. The exchange of information and feedback has been both productive and encouraging.

Finally, starting a new position in an unknown scientific environment at the onset of the COVID19 pandemic has not always been easy. I want to thank my (Swiss-based) girlfriend Lolita Ammann and my my (largely Austrian-based) core family – Erik, Felix, Gabriela and Josef Küng – for being there for me (and keeping up with me) throughout these extraordinary times.

To A.Univ.-Prof. DI Dr. Josef Küng
father, mentor and role model

Contents

Chapter 1

Introduction and summary of results

1.1 Motivation

Quantum computers are not the next generation of supercomputers. Rather, they are an entirely new type of computing hardware based on the rules of quantum mechanics – the laws of nature that govern physical systems at microscopic scales (e.g. on the level of individual atoms). And, although well-understood, these rules are radically different from everyday experience. Concepts like superposition, entanglement and, to some extent, true randomness occur naturally at these scales, but do not have macroscopic counterparts. It is these effects that render quantum mechanical problems challenging; both from a conceptual and a practical perspective. Many quantum-mechanical problems are notoriously difficult to solve, even for the largest supercomputers to date. Problems of paramount importance in material science, chemistry and pharmaceutics fall into this category. Quantum computers attempt to use quantum-mechanical effects in order to execute (certain) computations much faster than classical hardware ever could. Nowadays, the underlying vision is attributed to R. Fenyman [Fey82], who said "Nature isn't classical, dammit, and if you want to make a simulation of nature, you'd better make it quantum mechanical, and by golly it's a wonderful problem, because it doesn't look so easy."

This already hints at one of the most groundbreaking prospect of quantum computers: the accurate simulation of microscopical systems, e.g. to finally construct high-temperature superconductors, or ab initio quantum chemistry. In the 1990s (more than a decade after Feynman had shared his vision) researchers started to discover that quantum computers might also be able to solve certain combinatorial problems much faster than any known classical algorithms. These developments culminated in Shor's polynomial-time quantum algorithms for factoring and discrete logarithm [Sho94] – two combinatorial problems for which the best known classical algorithms have runtime exponential in system size. Polynomial-time algorithms for these particular problems could have far reaching implications for security. Many widespread cryptographic protocols, like RSA encryption or the Diffie-Hellman key exchange protocol, are built on the (conjectured) hardness of these number-theoretic problems. Since then, about 65 problems have been identified for which quantum computers do offer a noteworthy advantage. These are tabulated and explained in the *Quantum Algorithm*

Zoo[1]. Other prominent quantum algorithms include Grover's search algorithm in unstructured databases [Gro97], as well as faster algorithms for solving linear systems of equations [HHL09] and convex optimization [BS17, vAGGdW17a, BKF19].

For quite some time, these seminal insights were exclusively theoretical in nature and quantum advantages have been identified via a thorough mathematical analysis of runtime and memory requirements. But, the advent of ever larger and ever more accurate quantum hardware platforms is starting to change the field [Pre18]. The quantum computing platforms of today and the near future, so-called near-term devices, are becoming too large to simulate with conventional supercomputers. Doing so would incur an exponential overhead in memory and/or runtime. Google's sycamore chip, for example, works with 53 qubits – the fundamental carriers of quantum information. This translates into $4^{53} \approx 8.12 \times 10^{31}$ classical degrees of freedom; an astronomical number that is too large for even the largest supercomputers to handle. But, at the same time, these devices are still too small and too noisy to actually run any of the quantum algorithms mentioned above. Although we know, in principle, how to eventually build a fully-functional digital quantum computer (with negligible noise corruption), such devices are not yet on the horizon.

Near-term quantum computers do, however, seem large and intricate enough to do interesting stuff. Promising use cases are hybrid quantum-classical algorithms to heuristically solve the ground state problem in quantum chemistry (the Variational Quantum Eigensolver aka VQE) [PMS$^+$14, CAB$^+$20], nontrivial combinatorial problems (the Quantum Approximate Optimization Algorithm aka QAOA) [FGG14] like finding the maximum cut in a graph, as well as approaches to simulate the behavior of other quantum physical systems (quantum simulation), see e.g. [GAN14] and references therein. But despite plenty of activity and enthusiasm, rigorous evidence for an actual quantum advantage is very limited. Our understanding of hybrid quantum-classical algorithms is still in its infancy and there is plenty of room for improvements. In this thesis, we collect several mathematically rigorous contributions that address one or more of these challenges. We also put these results into a broader context.

Roadmap: The rest of this introductory chapter is organized as follows. Sub. 1.2 introduces the standard template for near-term (and far-term) quantum computers. Important challenges are identified in Sub. 1.3, where we also discuss novel ways to overcome them. In Sub. 1.4, we switch gears and focus on opportunities. We present rigorous theory contributions to *quantum advantage* (see also Chapter 4), *Variational Quantum Eigensolvers* (see also Chapter 3 and Chapter 5) and *quantum algorithms for optimization* (see also Chapter 6) and put them into context.

1.2 Hybrid quantum-classical computers

Understanding how a quantum computer actually works is not that easy. Although more successful than any other physical theory, quantum mechanics does not have the reputation of being either simple,

[1] `https://quantumalgorithmzoo.org`

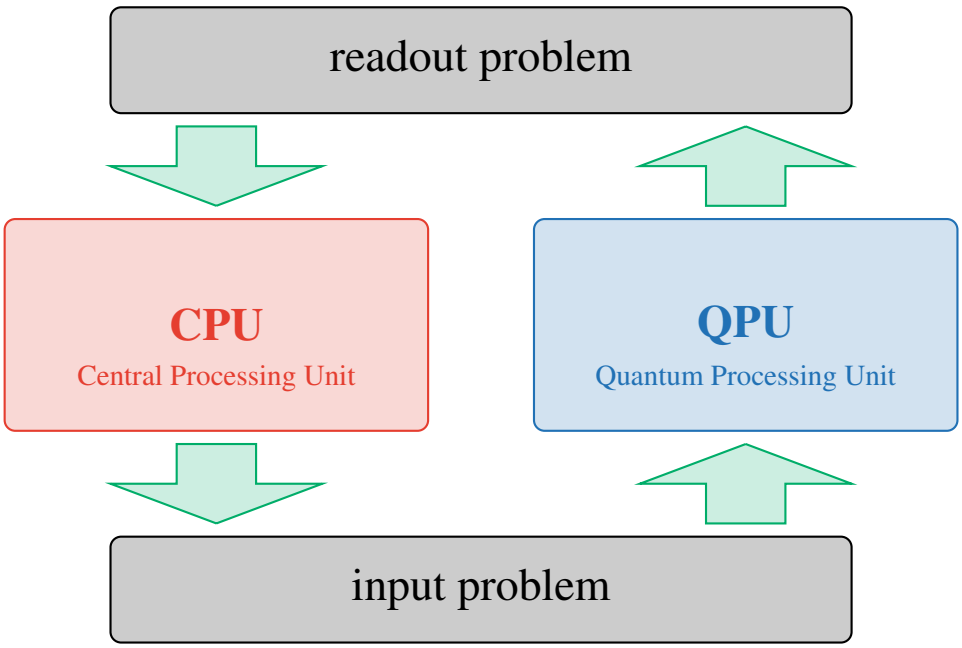


Figure 1.1: *Schematic illustration of a hybrid quantum-classical computer:* A conventional Central Processing Unit (CPU) can outsource certain computational task to a Quantum Processing Unit (QPU). The resulting hybrid architecture combines the strengths of both hardware platforms, but also suffers from information-transmission bottlenecks (input problem and readout problem).

or intuitive. Some quantum mechanical effects are responsible for the astonishing power of quantum computers, while other effects again limit their potential considerably. Balancing these blessings and curses against each other to still obtain a net gain is surprisingly tricky. And, as a result, we actually do not know that many problems for which quantum computers offer an unconditional (mathematically rigorous) advantage.

In order to get a first intuition about quantum computers, a high-level comparison with conventional hardware can be helpful. The core of most current computing devices is a central processing unit (CPU). It can be tasked to carry out any possible set of instructions we throw at it, but is not necessarily good at computing specific things (a jack of all trades, master of none). This is where alternative processing units come in. One important example are graphical processing units (GPUs). They are designed to solve specialized mathematical operations, in this case large matrix matrix multiplications, much more efficiently than traditional CPUs. The original motivation for this setup is computer graphics, but GPUs are also well-suited for training neural networks and simulating macroscopic physical systems.

However, even GPUs struggle with the excessive number of mathematical operations that would be required to accurately simulate physical and chemical processes beneath the nanoscale. Problems of this type occur naturally in material science (e.g. the search for high-temperature superconductors), pharmaceutics and chemistry (e.g. ab initio drug design) and fundamental physics (e.g. probing exotic field theories or constructing time crystals). All these problems have on thing in common. They adhere to the rules of quantum mechanics. And this renders them extremely difficult to handle with classical (in the sense of macroscopic; not quantum mechanical) computations and hardware. Hence,

it would be great if we had a different type of processing unit that is capable of handling these kind of problems. This is the conceptual origin of quantum computers [Fey82], but the term Quantum Processing Unit (QPUs) captures the intended purpose more accurately. QPUs are not designed to supersede conventional hardware (like CPUs or GPUs), but are specialized processing units that can further augment computing power. The result is a *hybrid quantum-classical computer*, schematically illustrated in Figure 1.1. This combination produces a completely new and different type of computing architecture that comes with novel opportunities, but also novel challenges. We refer to standard textbooks [Wat18, NC00] or the lecture notes [Kue19] composed by the applicant for further reading.

1.3 Challenges for hybrid quantum-classical computers

A hybrid quantum-classical computer unifies two types of radically different hardware. The CPU, on the one hand, uses registers comprised of bits encoded into electric currents. In turn, logical and arithmetic operations are realized by utilizing the laws of classical electromagnetism. The QPU, on the other hands, has registers comprised of qubits that are encoded into states of microscopic systems. Computation is based on physical interactions between qubits that are engineered utilizing the laws of quantum mechanics. This necessarily leads to information-transmission bottlenecks at the interface between the two processing units. This section discusses two aspects of this problem that are qualitatively very different.

1.3.1 The input problem

The CPU must be able to delegate computing tasks to the QPU. Since QPUs can only handle certain types of computation, this may involve converting the original task into a compatible form (synthesis). Subsequently, this standardized computation is decomposed into a sequence of more elementary operations (mapping) which can then be executed on the actual quantum hardware. Needless to say, conversion and decomposition should produce sequences of elementary quantum operations that are as short and noise-resilient as possible. The proceedings [BKW21, GKFW21, HKMW21] and preprints [CHKT20, FSK$^+$21, BWK20], co-authored by the applicant, address different aspects of and helpful subroutines for addressing these types of (input) problems.

We have chosen to introduce the input problem from a very practical, hardware-oriented, perspective. But there is also a conceptual dimension that is vital for quantum algorithm design. The precise workings of information access and storage, e.g. oracle access to input parameters that specify a given problem, can have a huge impact on the runtime of quantum algorithms. For instance, the development of intricate classical data structures [Tan19, Tan18, GLT18] have recently nullified widely-believed exponential advantages promised by quantum algorithms for recommendation systems [KP20], principal component analysis [LMR14] and clustering [LMR13], as well as stochastic regression [RSML18]. These aspects of the input problem also play an important role in Sub. 1.4.3 (see also Chapter 6), where

we discuss a quantum algorithm for convex optimization with provable speed-ups over the best known classical algorithms.

1.3.2 The readout problem

Once the QPU completes its task, the outcome of a quantum computation is stored within a register comprised of qubits, not conventional bits. And the laws of quantum mechanics severely restrict access to this type of quantum information. In order to retrieve any kind of actionable advice, measurements must be performed on the microscopic constituents that make up the quantum register. Alas, quantum mechanics dictates that informative measurements must be destructive ("collapse of the wavefunction"). A typical measurement of an n-qubit register produces a string of n conventional outcome bits, but also destroys the register in the process. What is more, these outcome bits themselves are random variables and concrete realizations do not carry any information by themselves. Instead, the actual result of the QPU computation is stored in the distribution over all possible outcome strings, not the actual realizations ("god does play dice"). This, in turn, implies that many identical repetitions of a given quantum computation are required to obtain sufficient statistics about this distribution of outcome bits and the actual result encoded within. The readout problem is an actual bottleneck that severely restricts the application range of hybrid quantum-classical computers. The number of repetitions required to readout sufficiently accurate QPU solutions typically grows with QPU size and necessarily slows down each quantum-classical cycle.

In Ref. [HKP20] we present and analyze a novel solution to this problem – the first of four journal publications that form the main part of this thesis (see also Chapter 3). The key idea is to repeatedly use randomized measurements to construct a succinct classical approximation of the underlying quantum system. This *classical shadow* can then be used to efficiently approximate (up to) exponentially many properties of the underlying quantum system – an exponential improvement over existing methods that is optimal in the sense that it saturates fundamental bounds from information theory. Further improvements are possible for readout problems with additional structure. We will discuss one important example in Section 1.4.2 below.

1.3.3 Perspective

It can not be overstated that qubits, the fundamental carriers of quantum information, are extremely delicate and hard to control. Stringent levels of precisions are required to correctly initialize a QPU and, subsequently, execute nontrivial computations. Several powerful tools for certification and characterization of quantum hardware have been developed to address these challenges, see e.g. [EHW$^+$20] for a recent overview which also discusses the journal articles [GKKT20, KLDF16, KKEG19, RKK$^+$18] co-authored by the applicant.

But accurate calibration can only go so far. The extremely fragile nature of qubits, as well as their analog degrees of freedom, imply that a perfect QPU cannot exist under realistic conditions. Some errors and noise fluctuations are inevitable. And, over the course of a long quantum computation, these

errors add up until the accumulated noise overpowers the underlying quantum signal, rendering the entire computation useless. Noise accumulation limits both the size and runtime of trustworthy QPU computations and is the limiting factor for today's hybrid quantum-classical computers [Pre18]. Fortunately and crucially, there are proposals on how to eventually overcome these limitations. *Quantum error correcting codes* distribute the state of individual logical qubits redundantly among a collection of many physical qubits, see e.g. [NC00]. This allows for protecting the encoded quantum information from essentially all kinds of errors, provided that each of them is sufficiently small (and they are not correlated in a malicious fashion). Moreover, proposals exist on how to process quantum information directly on the logical level, provided that the average error per elementary quantum operation is below a certain threshold. This leads to *fault-tolerant quantum computation*, the key stepping stone to construct fully scalable QPUs that are powerful enough to execute big, digital quantum algorithms like Shor's algorithms [Sho94] for factoring and discrete logarithm, as well as the HHL algorithm [HHL09] for solving linear systems. It is these quantum algorithms that promise exponential runtime savings over the best known existing algorithms, with far-reaching implications for security, data analysis and optimization. Having said this, fault-tolerant quantum computing is still a rather distant dream and years, perhaps even decades, of dedicated effort will be required to achieve the required level of control, accuracy and correction. The quest for building a scalable, fully functional QPU is a marathon, not a sprint.

1.4 Opportunities for near-term QPUs

We have seen that the quest for building large QPUs that are functional, as well as trustworthy, is an extremely ambitious goal. In recent years, important milestones have been achieved and several big tech companies (e.g. Amazon, Google, IBM), as well as startups (e.g. AlpineQuantum, IonQ, PsiQuantum and Rigetti) have announced ambitious plans for the near future. Paralleling these developments, fundamental research at academic institutions is also stronger than ever.

Still, scalable fault-tolerant quantum processors are unlikely to be available for years to come. In turn, the most prominent use cases for quantum computing, like polynomial-time algorithms for factoring and discrete logarithm [Sho94], quadratically faster search algorithms [Gro97] and exponentially faster linear system solvers [HHL09], are also off the table, at least for the foreseeable future. And this begs the question: what should we actually do with current and near-term quantum architectures? This section collects mathematically rigorous evidence for three different use cases.

We first introduce the quantum circuit model which then allows us to discuss *quantum advantage* (formerly also called quantum supremacy). For the first time, programmable QPUs are able to solve certain problems much faster than the best conventional supercomputers available. We then move on to discuss the *Variational Quantum Eigensolver* (VQE), a quantum-classical heuristic to solve challenging ground state problems in material science, quantum chemistry and physics. And finally, we present a hybrid quantum-classical algorithm for solving *convex optimization* problems more efficiently than the best known classical solvers.

1.4.1 Quantum circuits, quantum advantage and quantum complexity

The prevalent model for quantum computation is the *quantum circuit model*. It is a generalization of the Boolean circuit model used in theoretical computer science and chip design, see e.g. [AB09, Chapter 6] and [HH13]. QPUs work with n-qubit registers (the fundamental carriers of information in a conventional n-bit register are replaced by their quantum mechanical counterparts). In the quantum circuit model, each qubit is visually represented by a horizontal line. A quantum computation is a sequence of (elementary) quantum gates. These are reversible transformations that only affect a constant number of qubits each. And we read these instructions from left to right. We refer to Figure 1.2 for a visual illustration of a 15-qubit circuit comprised of, in total, 98 elementary 2-qubit gates (blue boxes) that are arranged in a brickwall geometry. The blue boxes are placeholders that can be replaced with any 2-qubit gate.

Crucially, the set of elementary quantum transformations is strictly larger than the set of elementary reversible transformations in conventional logic. This larger expressiveness is where the power of QPUs hails from. Elementary quantum transformations can be combined to produce more complicated quantum circuits. In fact, one of the fundamental results of quantum computation asserts that *any* quantum mechanical process that involves n qubits can be accurately approximated by a quantum circuit comprised only of elementary quantum gates, see e.g. [BBC$^+$95]. This includes all conceivable quantum computations, and the set of all reversible n-bit circuits is a strict subset thereof. Recall that conventional circuits, like those executed in a CPU, can be mapped to reversible circuits at the cost of (at most) a polynomial number of extra bits. This showcases that quantum circuits can, at least in theory, be at least as powerful as conventional hardware. In fact, the seminal algorithms by Shor [Sho94] (factoring and discrete logarithm), Harrow, Hassidim and Lloyd [HHL09] (fast linear system solver) and others do indicate that they are strictly more powerful.

An important summary parameter of quantum circuits is *depth*. That is, the number of steps required to execute all elementary gates that make up the circuit (after parallelization). Similar to runtime in the Turing machine model, circuit depth is a measure of cost associated with executing a given computation. The shorter the depth, the easier the associated quantum circuit. In fact, very shallow (i.e. constant-depth) quantum circuits are so easy that they can be efficiently simulated on conventional hardware, see e.g. [Vid03, BGM21, CC20]. On the other end of this spectrum are exceedingly wide (i.e. exponential-depth) quantum circuits that can become so complex that even a fully-functional QPU would require millions of years to sequentially execute all layers of elementary gates. The quantum-mechanical processes behind such circuits are too time-consuming and complicated to ever occur in nature.

The sweet spot for quantum computing lies between these two circuit-depth extremes. Powerful use cases, like Shor's algorithm or HHL, translate into quantum circuits whose circuit depth scales polynomially in the number of qubits. In stark contrast, the best known conventional algorithms translate into conventional circuits of (worst-case) exponential depth – an exponential quantum advantage. Alas, even the most accurate QPUs of today are far too noisy to reliably explore the regime of polynomial-depth quantum circuits. Gate errors, which are inevitable for today's devices,

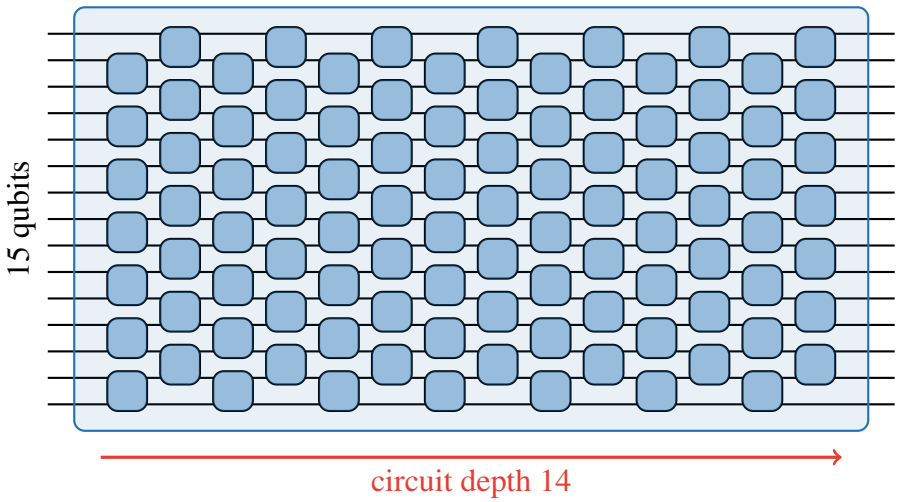


Figure 1.2: *Illustration of a quantum circuit diagram:* Qubits are represented by horizontal lines that are meant to be read from left to right. The little blue boxes are placeholders for elementary quantum gates that only ever act on two neighboring qubits. We can combine many of these elementary quantum gates to execute more complicated, global quantum computations. The circuit geometry depicted here illustrates a so-called *brickwork quantum circuit*. The circuit depth counts the total number of gate layers. It is an important summary parameter that tells us how complicated/expensive a given quantum computation is.

quickly accumulate and can overpower the actual signal. As a consequence, the circuit depth of near-term QPU computations needs to be shallow. And this raises the question of *quantum advantage*: Is there any computing problem (useful or useless), where near-term QPUs outperform conventional supercomputers?

It is widely believed that there is an affirmative, but also somewhat boring, answer: QPUs are much better at simulating themselves than conventional supercomputers could ever hope to be. The Google team used this type of reasoning to claim quantum advantage in 2019 [AAB$^+$19]. They used their 53-qubit QPU to execute random quantum circuits with depth up to 20 much faster than the world's most powerful supercomputer of the time: 200 seconds vs. at least 2.5 days[2].

But how can we be certain that conventional supercomputers must perform so much worse? On the one hand, there are credible obstructions from the theory of computational complexity. Roughly speaking, the ability of efficiently simulating a random quantum circuit computation on conventional hardware would lead to a collapse of the polynomial hierarchy at the third level (a hypothetical relation between different computational complexity classes that is widely believed to be false), see e.g. [AA11, BJS11].

On the other hand, there are also more practical considerations that support a quantum advantage. Several techniques have been developed that attempt to simulate quantum circuits on conventional hardware. And while some of them are designed to exploit latent structure, they all require an exponential overhead in the number of qubits once the circuits become too deep (c.f. [Vid03, BGM21, CC20] and/or lacks advantageous structure (c.f. [ZW19, WHB21, BBC$^+$19]). Importantly the circuits

[2]The original Google paper mentions 10,000 years, but IBM called these claims excessive and listed techniques that conventional supercomputers could use to achieve the task in 2.5 days, see `https://www.ibm.com/blogs/research/2019/10/on-quantum-supremacy/`. Up to our knowledge, the actual simulation has not been carried out yet.

used to demonstrate quantum advantage are comprised of elementary 2-qubit gates that are sampled uniformly at random from the collection of all possible 2-qubit gates. This has a serious downside: the resulting circuit executes a completely random, and therefore utterly useless, computation. But, on the upside, there can be no advantageous structure that classical simulators could exploit. And circuit depths of order 20 also do seem sufficient to derail short-depth simulation techniques.

Note, however, that the final argument only addresses quantum circuits that are incompressible. Indeed, suppose that it were possible to accurately approximate a quantum circuit functionality with another quantum circuit that is much shallower. Then, we could apply short-depth simulators, like [BGM21, CC20], to this compressed circuit and effectively nullify the presumed quantum advantage. Intuitively, such noteworthy compressions seem unlikely, but how can we be sure? Perhaps we were not clever enough to think of an ingenious shortcut that allows us to represent the underlying functionality much more efficiently. It is not possible in practice to enumerate all the quantum circuits that approximate another circuit to find one of minimal size. For that reason, it is quite difficult to obtain a useful lower bound on the minimal circuit size.

Yet it is possible, to derive powerful lower bounds for ensembles of random circuits, which hold with high probability when concrete circuits are selected from these ensembles. This is the main contribution of Ref. [BCHJ$^+$21] (see also Chapter 4) – the second of four journal publications that form the main part of this thesis. The key idea is to link shortest possible circuit depth, also called *quantum complexity*, to pseudorandom properties of the underlying random circuit ensemble. The stronger the pseudorandomness, the longer the circuit depth required to realize it. In a second step, we can then relate pseudorandomness to the depth of random quantum circuits with certain geometry constraints, like random brickwall circuits illustrated in Figure 1.2. Together, these arguments imply a direct relation between actual and minimal circuit depth. With extremely high probability (over the choice of individual 2-qubit gates), the shortest possible circuit depth can only be polynomially smaller than the actual depth of the original random circuit. In other words: substantial compressibility is extremely unlikely for random circuits. This insight supplies further evidence that the random quantum circuits used to demonstrate quantum advantage are indeed very hard to simulate on conventional hardware.

We find it worthwhile to point out that the study of quantum complexity also has implications beyond quantum computing. In quantum many-body physics, the shortest possible circuit depth required to prepare ground-state wave functions is used to classify topological phases of matter at zero temperature [CGW10]. Quantum complexity has recently also become a popular subject in high-energy physics, where complexity growth is conjectured to be related to the long-time growth of the interior of an eternal black hole [Sus16a, SS14, Sus16b]. These are encouraging synergies, where concepts and insights from quantum computing drive progress in other, seemingly unrelated, scientific communities.

1.4.2 Variational Quantum Eigensolver (VQE)

We have seen that QPUs can natively run computations that would require exponentially more resources on conventional hardware. In the last subsection, we have also pinpointed the reason: computing properties of a n-body quantum system does typically require (order) 2^n memory and runtime. This curse of dimensionality quickly becomes prohibitively expensive. QPUs, on the other hand, have the potential to bypass this issue entirely. This opens up new and interesting possibilities for solving quantum mechanical problems in material science, many-body physics and chemistry.

A prototypical example problem is the *ground state problem* in quantum many-body physics. The input is a spatial configuration of n qubits (spins), e.g. a one-dimensional chain or a two-dimensional lattice, as well as an energy function, called Hamiltonian, that is typically a sum of simple nearest-neighbor interactions: $H = \sum_{\langle ij \rangle} h_{ij}$, where $\langle i, j \rangle$ with $1 \leq i, j \leq n$ runs over pairs of qubits that are adjacent to each other. What is more, each nearest-neighbor interaction h_{ij} is simple and can be succinctly represented by a $2^2 \times 2^2$ matrix, because each of them only affects 2 qubits at a time. The total n-qubit Hamiltonian, however, is much more complicated. Mathematically, it is a self-adjoint matrix with 2^n rows and 2^n columns. The ground state problem asks for identifying the smallest eigenvalue of this enormous matrix:

$$\underset{\psi \in \mathbb{C}^{2^n}}{\text{minimize}} \quad \langle \psi, H\psi \rangle \quad \text{subject to} \quad \langle \psi, \psi \rangle = 1 \qquad \text{(ground state problem).} \tag{1.1}$$

Here, $\langle x, y \rangle = \bar{x}^T y = \sum_i \bar{x}_i y_i$ denotes the canonical inner product on the complex-valued vector space $\mathbb{C}^{2^n}$. The smallest possible value is called the ground state energy. Finding it, is not intrinsically difficult. Computing the eigenvalue decomposition of H determines $E_0 = \lambda_{\min}(H)$ and, by extension, solves Eq. (1.1) in a runtime that is (at most) cubic in matrix size. The problem is that H is an exponentially large matrix to begin with.

The variational method for computing ground state energies replaces general 2^n-dimensional vectors ψ by a family of ansatz vectors $\psi(\theta)$ that only depend on a polynomial number of (real-valued) parameters $\theta \in \mathbb{R}^m$ with $m = \text{poly}(n)$. Subsequently, we vary these parameters to minimize energy over this family of ansatz vectors:

$$\underset{\theta \in \mathbb{R}^m}{\text{minimize}} \quad \langle \psi(\theta), H\psi(\theta) \rangle \quad \text{subject to} \quad \langle \psi(\theta), \psi(\theta) \rangle = 1 \qquad \text{(variational method).}$$

By construction, any variational method produces upper bounds on the true ground state energy. The quality of approximation depends on the family of ansatz functions $\theta \mapsto \psi(\theta)$ and the way we update the parameters to (hopefully) approach a minimum. Indeed, there is a trade-off. Variational ansatz functions do suppress the degrees of freedom enormously, but they also complicate the optimization landscape. The function $\theta \mapsto \langle \psi(\theta), H\psi(\theta) \rangle$ typically has many local minima, as well as a large number of saddle points. But, the most glaring problem is that evaluating the objective function may still require matrix-vector multiplications in 2^n dimensions. Traditionally, this challenge is overcome by carefully selecting a family of ansatz vectors that plays nicely with the total Hamiltonian so that each $\langle \psi(\theta), H\psi(\theta) \rangle$ can be evaluated at $\text{poly}(n)$ cost. Prominent examples are the density matrix

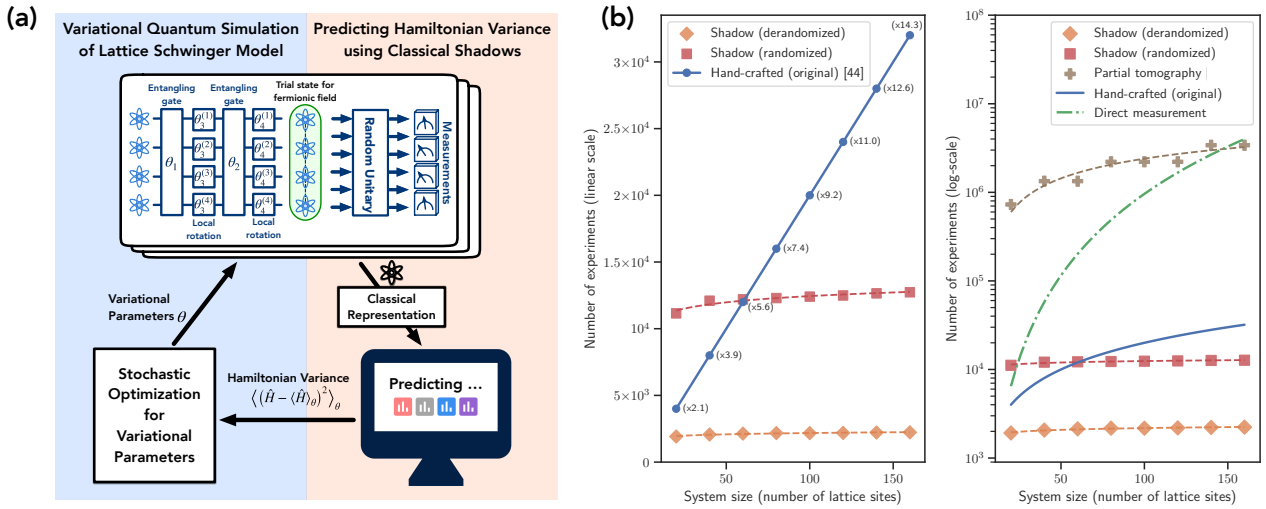


Figure 1.3: *Variational Quantum Eigensolver (VQE):* **(a)** Illustration for the particular case of simulating one-dimensional quantum electrodynamics (lattice Schwinger model) [KMvB$^+$19]. **(b)** Cost of converting relevant QPU result into actionable advice for a single variational parameter update. The plots show the number of QPU evaluations followed by measurement required to accurately estimate the cost function as a function of QPU size (number of qubits) for direct measurements (green), partial tomography (brown) [BMBO20], the original hand-crafted method (blue) [KMvB$^+$19], as well as classical shadows (red) [HKP20] and their derandomized counterpart (orange) [HKP21a].
This figure is imported from Ref. [HKP20], see also Chapter 3.

renormalization group formalism (DMRG) in many-body physics and the Hartree-Fock method in quantum chemistry.

The *variational Quantum Eigensolver* (VQE) instead uses a QPU to evaluate $\theta \mapsto \langle \psi(\theta), H\psi(\theta)\rangle$ directly. We refer to Ref. [CAB$^+$20] for a thorough introduction and important references. Outsourcing these computations to a QPU can circumvent the curse of dimensionality, but for a different family of ansatz vectors. Namely, those $\psi(\theta) \in \mathbb{C}^{2^n}$ that characterize the state of an n-qubit quantum register after a variational quantum circuit has been executed. This is visualized in Fig. 1.3 (left). The m variational parameters $\theta = (\theta_1, \ldots, \theta_m)$ specify simple single-qubit gates within a relatively shallow quantum circuit geometry. After encoding the parameters in a quantum circuit, we generate the associated ansatz state by applying the circuit to a fixed, and typically simple, input quantum state. So, $\psi(\theta)$ is an actual quantum state comprised of n qubits. Importantly, the preparation cost is only proportional to the circuit depth $\lceil 2(m/n) \rceil \ll 2^n$. We can repeatedly prepare this ansatz state to estimate $\langle \psi(\theta), H\psi(\theta)\rangle$ by decomposing $H = \sum_{\langle ij \rangle} h_{ij}$ into its elementary constituents and approximating each $\langle \psi(\theta), h_{ij}\psi(\theta)\rangle$ by direct quantum measurements. Approximately knowing $\langle \psi(\theta), H\psi(\theta)\rangle$ then allows us to execute a stochastic optimization for the variational parameters θ to obtain a new ansatz state that achieves smaller energy (Strictly speaking, we may actually have to estimate $\langle \psi(\theta), H\psi(\theta)\rangle$ for an entire collection of parameters to approximate, for instance, a stochastic gradient descent step). And multiple iterations of this hybrid quantum-classical update strategy yield better and better ansatz functions that hopefully converge to the true ground state energy.

By construction, many VQE iterations are required to (hopefully) converge to the ground state en-

ergy. And the cost of each iteration is actually dominated by the cost of approximating $\langle\psi(\theta), H\psi(\theta)\rangle$ for fixed parameters θ up to sufficient accuracy. A naive readout procedure, where we estimate individual terms h_{ij} in the Hamiltonian one after the other, scales linearly in the number of terms. And although polynomial in n, this overhead can quickly become a real bottleneck. The r.h.s. of Fig. 1.3 illustrates this for a VQE designed to probe physical theories for quantum electrodynamics [KMvB$^+$19]. There, the authors actually developed a specialized measurement procedure to efficiently estimate all terms of the problem-specific Hamiltonian. But, the associated cost still scales polynomially in QPU size n. This, in turn, limited their demonstration to $n = 9$ qubits, even though the quantum platform they used could have readily supported up to 20 qubits (and more).

This scaling problem can be overcome by developing better solutions to the readout problem. The works [HKP20] (see also Chapter 3) and [HKP21a] (see also Chapter 5) achieve just that. These two articles constitute two main pillars of this thesis. Ref. [HKP20] introduces and analyzes a novel quantum-to-classical converter based on randomly rotating the individual qubits just prior to measurement. Visualized in Fig. 1.3 (left), randomization ensures that the subsequent measurement can access all possible directions of the 2^n-dimensional quantum state space. And they do so in a random, yet unbiased fashion. This paves the way for Monte Carlo approximation: repeatedly perform randomized single-qubit quantum measurements (each applied to a freshly generated quantum state) and approximate the underlying quantum state by empirical averaging over the observed measurement outcomes. We call this approximation a *classical shadow* of the underlying quantum system. It is possible to rigorously prove that the number of randomized measurements required to accurately approximate a collection of $\mathrm{poly}(n)$ quantum state properties only scales logarithmically in n. Applying this to VQE can yield an exponential speed-up for the quantum-to-classical readout required within each iteration. Already (order) $\log(n)$ randomized measurements suffice to accurately approximate all $\mathrm{poly}(n)$ simple energy terms $\langle\psi(\theta), h_{ij}\psi(\theta)\rangle$ simultaneously. These numbers can then be combined to approximate $\langle\psi(\theta), H\psi(\theta)\rangle = \sum_{\langle ij\rangle}\langle\psi(\theta), h_{ij}\psi(\theta)\rangle$ orders of magnitudes faster than traditional methods. This scaling improvement is visualized in Fig. 1.3 (right) for the VQE problem considered in Ref. [KMvB$^+$19].

The same plot also showcases that additional improvements are possible if we derandomize the originally randomized measurement protocol behind classical shadows. Derandomization is a powerful procedure from theoretical computer science that can convert randomized algorithms into deterministic ones [MR95, AS08]. This general principle allows us to iteratively replace initially random single-qubit measurements with fixed deterministic measurement choices. We refer to Ref. [HKP21a] for details, see also Chapter 5 below. This greedy assignment procedure can be executed very efficiently on conventional hardware and effectively optimizes the quantum measurement procedure for the type of VQE Hamiltonian at hand. Runtime and memory scale linearly in the number of qubits and the number of target functions (which is optimal). We also prove that the resulting deterministic measurement protocol is guaranteed to perform at least as well as the randomized one. Complementary empirical studies paint an even more favorable picture. Fig. 1.3 (right), for instance, showcases a consistent improvement of about one order-of-magnitude.

Other VQE use cases yield even larger advantages for derandomization. This includes, in particular, the electronic ground state problem in quantum chemistry. There, one is interested in accurately computing the ground state energy of a small- to medium-size molecule. This is a challenging problem, because quantum mechanical effects (such as the Pauli exclusion principle for electrons) must be taken into account. It is, however, possible to encode the molecular energy function in a synthetic many-body Hamiltonian $H = \sum_i h_i$, much like the ones introduced and discussed above [JW28, BK02]. Once again, this Hamiltonian is comprised of only poly(n) terms h_i. Each of them is simple, but in contrast to before, their range is not confined to a small subset of qubits. This can cause problems for randomized measurement procedures, because the likelihood of obtaining useful outcome statistics for predicting $\langle \psi(\theta), h_i \psi(\theta) \rangle$ diminishes exponentially with the support size of h_i. Consequently, the original classical shadows protocol performs exceptionally poorly for these types of quantum chemistry Hamiltonians. Modifications of the original protocol, like biasing the originally random measurement settings [HBRM20, Had21, HHR$^+$21], do address this issue and improve the readout stage considerably. Derandomization may be viewed as another modification. Numerically, we observe that derandomization can keep up with these recent developments and often performs even better than biased, but still random, measurement techniques.

Before moving on, we find it worthwhile to point out that no rigorous performance guarantees have yet been established for VQE. On the contrary, it is possible to show that the classical optimization required in each iteration is intrinsically hard [BK21]. Complexity-theoretic obstructions for VQE are not too surprising. After all, the many-body ground state problem is known to be at least as hard as the satisfiability problem (SAT) and is probably even harder[3]. A provably efficient VQE solution for arbitrary ground state problems would therefore imply that the problem class NP is contained in BQP, the class of all problems that can efficiently be solved with high probability on a fully-functional quantum computer. And this inclusion is widely believed to be false. (We don't believe that quantum computers are able to solve NP-complete problems efficiently.) On the other hand, reductions of hard problem instances produce very particular, and typically even unphysical, Hamiltonians. And it is entirely possible that VQE does perform very well for more physically motivated problems that have advantageous structure. Numerical simulations, as well as proof-of-principle demonstrations on small-scale hardware do seem to point in this direction, see [CAB$^+$20] and references therein. For the time being, we should regard VQE as a heuristics, albeit a very interesting one.

1.4.3 Quantum algorithms for convex optimization

The most prominent quantum algorithms promise exponential quantum advantages over the best known conventional algorithms. But already polynomial speedups can make a substantial difference, especially if the best-known classical runtime scales like a bad polynomial in input size. Certain optimization algorithms fall into this category. An optimization problem is *convex* if it corresponds to

[3] It is possible to show that every problem whose solution can be efficiently checked on a quantum computer can be reduced to an instance of the ground state problem. This identifies the ground state problem as the quantum computing analog of the satisfiability problem [KSV02].

minimizing a convex function over a convex set of feasible solutions. Roughly speaking, convexity of the objective function ensures that every local minimum is also a global minimum, while convexity of the feasible set implies that iterative solvers don't get stuck near the boundary. Together, these desirable features ensure that most, though not all, convex optimization problems can be solved in polynomial runtime. We refer to standard textbooks [BV04, Bar02] for a thorough discussion.

Many important optimization problems can be rephrased as, or relaxed to, a convex optimization problem. Concrete examples include portfolio optimization [MOS14], network flow problems [GH62], constrained entropy maximization [Jay57], sparse [CRT06, Don06] and low-rank regression models [RFP10, Gro11, KRT17], phase retrieval [CSV13, GKK17], binary matrix decompositions [KT21], but also (optimal) relaxations of combinatorial optimization problems, such as finding the maximum cut in a graph (MAXCUT) [GW95]. This list is far from exhaustive.

Accurate general-purpose solvers for these types of optimization problems can scale with the fifth power of the input size. Although polynomial, this scaling quickly become prohibitively expensive. A seminal work by Brandão and Svore [BS17] addressed this polynomial scaling problem by introducing a novel quantum meta-algorithm for solving matrix-valued optimization problems with a positive semidefiniteness constraint (i.e. every feasible matrix must be self-adjoint and can only have nonnegative eigenvalues). A thorough theoretical analysis yields (worst-case) runtime guarantees that scale much more favorably with input size, but at the cost of a worse scaling an approximation accuracy. See also Ref. [vAGGdW17b] for a more thorough and improved analysis. Roughly speaking, the underlying idea is as follows. Absorb the positive semidefiniteness constraint by representing every feasible point as the matrix exponential of another self-adjoint matrix: $X \leftarrow \exp(-H)$ (recall that matrix exponentials have nonnegative eigenvalues by construction). Subsequently, we iteratively update the matrix exponent H to penalize directions where the current iterate $X = \exp(-H)$ is very far from being feasible or very far from being optimal. See Fig. 1.4 for a visualization of a related meta-algorithm that solves a semidefinite feasibility problem (i.e. determine wether the intersection of several convex sets is non-empty). Remarkably, it is possible to prove that this procedure must converge after a number of iterations that only scales logarithmically in problem size. This ensures that the total runtime is dominated by the cost for executing the update rule. And there, the main bottleneck is computing the matrix exponential $X = \exp(-H)$. For $d \times d$ matrices, this may require up to $\mathcal{O}(d^3)$ arithmetic operations on conventional hardware. Brandão and Svore realized that a QPU can sometimes perform matrix exponentiation more efficiently than a conventional CPU. A quantum subroutine called Gibbs sampling [TOV$^+$11, CS17] is capable of producing quantum states that encode structured matrix exponentials much more efficiently. This quantum advantage translates into highly competitive runtimes of the resulting hybrid quantum-classical algorithm. What is more, the QPU subroutine seems to be more stable with respect to quantum circuit imperfections than more traditional quantum algorithms, like Grover or Shor. Moreover, the number of required qubits only scales logarithmically in input size. These two features indicate that fast quantum solvers for convex optimization may constitute and interesting and useful application of near-term quantum architectures.

However, the Brandão-Svore algorithm is not perfect. The actual runtime scales rather unfavorably

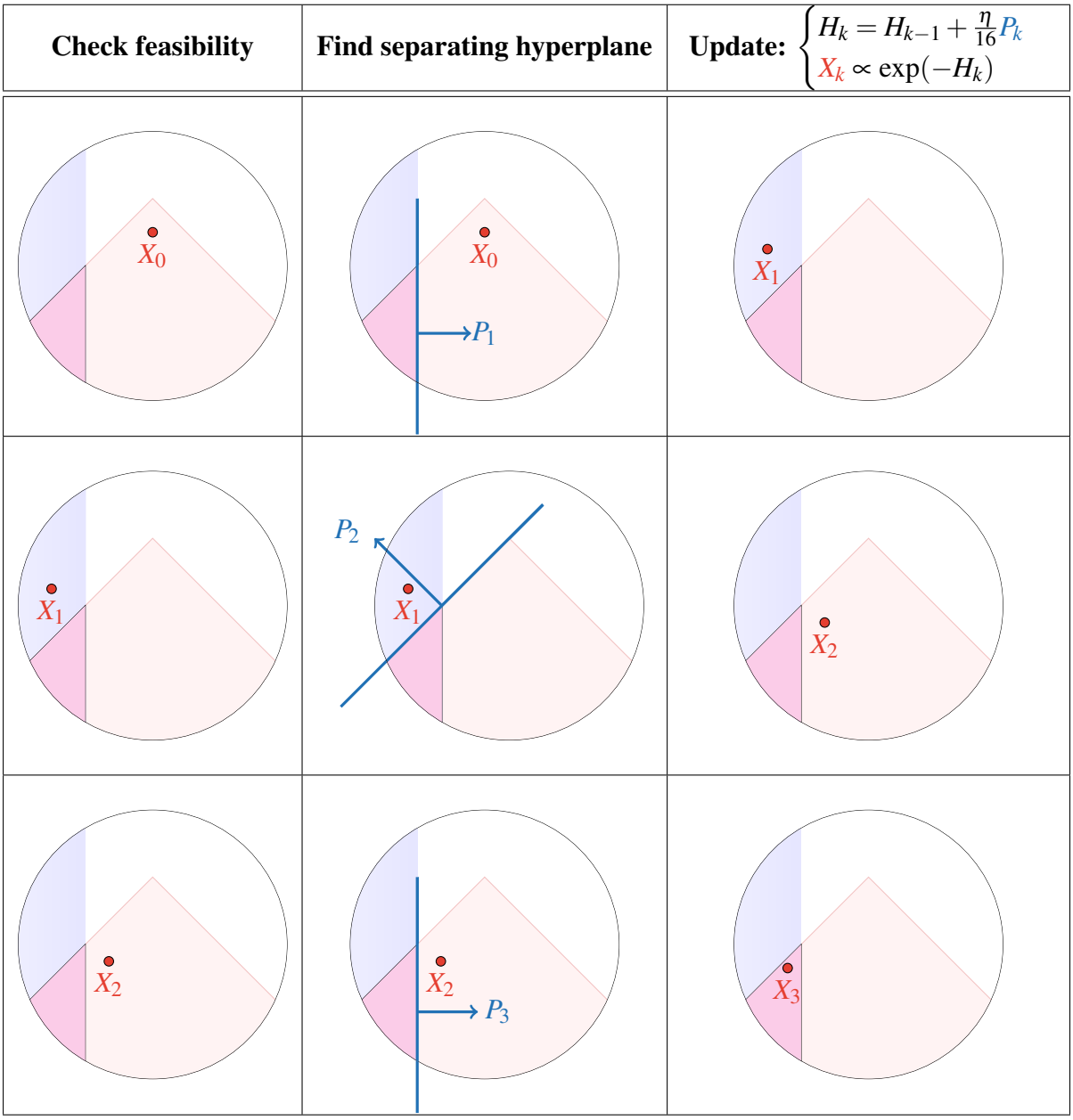


Figure 1.4: *Illustration of a meta-algorithm for convex optimization:* We visualize a semidefinite feasibility problem. The task is to find a point in the intersection of the set of positive semidefinite matrices (illustrated as a white circle), a linear half space (blue) and a non-linear wedge (orange). We first perform a change of variables $X_k \leftarrow \exp(-H_k)$ that ensures that we always stay within the set of positive semidefinite matrices (white circle). We then execute the following update rules iteratively: (i) check if the current iterate $X_k = \exp(-H_k)$ is contained in both the half space and the wedge. If this is the case, we have found a point in the intersection and are done. If this is not the case, there must be a hyperplane (blue line) that separates X_k from one of the sets. The update rule $H_{k+1} = H_k + \eta P_k$, where η is a small step size, introduces a penalty for this direction of violation, such that the next iterate $X_{k+1} = \exp(-H_{k+1})$ will be closer to the convex set in question.
In Ref. [BKF19] (see also Chapter 6) we rigorously prove that this procedure must terminate after only a logarithmic number of steps. The most expensive subroutine is computing the matrix exponentials $X_k = \exp(-H_k)$. Outsourcing this subroutine to a QPU would result in a noteworthy quantum speed-up.

with a problem-specific parameter, called the width of the optimization problem. And for many interesting use cases, like relaxations of combinatorial optimization problems and spin ground state problems, the problem width is so large that it negates any potential for a quantum advantage. In Ref. [BKF19], we address this problem by developing a new overarching meta-algorithm that is better suited for dealing with these types of problem geometries. Illustrated in Fig. 1.4, this meta-algorithm is already competitive on conventional hardware. But, similarly to the original Brandão-Svore hybrid algorithm, an additional quantum advantage is possible by delegating the computation of matrix exponentials to a QPU. As a result, we obtain rigorous runtime guarantees that outperform the best known algorithms for generic MAXCUT and spin glass problem instances. This work is the last of four journal publications that form the main part of this thesis[4].

The framework we develop turns out to be remarkably flexible and can be adapted to cover other use cases. In Ref. [FBK21], we use similar ideas to develop a procedure for reconstructing classical descriptions of quantum systems in a resource-optimal fashion. Numerical studies suggest that a conventional and naive implementation of this algorithms is already very competitive. Empirically, we find that the algorithm performs much better than the theory suggests.

1.4.4 Perspective

The last subsections illustrated three promising use cases for existing and near-term quantum processing units. In Sub. 1.4.1 we have illustrated how existing quantum processors, like Google's sycamore chip, can execute certain, admittedly contrived, computations much quicker than conventional hardware ever could (quantum advantage). The Variational Quantum Eigensolver (VQE), discussed in Sub. 1.4.2, attempts to harness this potential for solving challenging ground state problems in many-body physics and quantum chemistry. In Sub. 1.4.3, we finally discussed the possibility of speeding up solvers for certain convex optimization problems by executing hybrid quantum-classical algorithms that outsource crucial subroutines to a QPU for a net gain in runtime.

There are other near-term applications that we haven't discussed (yet). One of them is the *Quantum Approximate Optimization Algorithm (QAOA)* [FGG14]. Somewhat similar in spirit to VQE, this is a hybrid quantum-classical algorithm designed for (approximately) solving combinatorial optimization problems (MAXCUT). In contrast to VQE, rigorous theory support does exist for several important use cases, like finding the maximum cut in a (triangle-free) graph [FGG14]. Alas, these turn out to be weaker than the strongest convergence guarantees available for conventional algorithms[5]. Rigorous quantum speed-ups are currently not available for this algorithm, but empirical results do look promising.

Another potential application for near-term quantum algorithms is machine learning with *quantum-enhanced feature spaces*, see e.g. [HCT+19]. The underlying idea is to use quantum circuits to define

[4]Note that it is joint work with F. Brandão, one of the inventors of this type of hybrid quantum-classical algorithms. It also features in the Quantum Algorithm Zoo (`https://quantumalgorithmzoo.org`).

[5]Interestingly, these classical guarantees have only been established after Ref. [FGG14] pointed towards a rigorous quantum speedup for approximating MAXCUT.

novel types of feature maps for supervised and unsupervised learning. Coordinates $x_k \in \mathbb{R}$ of a data point $x = (x_1, \ldots, x_D) \in \mathbb{R}^D$ parametrize pre-specified single-qubit gates within a fixed circuit geometry (e.g. a single-qubit Pauli-X rotation by angle x_k, $U(x_k) = \exp(-\mathrm{i}x_k\sigma_X)$, embedded into a larger, n-qubit quantum circuit geometry). This produces an n-qubit quantum circuit $U(x_1, \ldots, x_D)$ that is applied to a fixed input quantum state. The resulting QPU state $\psi(x_i)$ implicitly describes a quantum feature vector with 2^n degrees of freedom. What is more, the QPU can also be used to compute similarity measures between feature vectors belonging to different data points. The most popular choice is the quantum embedding kernel $k_{\text{quantum}}(x,y) = |\langle \psi(x), \psi(y)\rangle|^2$ for data points $x, y \in \mathbb{R}^D$. The kernel trick then allows for integrating these quantum-enhanced similarity measures in a conventional learning algorithm, e.g. a support vector machine for binary classification.

Conceptually, quantum embedding kernels are reminiscent of VQE. We use QPUs as co-processors to compute functions that live in an exponentially large space. These computations are then used to execute powerful conventional algorithms. And, like with VQE, we are not yet aware of any rigorous quantum advantages. This, however, is a very active field of research – both from an empirical and a theoretical perspective. We refer to Refs. [ASZ+21, HBM+21] for interesting recent contributions.

1.5 Summary and outlook

In this introductory chapter, we have attempted to provide a broad and, whenever possible, nontechnical overview of the state of quantum computing in the year 2021. We have pointed out ultimate goals, like polynomial-time algorithms for factoring and discrete logarithm, but also emphasized the restricted applicability of quantum computers. They should be viewed as special-purpose co-processors, quantum processing units (QPU), that can be used to further empower conventional hardware. This motivates the concept of hybrid quantum-classical computers which comes with novel challenges as well as opportunities. The study of both is fascinating and timely. Unprecedented progress in the design and control of quantum architectures has led to the advent of the first interesting QPUs; interesting in the sense that they are too large to be (easily) simulated on conventional hardware.

In Section 1.3, we discussed (some of the) challenges that come with a hybrid architecture. Two bottlenecks concern information transmission. We must be able to efficiently map conventional instructions, like execute this quantum computation, onto the QPU (*input problem*). And, once a quantum computation has completed, we must also be able to access its result (*readout problem*). While the former problem bears conceptual similarities with conventional circuit design, the readout problem is plagued by genuine quantum effects like probabilistic measurement outcomes and wavefunction collapse. These effects can add up and lead to real bottlenecks. Sometimes, though not always, they can be overcome by a combination of randomized measurements (quantum hardware) and data processing (conventional software). This quantum-to-classical conversion procedure, known as *classical shadows*, has empowered many recent QPU applications.

In Section 1.4 we shifted our attention to opportunities instead. Near-term hybrid quantum-classical computers have been used to demonstrate *quantum advantage*: quantum computers can solve certain

very specialized, and admittedly also very useless, problems in milliseconds for which the world's largest supercomputers would need days at least. Near-term QPUs can also be used to execute the Variational Quantum Eigensolver (VQE), a hybrid quantum-classical heuristic to solve challenging ground state problems in quantum chemistry, many-body physics and material science. Other quantum-enhanced hybrid algorithms, like the Quantum Approximate Optimization Algorithm (QAOA), or quantum-enhanced kernels and neural networks utilize a similar set of ideas and have also attracted a lot of attention recently. Finally, we have also discussed how QPUs can be used to speed up powerful meta-algorithms for solving convex optimization problems. The required QPU subroutine, computing matrix exponentials, is more demanding than the quantum subroutines for VQE, QAOA and the like. But it is still a far cry away from the intricacy of traditional quantum algorithms, like those of Shor and Grover.

The advent of intermediate-scale (50 – 100 qubits) QPUs is in the process of changing quantum computing research. Until very recently, quantum computers have mostly been a theoretical concept. Likewise, progress and new insights have typically been the product of either rigorous mathematical arguments and/or simulations of (small-scale) quantum architectures on conventional hardware. But the new era also opens up new possibilities. Namely, empirical studies on hybrid quantum-classical computers. Google's 53-qubit chip, for instance, is capable of performing computations that we cannot hope to simulate on conventional supercomputers. And even larger quantum architectures are currently in the making. More so than ever, this potential has attracted attention and talent from other scientific communities. And it is difficult to tell, where this accumulation of new possibilities and talent will lead to in the upcoming years. Just one thing seems relatively certain: Fully functional, fault-tolerant, quantum computers will not be available for years to come. And it will require continued attention, focus and dedication to get there at all.

But, in the meantime, there is still a lot of potential for interesting developments. The intersection between *machine learning* (ML) and *quantum computing* (QC) looks like a particularly promising junction for the exchange of powerful ideas. Partly, because both quantum computing and machine learning seem to already be potent in their own right. But also, because both fields are fashionable and substantial attention has already gravitated towards them individually. Machine learning methods have already been successfully applied to problems in quantum physics, see [CT17, vNLH17, BCJ$^+$19]. And modern QPUs could only become as accurate as they currently are, because sophisticated (and expensive) machine learning procedures are used to calibrate them on a daily basis. In the converse direction, there has been much hope that QPUs may have the potential to speed up expensive subroutines within the training stage of a ML model exponentially, see e.g. [KP20, LMR14, LMR13, RSML18]. But, more recently, all these works have been superseded by conventional algorithms empowered by quantum-inspired data structures [Tan19, Tan18, GLT18, KP20]. As of now, these recent developments have nullified all known exponential quantum speed-ups for ML subroutines.

But there are other ways to combine quantum computing and machine learning. Quantum embedding kernels, discussed in Sec. 1.4.4, for instance, currently receive a lot of attention. Even more recently, a preprint co-authored by the applicant proposes to use near-term QPUs in order to generate

training data that can subsequently be used to empower conventional ML models [HKT+21]. This approach combines the core strengths of both fields: QPUs are good at simulating quantum physical properties while (conventional) machine learning models, like support vector machines or neural networks, excel at learning and generalizing from training data. The quantum-to-classical converters introduced in Sub. 1.3.2 (see also Chapter 3) provide a sufficiently strong link between the two realms and actually allows us to rigorously prove powerful convergence guarantees that point towards a novel window of opportunity, as well as potential quantum advantages. The companion paper [HKP21b] highlights that such a combination is surprisingly powerful. There, we prove that even a fully quantum ML algorithm running on a fully-functional quantum computer with an arbitrary number of qubits could not substantially outperform this near-term ML setup in terms of training data size and average prediction error. Exponential runtime improvements may still be possible, though.

To summarize: synergies between quantum-mechanical experiments (simulated on a QPU) and conventional ML (executed on conventional hardware) can be even more powerful than one might think [HKP21b]. And, moreover, they can even be equipped with rigorous performance guarantees [HKT+21] that scale (quasi-) polynomially in quantum system size. These quantum machine learning ideas are still at an early stage. But, we are confident that they open up completely new possibilities for near-term quantum computers that can be backed up by rigorous and non-asymptotic theory support.

Chapter 2

List of publications and core contributions

2.1 List of publications

Disclaimer: the applicant is a proponent of alphabetical author ordering. Whenever authors appear in alphabetical order, this order does not necessarily reflect actual contributions.

2.1.1 Publications included in this thesis

1. [HKP20] H.Y. Huang, R. Kueng, J. Preskill, Predicting many properties of a quantum system from very few measurements, *Nature Physics* **16**:1050–1057 (2020)

2. [BCHJ$^+$21] F.G.S.L. Brandão, W. Chemissany, N. Hunter-Jones, R. Kueng, J. Preskill, Models of quantum complexity growth, *PRX Quantum* **2**:030316 (2021) [editor's suggestion]

3. [HKP21a] H.Y. Huang, R. Kueng, J. Preskill, Efficient estimation of Pauli observables by derandomization, *Physical Review Letters* **127**:030503 (2021)

4. [BKF19] F.G.S.L. Brandão, R. Kueng, D. Stilck França, Faster quantum and classical SDP approximations for quadratic binary optimization, to appear in *Quantum* (2021)

2.1.2 Other publications during candidature

Journal articles (peer-reviewed)

5. CF. Chen, HY. Huang, R. Kueng, JA. Tropp, Quantum simulation via randomized product formulas: Low gate complexity with accuracy guarantees, to appear in *PRX Quantum* (2021)

6. [HKP21b] H.Y. Huang, R. Kueng, J. Preskill, Information-theoretic bounds on quantum advantage in machine learning, *Physical Review Letters* **126**:190505 (2021) [editor's suggestion]

7. [KT21] R. Kueng and J.A. Tropp, Binary Component Decomposition Part I: The Positive-Semidefinite Case, *SIAM Journal on Mathematics of Data Science* **3**(2):544–572 (2021)

8. [EKH+20] A. Elben, R. Kueng H.Y. Huang, R. van Bijnen, C. Kokail, M. Dalmonte, P. Calabrese, B. Kraus, J. Preskill, P. Zoller, B. Vermersch, Mixed-state entanglement from local randomized measurements, *Physical Review Letters* **125**:200501 (2020)

9. [GKKT20] M. Guţă, J. Kahn, R. Kueng J.A. Tropp, Fast state tomography with optimal error bounds, *Journal of Physics A* **53**:204001 (2020)

10. [KMV19] R. Kueng D.G. Mixon, S. Villar, Fair redistricting is hard, *Theoretical Computer Science* **791**:28–35 (2019)

11. [KKEG19] M. Kliesch, R. Kueng J. Eisert, D. Gross, Guaranteed recovery of quantum processes from few measurements, *Quantum* **3**:171 (2019)

12. [JKM19] P. Jung, R. Kueng, D.G. Mixon, Derandomizing compressed sensing with combinatorial design, *Frontiers in Applied Mathematics and Statistics* **5**:26 (2019)

13. [RKK+18] I. Roth, R. Kueng S. Kimmel, Y.K. Liu, D. Gross, J. Eisert, M. Kliesch, Recovering quantum gates from few average gate fidelities, *Physical Review Letters* **121**:170502 (2018) [editor's suggestion]

Conference proceedings (peer-reviewed)

14. [FBK21] F.G.S.L. Brandão, R. Kueng, D. Stilck França, Fast and robust quantum state tomography from few basis measurements, *Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC)* (2021)

15. [GKFW21] T. Grurl, R. Kueng, J. Fuß, R. Wille, Stochastic quantum circuit simulation using decision diagrams, *Design, Automation and Test in Europe (DATE) Conference* (2021)

16. [HKMW21] S. Hillmich, R. Kueng, I.L. Markov, R. Wille, As Accurate as Needed, as Efficient as Possible: Approximations in DD-based Quantum Circuit Simulation, *Design, Automation and Test in Europe (DATE) Conference* (2021)

17. [BKW21] L. Burgholzer, R. Kueng R. Wille, Random stimuli generation for the verification of quantum circuits, *Asia and South Pacific Design Automation Conference (ASP-DAC)* (2021)

18. [RFK+18] I. Roth, A. Flinth, R. Kueng J. Eisert, G. Wunder, Hierarchical restricted isometry property for Kronecker product measurements, *56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (2018)

Book chapters (peer-reviewed)

19. T. Fuchs, D. Gross, P. Jung, F. Krahmer, R. Kueng, D. Stöger, Proof methods for robust low-rank matrix recovery, to appear in *Compressed Sensing in Information Processing (CoSIP)* (2021)

In the pipeline (under review)

20. T. Surawy-Stepney, J. Kahn, R. Kueng, M. Guţă, Projected Least-Squares Quantum Process Tomography, under review at *PRX Quantum* (2021)

21. H.Y. Huang, R. Kueng, G. Torlai, V.V. Albert, J. Preskill, Provably efficient machine learning for quantum many-body problems, under review at *Science* (2021)

22. T. Fuchs, D. Gross, P. Jung, F. Krahmer, R. Kueng, D. Stöger, Sketching with Kerdock's crayons: Fast sparsifying transforms for arbitrary linear maps , under review at *SIAM Journal on Matrix Analysis and Applications* (2021)

23. A. Neven, J. Carrasco, V. Vitale, C. Kokail, A. Elben, M. Dalmonte, P. Calabrese, P. Zoller, B. Vermersch, R. Kueng, B. Kraus, Symmetry-resolved entanglement detection using partial transpose moments, accepted for publication in *npj Quantum Information* (2021)

24. PK. Faehrmann, M. Steudtner, R. Kueng, M. Kieferova, J. Eisert, Randomizing multi-product formulas for improved Hamiltonian simulation, under review at *Physical Review X* (2021)

25. V. Vitale, A. Elben, R. Kueng, A. Neven, J. Carrasco, B. Kraus, P. Zoller, P. Calabrese, B. Vermersch, M. Dalmonte, Symmetry-resolved dynamical purification in synthetic quantum matter under review at *Physcial Review X* (2021)

26. L. Burgholzer, R. Wille, R. Kueng, Characteristics of Reversible Circuits for Error Detection, under review at *Array* (2021)

27. D. Suess, N. Maraviglia, R. Kueng, A. Maïnos, C. Sparrow, T. Hashimoto, N. Matsuda, D. Gross, A. Laing, Rapid characterisation of linear-optical networks via PhaseLift, under review at *Optica* (2021)

28. A. Haim, R. Kueng, G. Refael, Variational-Correlations Approach to Quantum Many-body Problems, under review at *Physical Review B* (2020)

29. R. Kueng, JA. Tropp, Binary component decomposition Part II: The asymmetric case, under review at *Linear Algebra and its Applications* (2021)

2.2 Core contributions

Since his graduation in 2016, the applicant has co-authored more than 21 scientific articles that address a broad range of challenges and opportunities within the umbrella term of near-term quantum computing. Four among them have been selected to form the main body of this thesis. This subselection is far from complete and does not necessarily include the strongest or most influential contributions either. Instead, these four journal articles are intended to be exemplars for the applicant's method-oriented approach to research. In a first step, familiarity with the theory behind quantum computing and detailed knowledge about existing results in the literature are combined to identify well-posed problems that seem interesting, timely and, perhaps most importantly, solvable. Subsequently, advanced proof techniques from theoretical computer science (e.g. convex optimization, pseudorandomness, computational complexity and algorithm design), mathematics (e.g. high-dimensional probability, matrix analysis and representation theory) and information theory (e.g. communication complexity and channel coding) are used to make rigorous assertions. Whenever possible, complementary numerical studies are conducted to confirm the theory or to underscore the feasibility of a newly developed method.

2.2.1 Efficient quantum-to-classical converters

We present an efficient method for constructing an approximate classical description of a quantum system using randomized quantum measurements. Additional randomization turns out to substantially reduce the cost of converting quantum information into classical information. We prove that the number of measurements is independent of system size and even saturates fundamental lower bounds from information theory. Experiments (both in silico and in vitro) highlight the advantages, which can even be exponential, relative to previously known methods.

2.2.2 Incompressibility of generic quantum circuits

The concept of quantum complexity has far-reaching implications spanning theoretical computer science, quantum many-body physics, and high energy physics. The complexity of a quantum evolution (or computation) is defined as the size of the shortest quantum circuit that accurately approximates it. In quantum many-body physics, for instance, it s reasonable to expect that the chaotic many-body Hamiltonians generate time evolutions whose complexity grows linearly in time. Likewise, we expect that the complexity of a generic quantum circuit is directly proportional to circuit depth. In other words, we believe that these circuits are incompressible.

However, because it is hard to rule out short-cuts, it is notoriously difficult to derive lower bounds on quantum complexity. To go further, one may study more generic models of complexity growth. We prove that local random quantum circuits generate unitary transformations whose complexity grows linearly for a long time, mirroring the behavior one expects in chaotic quantum systems, verifying conjectures in the context of holography (AdS/CFT duality). Our results also lend credence to the

claim that random quantum circuits are extremely hard to simulate on conventional hardware (quantum advantage).

2.2.3 Improving near-term quantum algorithms by derandomization

The ground state problem, i.e. accurately compute the exact ground state energy of a medium-sized molecule, is one of the most fundamental problems in quantum chemistry. And it is a difficult problem, largely because quantum mechanical interactions between the electrons must be taken into account. And conventional methods struggle with this, despite decades of dedicated research and the availability of powerful supercomputers. Near-term hybrid quantum-classical computers offer a tantalizing alternative that is based on two steps: (i) encode the electronic energy function for the molecule of interest into a synthetic energy function on n qubits; (ii) use a Variational Quantum Eigensolver (VQE) to (hopefully) find the ground state energy of this synthetic n-qubit system. What is more, the encoded energy function is not arbitrary. It is a sum of many Pauli observables (i.e. multi-linear functions that come with a particularly rich, algebraic structure).

We consider the readout problem for precisely this type of VQE and propose an efficient derandomization procedure that iteratively replaces random single-qubit measurements with fixed Pauli measurements. By construction, the resulting deterministic measurement procedure is guaranteed to perform at least as well as the randomized one. In some cases, e.g. VQE for quantum chemistry, the derandomized procedure is substantially better than the randomized one. Numerical experiments highlight the advantages of our derandomized protocol over various previous methods for estimating the ground-state energies of small molecules.

Whenever applicable, derandomization can be viewed as an unconditional improvement of the quantum-to-classical converters discussed in Sub. 2.2.1.

2.2.4 Quantum algorithms for convex optimization

The most well-known quantum algorithms offer provable exponential speed-ups over the best known conventional algorithm. These algorithms and their use-cases are rare and far between, though. But provable polynomial speed-ups should not be underestimated either, especially for problems where the best known conventional algorithm scales with a large power of the input size. General purpose solvers for convex optimization fall into this category. Many important problems in combinatorial optimization, data analysis, but also logistics and finance, can be reduced (or relaxed) to instances of semidefinite programs – a special class of convex optimization problems.

We develop a hybrid quantum-classical algorithm for solving semidefinite relaxations of binary quadratic optimization problems. This class of relaxations for combinatorial optimization has so far eluded rigorous quantum advantages. For generic instances, we rigorously prove that our quantum solver gives a nearly quadratic speedup over the best known state-of-the-art algorithms. To achieve this, we develop a meta-algorithm based on iterative matrix exponent updates (a variant of mirror descent) that is to be executed on conventional hardware. Expensive subroutines, most notably computing a

matrix exponential in each iteration, are outsourced to a quantum co-processor. This subroutine is probably too demanding for today's quantum architectures. But, at the same time, it is much simpler, and more noise resilient, than most quantum algorithms. There is hope that a variant of this hybrid quantum-classical algorithm will become tractable in the not too distant future.

2.3 Noteworthy contributions outside quantum computing

The main focus of this thesis is quantum computing. This is a new, fascinating and interdisciplinary field of research that comes with unique challenges and opportunities. In recent years, most of the applicant's efforts and passion have been devoted to finding new ways to overcome quantum computing challenges and harness quantum advantages. But these methods and proof techniques readily extend to other timely research areas. In fact, the list of publications in Sec. 2.1 contains at least 8 scientific articles that address topics in computer science, wireless communication and data science. Similar to before, we choose to briefly illustrate scope and range of these non-quantum results by means of two examples.

2.3.1 Complexity-theoretic obstacles for fair districting

[KMV19] R. Kueng D.G. Mixon, S. Villar, *Fair redistricting is hard*, Theoretical Computer Science **791**:28–35 (2019)

Districting, also known as Gerrymandering, is one of many peculiarities of the US voting system. In regular time intervals, state governors have the power to redraw borders of voting districts within the entire state. This practice has been used by both Democrats and Republicans to either spread voters of the opposite party among as many districts as possible ("cracking") or to concentrate many voters of their own party into single voting districts ("packing"). The ultimate goal is to ensure that as many voting districts as possible have a majority of likeminded voters. And because the US president is not elected directly, but appointed by an Electoral College, this can make a big difference. Districting, for instance, has been the reason why Donald Trump won against Hillary Clinton in the 2016 election despite having less than 50% of American voters behind him.

In Ref. [KMV19], we approach Gerrymandering from a computational complexity perspective. More precisely, we consider the problem of *fair districting*: Given a distribution of voters, separate them into districts such that the ratio of districts where party A has a majority is proportional to the actual percentage of voters in favor of the same party. (We also impose some reasonable constraints on the size and shape of districts allowed). Subsequently, we show that the question of deciding whether a fair districting exists is at least as difficult as solving the satisfiability problem. This in turn implies that fair districting is NP-hard in general. The proof follows by Karp reduction from planar 3-SAT.

2.3.2 Semi-discrete matrix factorization

[KT21] R. Kueng and J.A. Tropp, Binary Component Decomposition Part I: The Positive-Semidefinite Case, *SIAM Journal on Mathematics of Data Science*, **3**(2):544–572 (2021)

[KT19] R. Kueng, J.A. Tropp, Binary component decomposition Part II: The asymmetric case, under review at *Linear Algebra and its Applications* (2021)

A *matrix factorization* represents a given matrix B as the product of two more structured matrices: $B = VW^T$. In data analysis, matrix factorizations are an indispensable tool for exposing latent structure. What is more, the columns of V, often known as factors, and the columns of W, often called loadings, can also reveal important structural information. The celebrated Principal Component Analysis (PCA), for instance, follows from a special case of matrix factorization, namely the singular value decomposition. Such types of matrix factorization are known to always exist. And, what is more, we have efficient algorithms to compute them.

But PCA is not perfect. For instance, it requires that the factors (columns of V) must be orthogonal to each other. And, more often than not, this seems like an artificial constraint that forces factors to become dense vectors whose entries are floating point numbers with both positive and negative signs. This, in turn, makes it extremely difficult to interpret PCA factors. In Refs. [KT21, KT19], we develop the theoretical foundation of *binary component decompositions* (BCD): $B = SW^T$ and the entries of S are constrained to take values in the binary set $\{\pm 1\}$ (or $\{0, 1\}$). This matrix factorization model is appropriate when the underlying factors (columns of S) are meant to reflect an exclusive choice. Concrete examples are ‘yes’ and ‘no’ in survey data, ‘like’ and ‘dislike’ in collaborative filtering, ‘active’ and ‘inactive’ in genomics, etc.

Our research answers fundamental questions about existence and uniqueness of BCDs. Moreover, we also develop a tractable factorization algorithm that is guaranteed to succeed under a mild deterministic condition. This is remarkable, because the problem of computing BCDs looks like a daunting combinatorial optimization problem. Most structured matrix factorization problems are, in fact, NP hard in general.

Chapter 3

Efficient quantum-to-classical converters

or: Predicting many properties of a quantum system from very few measurements

Abstract

Predicting the properties of complex, large-scale quantum systems is essential for developing quantum technologies. We present an efficient method for constructing an approximate classical description of a quantum state using very few measurements of the state. This description, called a 'classical shadow', can be used to predict many different properties; order $\log(M)$ measurements suffice to accurately predict M different functions of the state with high success probability. The number of measurements is independent of the system size and saturates information-theoretic lower bounds. Moreover, target properties to predict can be selected after the measurements are completed. We support our theoretical findings with extensive numerical experiments. We apply classical shadows to predict quantum fidelities, entanglement entropies, two-point correlation functions, expectation values of local observables and the energy variance of many-body local Hamiltonians. The numerical results highlight the advantages of classical shadows relative to previously known methods.

Authors

Hsin-Yuan (Robert) Huang, Richard Kueng, John Preskill.

Journal

Nature Physics, **16**:1050–1057 (2020).

Confirmation of declaration of author contributions (Hsin-Yuan Huang)

Publication:

H.Y. Huang, R. Kueng, J. Preskill, Predicting many properties of a quantum system from very few measurements, *Nature Physics*, **16**:1050–1057 (2020)

Declaration of author contributions:

Hsin-Yuan Huang and Richard Kueng developed the theoretical aspects of this work. Hsin-Yuan Huang conducted the numerical experiments and wrote the open-source code. John Preskill conceived the applications of classical shadows. All authors wrote the manuscript.

Confirmation by co-author:

I confirm this declaration of author contributions, as well as my co-authorship.

Hsin-Yuan Huang

Hsin-Yuan Huang

Confirmation of declaration of author contributions (John Preskill)

Publication:

H.Y. Huang, R. Kueng, J. Preskill, Predicting many properties of a quantum system from very few measurements, *Nature Physics*, **16**:1050–1057 (2020)

Declaration of author contributions:

Hsin-Yuan Huang and Richard Kueng developed the theoretical aspects of this work. Hsin-Yuan Huang conducted the numerical experiments and wrote the open-source code. John Preskill conceived the applications of classical shadows. All authors wrote the manuscript.

Confirmation by co-author:

I confirm this declaration of author contributions, as well as my co-authorship.

John Preskill

ARTICLES
https://doi.org/10.1038/s41567-020-0932-7

nature physics

Predicting many properties of a quantum system from very few measurements

Hsin-Yuan Huang[1,2]✉, Richard Kueng[1,2,3] and John Preskill[1,2,4]

Predicting the properties of complex, large-scale quantum systems is essential for developing quantum technologies. We present an efficient method for constructing an approximate classical description of a quantum state using very few measurements of the state. This description, called a 'classical shadow', can be used to predict many different properties; order $\log(M)$ measurements suffice to accurately predict M different functions of the state with high success probability. The number of measurements is independent of the system size and saturates information-theoretic lower bounds. Moreover, target properties to predict can be selected after the measurements are completed. We support our theoretical findings with extensive numerical experiments. We apply classical shadows to predict quantum fidelities, entanglement entropies, two-point correlation functions, expectation values of local observables and the energy variance of many-body local Hamiltonians. The numerical results highlight the advantages of classical shadows relative to previously known methods.

Making predictions based on empirical observations is a central topic in statistical learning theory and is at the heart of many scientific disciplines, including quantum physics. For this latter field, predictive tasks, such as estimating target fidelities, verifying entanglement and measuring correlations, are essential for building, calibrating and controlling quantum systems. Recent advances in the size of quantum platforms[1] have pushed traditional prediction techniques—like quantum state tomography—to the limit of their capabilities. This is mainly due to the curse of dimensionality: the number of parameters needed to describe a quantum system scales exponentially with the number of its constituents. Moreover, these parameters cannot be accessed directly, but must be estimated by measuring the system. An informative quantum-mechanical measurement is both destructive (wavefunction collapse) and yields only probabilistic outcomes (Born's rule). Hence, many identically prepared samples are required to estimate accurately even a single parameter of the underlying quantum state. Furthermore, all of these measurement outcomes must be processed and stored in memory for subsequent prediction of relevant features. In summary, reconstructing a full description of a quantum system with n constituents (for example, qubits) necessitates a number of measurement repetitions exponential in n, as well as an exponential amount of classical memory and computing power.

Several approaches have been proposed to overcome this fundamental scaling problem. These include matrix product state (MPS) tomography[2] and neural network tomography[3,4]. Both require only a polynomial number of samples, provided that the underlying state has suitable properties. However, for general quantum systems, these techniques still require an exponential number of samples. See Supplementary Section 3 for details.

Pioneering a conceptually very different line of research, Aaronson[5] pointed out that demanding full classical descriptions of quantum systems may be excessive for many concrete tasks. Instead it is often sufficient to accurately predict certain properties of the quantum system. In quantum mechanics, interesting properties are often linear functions of the underlying density matrix ρ, such as the expectation values $\{o_i\}$ of a set of observables $\{O_i\}$:

$$o_i(\rho) = \mathrm{trace}(O_i\rho) \quad 1 \le i \le M \tag{1}$$

The fidelity with a pure target state, entanglement witnesses and the probability distribution governing the possible outcomes of a measurement are all examples that fit this framework. A nonlinear function of ρ, such as entanglement entropy, may also be of interest. Aaronson coined the term[5,6] 'shadow tomography' for the task of predicting properties without necessarily fully characterizing the quantum state, and he showed that a polynomial number of state copies already suffice to predict an exponential number of target functions. Although very efficient in terms of samples, Aaronson's procedure is very demanding in terms of quantum hardware; a concrete implementation of the proposed protocol requires exponentially long quantum circuits that act collectively on all the copies of the unknown state stored in a quantum memory.

In this Article, we combine the mindset of shadow tomography[5] (predict target functions, not the full state) with recent insights from quantum state tomography[7] (rigorous statistical convergence guarantees) and the stabilizer formalism[8] (efficient implementation). The result is a highly efficient protocol that learns a minimal classical sketch S_ρ—the classical shadow—of an unknown quantum state ρ that can be used to predict arbitrary linear function values (equation (1)) by a simple median-of-means protocol. A classical shadow is created by repeatedly performing a simple procedure: apply a unitary transformation $\rho \mapsto U\rho U^\dagger$, and then measure all the qubits in the computational basis. The number of times this procedure is repeated is called the 'size' of the classical shadow. The transformation U is randomly selected from an ensemble of unitaries, and different ensembles lead to different versions of the procedure that have characteristic strengths and weaknesses. In a practical scheme, each ensemble unitary should be realizable as an efficient quantum circuit. We consider random n-qubit Clifford circuits and tensor products of random single-qubit Clifford circuits as important

[1]Institute for Quantum Information and Matter, California Institute of Technology, Pasadena, CA, USA. [2]Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, USA. [3]Institute for Integrated Circuits, Johannes Kepler University Linz, Linz, Austria. [4]Walter Burke Institute for Theoretical Physics, California Institute of Technology, Pasadena, CA, USA. ✉e-mail: hsinyuan@caltech.edu

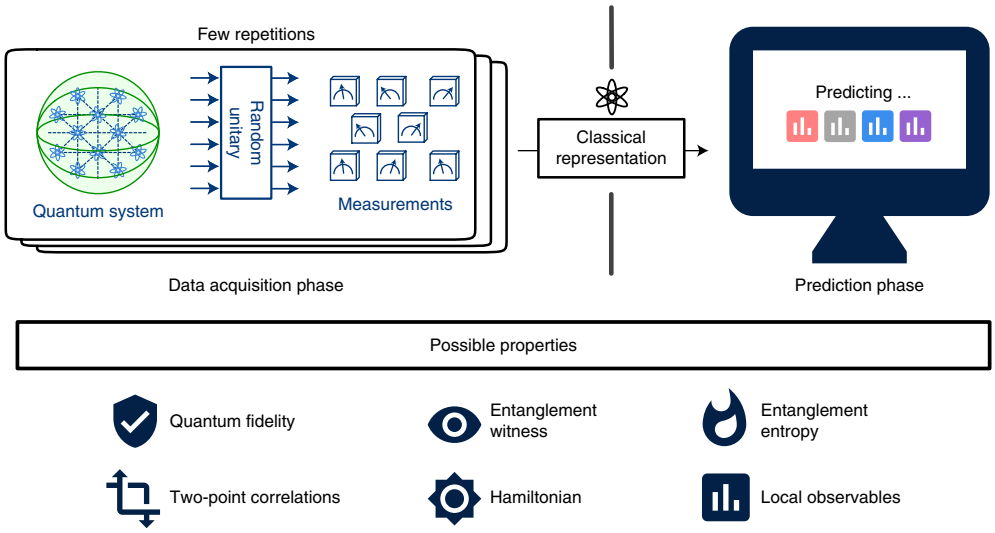


Fig. 1 | An illustration for constructing a classical representation, the classical shadow, of a quantum system from randomized measurements. In the data acquisition phase, we perform a random unitary evolution and measurements on independent copies of an n-qubit system to obtain a classical representation of the quantum system—the classical shadow. Such classical shadows facilitate accurate prediction of a large number of different properties using a simple median-of-means protocol.

special cases. These two procedures turn out to complement each other nicely. Figure 1 provides a visualization and a list of important properties that can be predicted efficiently.

Our main theoretical contribution equips this procedure with rigorous performance guarantees. Classical shadows with size of order $\log(M)$ suffice to predict M target functions in equation (1) simultaneously. Most importantly, the actual system size (number of qubits) does not enter directly. Instead, the number of measurement repetitions N is determined by a (squared) norm $\|O_i\|^2_{\text{shadow}}$. This norm depends on the target functions and the particular measurement procedure used to produce the classical shadow. For example, random n-qubit Clifford circuits lead to the Hilbert–Schmidt norm. On the other hand, random single-qubit Clifford circuits produce a norm that scales exponentially in the locality of target functions, but is independent of system size. The resulting prediction technique is applicable to current laboratory experiments and facilitates the efficient prediction of few-body properties, such as two-point correlation functions, entanglement entropy of small subsystems and expectation values of local observables.

In some cases, this scaling may seem unfavourable. However, we rigorously prove that this is not a flaw of the method, but an unavoidable limitation rooted in quantum information theory. By relating the prediction task to a communication task[9], we establish fundamental lower bounds highlighting that classical shadows are (asymptotically) optimal.

We support our theoretical findings by conducting numerical simulations for predicting various physically relevant properties over a wide range of system sizes. These include quantum fidelity, two-point correlation functions, entanglement entropy and local observables. We confirm that prediction via classical shadows scales favourably and improves on powerful existing techniques—such as machine learning—in a variety of well-motivated test cases. An open-source release for predicting many properties from very few measurements is available at https://github.com/momohuang/predicting-quantum-properties.

Procedure

Throughout this work we restrict attention to n-qubit systems and ρ is a fixed, but unknown, quantum state in $d=2^n$ dimensions. To extract meaningful information, we repeatedly perform a simple measurement procedure: apply a random unitary to rotate the state ($\rho \mapsto U\rho U^\dagger$) and perform a computational-basis measurement. The unitary U is selected randomly from a fixed ensemble. On receiving the n-bit measurement outcome $|\hat{b}\rangle \in \{0,1\}^n$, we store an (efficient) classical description of $U^\dagger|\hat{b}\rangle\langle\hat{b}|U$ in classical memory. It is instructive to view the average (over both the choice of unitary and the outcome distribution) mapping from ρ to its classical snapshot $U^\dagger|\hat{b}\rangle\langle\hat{b}|U$ as a quantum channel:

$$\mathbb{E}\left[U^\dagger|\hat{b}\rangle\langle\hat{b}|U\right] = \mathcal{M}(\rho) \;\Rightarrow\; \rho = \mathbb{E}\left[\mathcal{M}^{-1}\left(U^\dagger|\hat{b}\rangle\langle\hat{b}|U\right)\right] \quad (2)$$

This quantum channel $\mathcal{M}$ depends on the ensemble of (random) unitary transformations. Although the inverted channel $\mathcal{M}^{-1}$ is not physical (it is not completely positive), we can still apply $\mathcal{M}^{-1}$ to the (classically stored) measurement outcome $U^\dagger|\hat{b}\rangle\langle\hat{b}|U$ in a completely classical post-processing step. ($\mathcal{M}$ is invertible if the ensemble of unitary transformations defines a tomographically complete set of measurements; see Supplementary Section 1.) In doing so, we produce a single classical snapshot $\hat{\rho} = \mathcal{M}^{-1}\left(U^\dagger|\hat{b}\rangle\langle\hat{b}|U\right)$ of the unknown state ρ from a single measurement. By construction, this snapshot exactly reproduces the underlying state in expectation (over both unitaries and measurement outcomes): $\mathbb{E}[\hat{\rho}] = \rho$. Repeating this procedure N times results in an array of N independent, classical snapshots of ρ:

$$\begin{aligned} S(\rho; N) = \Big\{&\hat{\rho}_1 = \mathcal{M}^{-1}\left(U_1^\dagger|\hat{b}_1\rangle\langle\hat{b}_1|U_1\right), \ldots, \\ &\hat{\rho}_N = \mathcal{M}^{-1}\left(U_N^\dagger|\hat{b}_N\rangle\langle\hat{b}_N|U_N\right)\Big\} \end{aligned} \quad (3)$$

We call this array the classical shadow of ρ. Classical shadows of sufficient size N are expressive enough to predict many properties of the unknown quantum state efficiently. To avoid outlier corruption, we split the classical shadow into equally sized chunks and construct several, independent sample mean estimators. Subsequently, we predict linear function values (1) via median of means estimation[10,11]. This procedure is summarized in Algorithm 1. For many

physically relevant properties O_i and measurement channels $\mathcal{M}$, Algorithm 1 can be carried out very efficiently without explicitly constructing the large matrix $\hat{\rho}_i$.

Median of means prediction with classical shadows can be defined for any distribution of random unitary transformations. Two prominent examples are (1) random n-qubit Clifford circuits and (2) tensor products of random single-qubit Clifford circuits. Example (1) results in a clean and powerful theory, but also practical drawbacks, because $n^2/\log(n)$ entangling gates are needed to sample from n-qubit Clifford unitaries. The corresponding inverted quantum channel is $\mathcal{M}_n^{-1}(X) = (2^n+1)X - \mathbb{1}$. Example (2) is equivalent to measuring each qubit independently in a random Pauli basis. Such measurements can be routinely carried out in many experimental platforms. The corresponding inverted quantum channel is $\mathcal{M}_P^{-1} = \bigotimes_{i=1}^n \mathcal{M}_1^{-1}$. We refer to examples (1)/(2) as random Clifford/Pauli measurements, respectively. In both cases, the resulting classical shadow can be stored efficiently in a classical memory using the stabilizer formalism.

Algorithm 1. Median of means prediction based on a classical shadow $S(\rho, N)$.

1 **function** LinearPredictions($O_1, \ldots, O_M$, $S(\rho; N)$, K)

2 Import $S(\rho; N) = [\hat{\rho}_1, \ldots, \hat{\rho}_N]$ ▷ Load classical shadow

3 Split the shadow into K equally-sized parts and set ▷ Construct K estimators of ρ

$\hat{\rho}_{(k)} = \frac{1}{\lfloor N/K \rfloor} \sum_{l=(k-1)\lfloor N/K \rfloor + 1}^{k\lfloor N/K \rfloor} \hat{\rho}_l$

4 **for** $i = 1$ to M **do**

5 Output $\hat{o}_i(N,K) = \text{median}\left\{ \text{tr}\left(O_i\hat{\rho}_{(1)}\right), \ldots, \text{tr}\left(O_i\hat{\rho}_{(K)}\right) \right\}$. ▷ Median of means estimation

Rigorous performance guarantees

Theorem 1 (informal version). *Classical shadows of size N suffice to predict M arbitrary linear target functions* $\text{tr}(O_1\rho), \ldots, \text{tr}(O_M\rho)$ *up to additive error ϵ given that* $N \geq$ (order) $\log(M) \max_i ||O_i||_{\text{shadow}}^2/\epsilon^2$. *The definition of the norm $||O_i||_{\text{shadow}}$ depends on the ensemble of unitary transformations used to create the classical shadow.*

We refer to Supplementary Section 1 for background, a detailed statement and proofs. Theorem 1 is most powerful when the linear functions have a bounded norm that is independent of system size. In this case, classical shadows allow for predicting a large number of properties from only a logarithmic number of quantum measurements.

The norm $||O_i||_{\text{shadow}}$ in Theorem 1 plays an important role in defining the space of linear functions that can be predicted efficiently. For random Clifford measurements, $||O||_{\text{shadow}}^2$ is closely related to the Hilbert–Schmidt norm $\text{tr}(O^2)$. As a result, a large collection of (global) observables with a bounded Hilbert–Schmidt norm can be predicted efficiently. For random Pauli measurements, the norm scales exponentially in the locality of the observable, not the actual number of qubits. For an observable O_i that acts non-trivially on (at most) k qubits, $||O_i||_{\text{shadow}}^2 \leq 4^k ||O_i||_\infty^2$, where $||\cdot||_\infty$ denotes the operator norm. (This scaling can be further improved to 3^k if O_i is a tensor product of k single-qubit observables.) This guarantees the accurate prediction of many local observables from a much smaller number of measurements.

Illustrative example applications

Quantum fidelity estimation. Suppose we wish to certify that an experimental device prepares a desired n-qubit state. Typically, this target state $|\psi\rangle\langle\psi|$ is pure and highly structured, for example, a Greenberger–Horne–Zeilinger (GHZ) state[12] for quantum communication protocols or a toric code ground state[13] for fault-tolerant quantum computation. Theorem 1 asserts that a classical shadow (Clifford measurements) of dimension-independent size suffices to accurately predict the fidelity of any state in the lab with any pure target state. This improves on the best existing result on direct fidelity estimation[14] which requires $O(2^n/\epsilon^4)$ samples in the worst case. Moreover, a classical shadow of polynomial size allows for estimating an exponential number of (pure) target fidelities all at once.

Entanglement verification. Fidelities with pure target states can also serve as (bipartite) entanglement witnesses[15]. For many (but not all[16]) bipartite entangled states ρ, there exists a constant α and an observable $O = |\psi\rangle\langle\psi|$ such that $\text{tr}(O\rho) > \alpha \geq \text{tr}(O\rho_s)$, for all (bipartite) separable states ρ_s. Establishing $\text{tr}(O\rho) > \alpha$ verifies the existence of entanglement in the state ρ. Any $O = |\psi\rangle\langle\psi|$ that satisfies the above condition is known as an entanglement witness for the state ρ. Classical shadows (Clifford measurements) of logarithmic size allow for checking a large number of potential entanglement witnesses simultaneously.

Predicting expectation values of local observables. Many near-term applications of quantum devices rely on repeatedly estimating a large number of local observables. For example, low-energy eigenstates of a many-body Hamiltonian may be prepared and studied using a variational method, in which the Hamiltonian, a sum of local terms, is measured many times. Classical shadows constructed from a logarithmic number of random Pauli measurements can efficiently estimate polynomially many such local observables. Because only single-qubit Pauli measurements suffice, this measurement procedure is highly efficient. Potential applications include quantum chemistry[17] and lattice gauge theory[18].

Predicting expectation values of global observables (non-example). Classical shadows are not without limitations. In our examples, the size of classical shadows must either scale with $\text{tr}(O_i^2)$ (Clifford measurements) or must scale exponentially in the locality of O_i (Pauli measurements). Both quantities can simultaneously become exponentially large for non-local observables with large Hilbert–Schmidt norm. A concrete example is the Pauli expectation value of a spin chain: $\langle P_{i_1} \otimes \cdots \otimes P_{i_n} \rangle_\rho = \text{tr}(O_1\rho)$, where $\text{tr}(O_1^2) = 2^n$ and $k = n$ (non-local observable). In this case, classical shadows of exponential size may be required to accurately predict a single expectation value. In contrast, a direct spin measurement achieves the same accuracy with only of order $1/\epsilon^2$ copies of the state ρ.

Matching information-theoretic lower bounds

The non-example above raises an important question: does the scaling of the required number of measurements with Hilbert–Schmidt norm or with the locality of observables arise from a fundamental limitation, or is it merely an artefact of prediction with classical shadows? A rigorous analysis reveals that this scaling is no mere artefact; rather, it stems from information-theoretic reasons.

Theorem 2 (informal version). *Any procedure based on single-copy measurements, that can predict any M linear functions* $\text{tr}(O_i\rho)$ *up to additive error ϵ, requires at least* (order) $\log(M)\ \max_i ||O_i||_{\text{shadow}}^2/\epsilon^2$ *measurements.*

Here, $||O_i||_{\text{shadow}}^2$ could be taken as the Hilbert–Schmidt norm $\text{tr}(O_i^2)$ or as a function scaling exponentially in the locality of O_i. The proof results from embedding the abstract prediction procedure into a communication protocol. Quantum information theory imposes fundamental restrictions on any quantum communication protocol and allows us to deduce stringent lower bounds. See Supplementary Sections 7 and 8 for details and proofs.

The two main technical results complement each other nicely. Theorem 1 equips classical shadows with a constructive performance

guarantee: an order of $\log(M)\max_i\|O_i\|^2_{\text{shadow}}/\epsilon^2$ single-copy measurements suffice to accurately predict an arbitrary collection of M target functions. Theorem 2 highlights that this number of measurements is unavoidable in general.

Predicting nonlinear functions

The classical shadow $S(\rho;N)=\{\hat{\rho}_1,\ldots,\hat{\rho}_N\}$ of the unknown quantum state ρ may also be used to predict nonlinear functions $f(\rho)$. We illustrate this with a quadratic function $f(\rho)=\mathrm{tr}(O\rho\otimes\rho)$, where O acts on two copies of the state. Because $\hat{\rho}_i$ is equal to the quantum state ρ in expectation, one could predict $\mathrm{tr}(O\rho\otimes\rho)$ using two independent snapshots $\hat{\rho}_i,\hat{\rho}_j$, $i\neq j$. Because of independence, $\mathrm{tr}(O\hat{\rho}_i\otimes\hat{\rho}_j)$ correctly predicts the quadratic function in expectation:

$$\mathbb{E}\mathrm{tr}(O\hat{\rho}_i\otimes\hat{\rho}_j)=\mathrm{tr}(O\mathbb{E}\hat{\rho}_i\otimes\mathbb{E}\hat{\rho}_j)=\mathrm{tr}(O\rho\otimes\rho) \tag{4}$$

To reduce the prediction error, we use N independent snapshots and symmetrize over all possible pairs: $\frac{1}{N(N-1)}\sum_{i\neq j}\mathrm{tr}(O\hat{\rho}_i\otimes\hat{\rho}_j)$. We then repeat this procedure several times and form their median to further reduce the likelihood of outlier corruption (similar to median of means). Rigorous performance guarantees are presented in Supplementary Section 1.C. This approach readily generalizes to higher-order polynomials using U-statistics[19].

One particularly interesting nonlinear function is the second-order Rényi entropy, $-\log(\mathrm{tr}(\rho_A^2))$, where A is a subsystem of the n-qubit quantum system. We can rewrite the argument in the log as $\mathrm{tr}(\rho_A^2)=\mathrm{tr}(S_A\rho\otimes\rho)$, where S_A is the local swap operator of two copies of subsystem A, and use classical shadows to obtain very accurate predictions. The required number of measurements scales exponentially in the size of subsystem A, but is independent of total system size. Probing this entanglement entropy is a useful task and a highly efficient specialized approach has been proposed in ref. [20]. We compare this method of Brydges and colleagues to classical shadows in the numerical experiments.

For nonlinear functions, unlike linear ones, we have not derived an information-theoretic lower bound on the number of measurements needed, although it may be possible to do so by generalizing our methods.

Numerical experiments

One of the key features of prediction with classical shadows is scalability. The data acquisition phase is designed to be tractable for state-of-the-art platforms (Pauli measurements) and future quantum computers (Clifford measurements), respectively. The resulting classical shadow can be stored efficiently in classical memory. For many important features—such as local observables or global features with efficient stabilizer decompositions—scalability, moreover, extends to the computational cost associated with median of means prediction.

These design features allowed us to conduct numerical experiments for a wide range of problems and system sizes (up to 160 qubits). The computational bottleneck is not feature prediction with classical shadows, but generating synthetic data, that is, classically generating target states and simulating quantum measurements. Needless to say, this classical bottleneck does not occur in actual experiments. We then use this synthetic data to learn a classical representation of ρ and use this representation to predict various interesting properties.

Machine-learning-based approaches[3,4] are among the most promising alternative methods that have applications in this regime, where the Hilbert space dimension is roughly comparable to the total number of silicon atoms on earth ($2^{160}\simeq 10^{48}$). For example, a recent version of neural network quantum state tomography (NNQST) is a generative model that is based on a deep neural network trained on independent quantum measurement outcomes with local SIC/tetrahedral positive-operator valued measures (POVMs)[21]. In this section, we consider the task of learning a classical representation of an unknown quantum state, and using the representation to predict various properties, addressing the relative merit of classical shadows and alternative methods.

Predicting quantum fidelities via Clifford measurements. Here we focus on classical shadows based on random Clifford measurements, which are designed to predict observables with a bounded Hilbert–Schmidt norm. When the observables have efficient representations—such as efficient stabilizer decompositions—the computational cost for performing median of means prediction can also be efficient. (The runtime of Algorithm 1 is dominated by the cost of computing quadratic functions $\left\langle\hat{b}\middle|UOU^\dagger\middle|\hat{b}\right\rangle$ in 2^n dimensions. If $O=|\psi\rangle\langle\psi|$ is a stabilizer state, the Gottesman–Knill theorem allows for evaluation in $\mathcal{O}(n^2)$-time.) An important example is the quantum fidelity with a target state. In ref. [3], the viability of NNQST is demonstrated by considering GHZ states with a varying number of qubits n. Numerical experiments highlight that the number of measurement repetitions (size of the training data) to learn a neural network model of the GHZ state that achieves a target fidelity of 0.99 scales linearly in n. We have also implemented NNQST for GHZ states and compared it to median of means prediction with classical shadows. Figure 2a confirms the linear scaling of NNQST and the assertion of Theorem 1: classical shadows of constant size suffice to accurately estimate GHZ target fidelities, regardless of the actual system size. In addition, we have also tested the ability of both approaches to detect potential state preparation errors. More precisely, we consider a scenario where the GHZ source introduces a phase error with probability $p\in[0,1]$:

$$\rho_p=(1-p)\left|\psi^+_{\mathrm{GHZ}}(n)\right\rangle\left\langle\psi^+_{\mathrm{GHZ}}(n)\right|+p\left|\psi^-_{\mathrm{GHZ}}(n)\right\rangle\left\langle\psi^-_{\mathrm{GHZ}}(n)\right|,$$
$$\left|\psi^\pm_{\mathrm{GHZ}}(n)\right\rangle=\tfrac{1}{\sqrt{2}}\left(|0\rangle^{\otimes n}\pm|1\rangle^{\otimes n}\right) \tag{5}$$

We learn a classical representation of the GHZ source and subsequently predict the fidelity with the pure GHZ state. Figure 2b highlights that the classical shadow prediction accurately tracks the decrease in target fidelity as the error parameter p increases. NNQST, in contrast, seems to consistently overestimate this target fidelity. In the extreme case ($p=1$), the true underlying state is completely orthogonal to the target state, but NNQST nonetheless reports fidelities close to one. This shortcoming arises because the POVM-based machine-learning approach can only efficiently estimate an upper bound on the true quantum fidelity efficiently. To estimate the actual fidelity, an exceedingly large number of measurements is needed. Similar experiments are described in Supplementary Section 2, where we focus on toric code ground states and entanglement witnesses, respectively.

Predicting two-point correlation and subsystem entanglement entropy (Pauli measurements). Classical shadows based on random Clifford measurements excel at predicting quantum fidelities. However, random Clifford measurements can be challenging to implement in practice, because many entangling gates are needed to implement general Clifford circuits. Next we consider classical shadows based on random local Pauli measurements, which are easier to perform experimentally. The subsystem properties can be predicted efficiently by constructing the reduced density matrix from the classical shadow. Therefore, the computational complexity scales exponentially only in the subsystem size, rather than the size of the entire system. Our numerical experiments confirm that classical shadows obtained using random Pauli measurements excel at predicting few-body properties of a quantum state, such as two-point correlation functions and subsystem entanglement entropy.

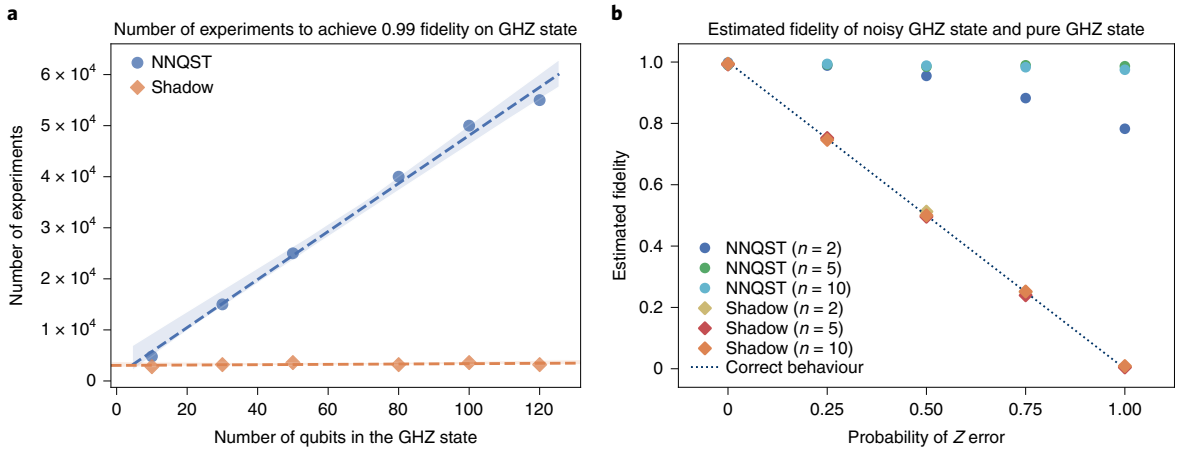


Fig. 2 | Predicting quantum fidelities using classical shadows (Clifford measurements) and NNQST. a, Number of measurements required to identify an *n*-qubit GHZ state with 0.99 fidelity. The shaded regions show the s.d. of the needed number of experiments over 10 independent runs. The dashed lines are the linear regression lines for the number of experiments under different system sizes. **b**, Estimated fidelity between a perfect GHZ target state and a noisy preparation, where *Z* errors can occur with probability $p \in [0, 1]$, under 6×10^4 experiments. The dotted line represents the true fidelity as a function of *p*. NNQST can only estimate an upper bound on quantum fidelity efficiently, so we consider this upper bound for NNQST and use quantum fidelity for the classical shadow.

Two-point correlation functions. NNQST has been shown to predict two-point correlation functions effectively[3]. Here, we compare classical shadows with NNQST for two physically motivated test cases: ground states of the antiferromagnetic transverse field Ising model in one dimension (TFIM) and the antiferromagnetic Heisenberg model in two dimensions. The Hamiltonian for TFIM is $H = J\sum_i \sigma_i^Z \sigma_{i+1}^Z + h\sum_i \sigma_i^X$, where $J > 0$, and we consider a chain of 50 lattice sites. The critical point occurs at $h = J$ and exhibits power-law decay of correlations rather than exponential decay. The Hamiltonian for the two-dimensional (2D) Heisenberg model is $H = J\sum_{\langle i,j\rangle} \vec{\sigma}_i \cdot \vec{\sigma}_j$, where $J > 0$, and we consider an 8×8 triangular lattice. We follow the approach in ref. [3], where the ground state is approximated by a tensor network found using the density matrix renormalization group (DMRG). Random Pauli measurements on the ground state may then be simulated using this tensor network. The two methods are compared in Fig. 3. In Fig. 3a,b, we can see that both the classical shadow (with Pauli measurements) and NNQST perform well at predicting two-point correlations. However, NNQST has a larger error for the 2D Heisenberg model; note that, for larger separations (the lower right corner of the surface plot), NNQST produces some fictitious oscillations that are not visible in the results from DMRG and classical shadows. The two approaches use the same quantum measurement data; the only difference is the classical post-processing. In Fig. 3c we compare the cost of this classical post-processing, finding roughly a 10^4 times speed-up in classical processing time using the classical shadow instead of NNQST.

Subsystem entanglement entropies. An important nonlinear property that can be predicted with classical shadows is subsystem entanglement entropy. The required number of measurements scales exponentially in subsystem size, but is independent of the total number of qubits. Moreover, these measurements can be used to predict many subsystem entanglement entropies at once. This problem has also been studied extensively in ref. [20], where a specialized approach (which we refer to here as the 'Brydges et al. protocol') was designed to efficiently estimate second-order Rényi entanglement entropies using random local measurements. In ref. [20], a random unitary rotation is reused several times. Predictions using classical shadows could also be slightly modified to adapt to this scenario. Results from our numerical experiments are shown in Fig. 4. In Fig. 4a, we predict the entanglement entropy for all subsystems of size ≤ 2 from only 2,500 measurements of the approximate ground state of the disordered Heisenberg model in one dimension. This is a prototypical model for studying many-body localization[22]. The ground state is approximated by a set of singlet states $\{\frac{1}{\sqrt{2}}(|01\rangle - |10\rangle)\}$ found using the strong-disorder renormalization group[23,24]. Both the classical shadow protocol and the Brydges et al. method use random single-qubit rotations and basis measurements to find a classical representation of the quantum state; the only difference between the methods is in the classical post-processing. For these small subsystems, we find that the prediction error of the classical shadow is smaller than the error of the Brydges et al. protocol. In Fig. 4b, we consider predicting the entanglement entropy in a GHZ state for system sizes ranging from $n = 4$ to $n = 10$ qubits. We focus on the entanglement entropy of the subsystem with system size $n/2$ on the left side. Note that this entanglement entropy is equal to one bit for any system size *n*. To achieve an error of 0.05, classical shadows require several times fewer measurements and the discrepancy increases as we require smaller error.

Application to quantum simulation of the lattice Schwinger model (Pauli measurements). Simulations of quantum field theory using quantum computers may someday advance our understanding of fundamental particle physics. Although high-impact discoveries may still be a way off, notable results have already been achieved in studies of 1D lattice gauge theories using quantum platforms.

For example, in ref. [18], a 20-qubit trapped ion analogue quantum simulator was used to prepare low-energy eigenstates of the lattice Schwinger model (1D quantum electrodynamics). The authors prepared a family of quantum states $\{|\psi(\theta)\rangle\}$, where θ is a variational parameter, and computed the variance of the energy $\langle(\hat{H} - \langle\hat{H}\rangle_\theta)^2\rangle_\theta$ for each value of θ. Here, $\hat{H}$ is the Hamiltonian of the model, and $\langle\hat{O}\rangle_\theta = \langle\psi(\theta)|\hat{O}|\psi(\theta)\rangle$ is the expectation value of the operator $\hat{O}$ in the state $|\psi(\theta)\rangle$. Because energy eigenstates, and only energy eigenstates, have vanishing energy dispersion, adjusting θ to minimize the variance of energy prepares an energy eigenstate.

After solving the Gauss law constraint to eliminate the gauge fields, the Hamiltonian $\hat{H}$ of the Schwinger model is 2-local, though not geometrically local in one dimension. Hence the quantity $\langle(\hat{H} - \langle\hat{H}\rangle_\theta)^2\rangle_\theta$ is a sum of expectation values of 4-local observables, which can be measured efficiently using a classical shadow

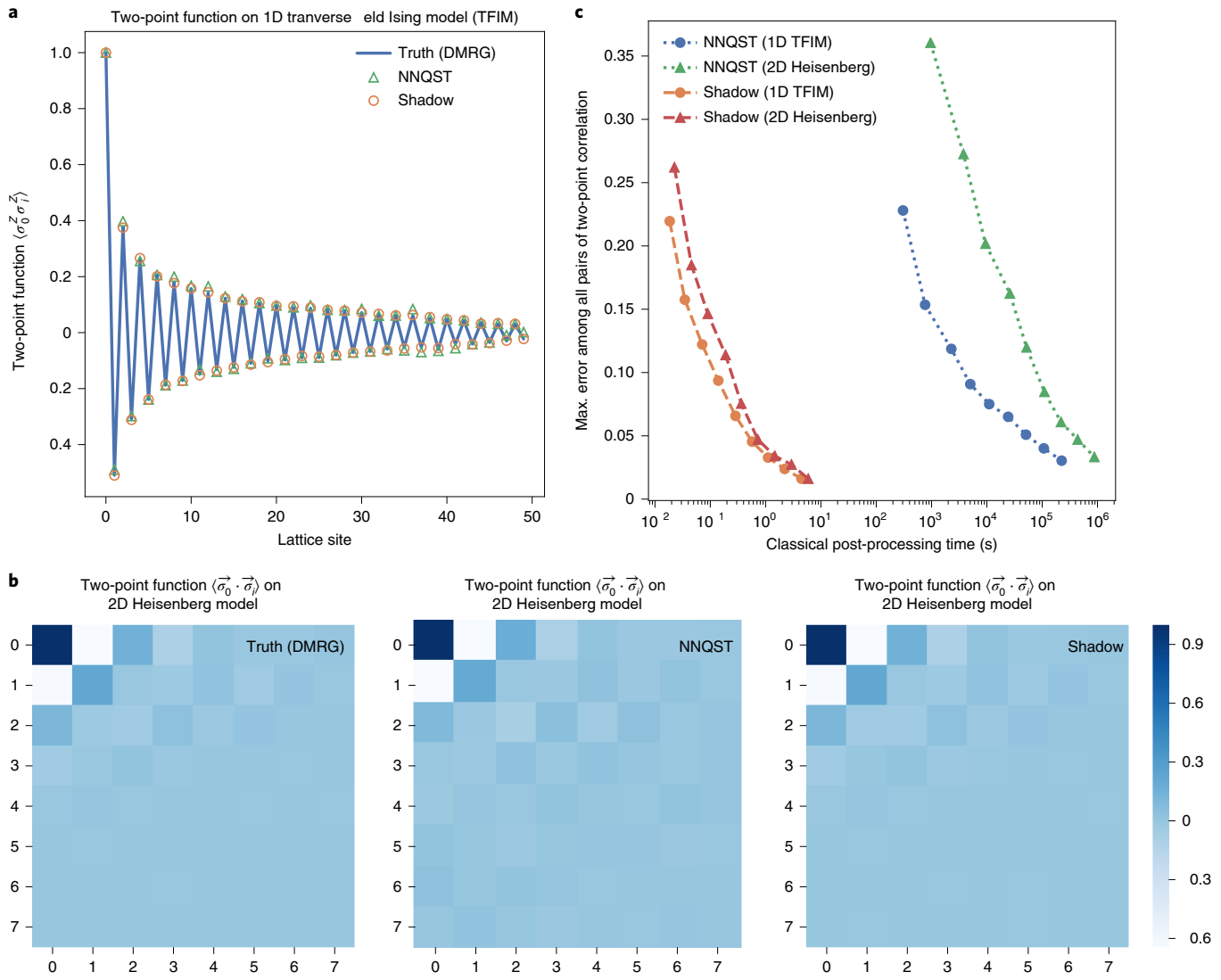


Fig. 3 | Predicting two-point correlation functions using classical shadows (Pauli measurements) and NNQST. a, Predictions of two-point functions $\langle\sigma_0^Z\sigma_i^Z\rangle$ for ground states of the 1D critical antiferromagnetic TFIM with 50 lattice sites. These are based on 2^{19} random Pauli measurements. **b**, Predictions of two-point functions $\langle\vec{\sigma}_0\cdot\vec{\sigma}_i\rangle$ for the ground state of the 2D antiferromagnetic Heisenberg model with 8×8 lattice sites. The predictions are based on 2^{19} random Pauli measurements. **c**, Classical processing time (CPU time in seconds) and prediction error (the largest among all pairs of two-point correlations) over different numbers of measurements: $\{2^{11}, ..., 2^{19}\}$. The quantum measurement scheme in classical shadows (Pauli) is the same as the POVM-based neural network tomography (NNQST) in ref. [3]. The only difference is the classical post-processing. As the number of measurements increases, the processing time increases, while the prediction error decreases.

derived from random Pauli measurements. This is illustrated in Fig. 5a. Figure 5b compares the performance of classical shadows to the measurement scheme for 4-local observables designed in ref. [18], and also to a recent method[25] for measuring local observables, as well as the standard approach that directly measures all observables independently.

The results show, for the methods we considered, the number of copies of the quantum state needed to measure the expectation value of all 4-local Pauli observables in $\langle(\hat{H}-\langle\hat{H}\rangle_\theta)^2\rangle_\theta$ with an error equivalent to measuring each of these observables at least 100 times. In ref. [18], such a relatively small number of measurements per local observable already yielded results comparable to theoretical predictions based on exact diagonalization. We find that the performance of the classical shadow method is better than the method used in ref. [18] only for system size larger than 50 qubits, and may actually be worse for small system sizes. However, classical shadows provide a good prediction for any set of local observables, while the method of ref. [18] was hand-crafted for the particular task of estimating the variance of the energy in the Schwinger model.

To make a more apt comparison, we constructed a deterministic version of classical shadows, using a fixed set of measurements rather than random Pauli measurements, specifically adapted for the purpose of estimating $\langle(\hat{H}-\langle\hat{H}\rangle_\theta)^2\rangle_\theta$ in the lattice Schwinger model. This deterministic collection of Pauli measurements is obtained by a powerful technique called derandomization[26,27]. This procedure simulates the classical shadow scheme based on randomized measurements and makes use of the rigorous performance bound we developed. When a coin is tossed in the randomized scheme to decide which measurement to perform next, the next measurement in the derandomized version is chosen to have the best possible performance bound for the rest of the protocol. It turns out that this derandomization of the classical shadow method can be carried out very efficiently (full details will appear in upcoming work). Not surprisingly, the derandomized version, also included in Fig. 5,

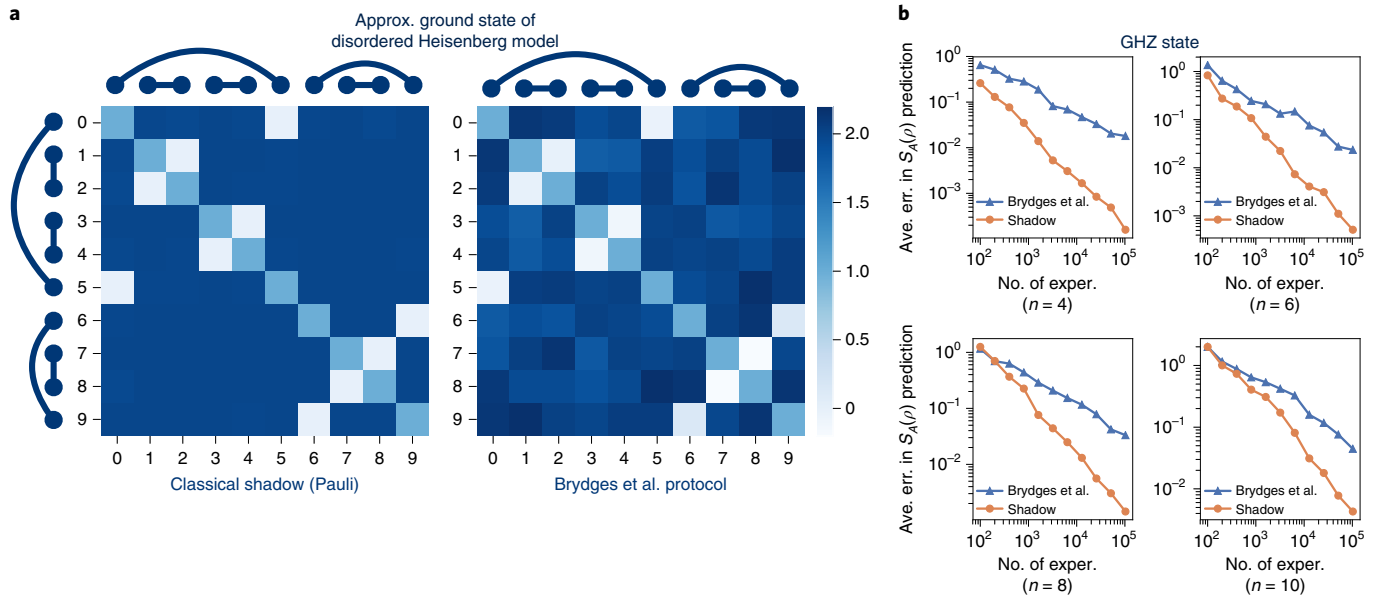


Fig. 4 | Predicting entanglement Rényi entropies using classical shadows (Pauli measurements) and the Brydges et al. protocol. a, Prediction of second-order Rényi entanglement entropy for all subsystems of size at most two in the approximate ground state of a disordered Heisenberg spin chain with 10 sites and open boundary conditions. The classical shadow is constructed from 2,500 quantum measurements. The predicted values using the classical shadow visually match the true values with a maximum prediction error of 0.052. The Brydges et al. protocol[20] results in a maximum prediction error of 0.24. **b**, Comparison of classical shadows and the Brydges et al. protocol[20] for estimating second-order Rényi entanglement entropy in GHZ states. We consider the entanglement entropy of the subsystem with size $n/2$ on the left side.

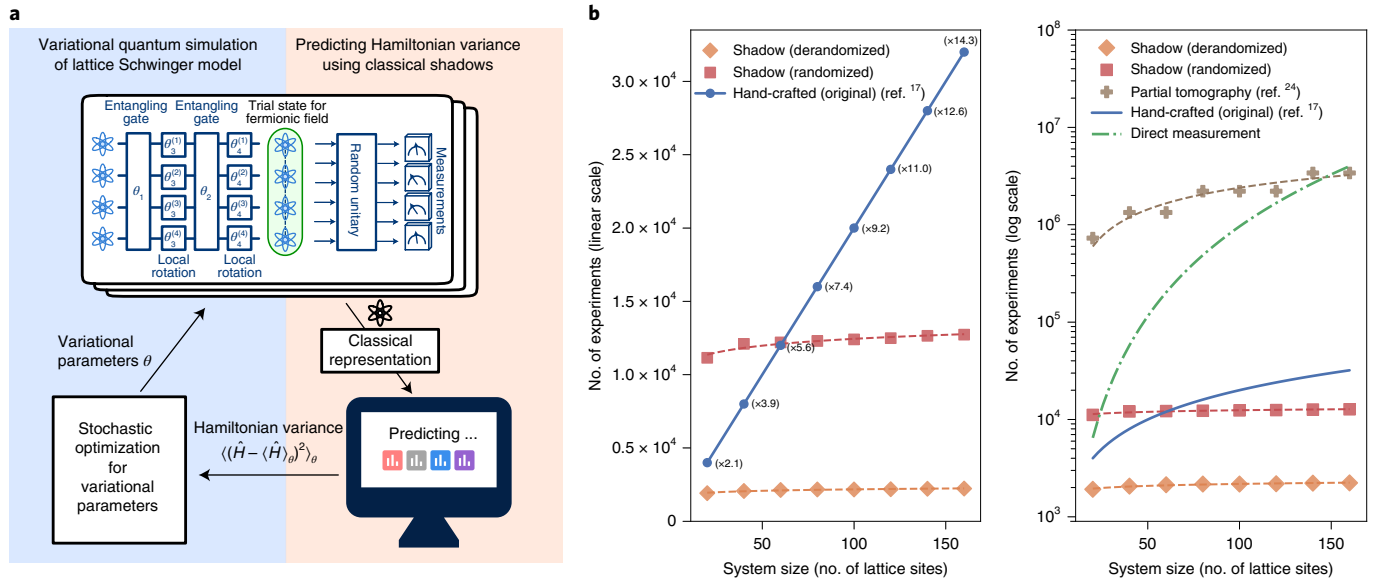


Fig. 5 | Application of classical shadows (Pauli measurements) to variational quantum simulation of the lattice Schwinger model. a, An illustration of variational quantum simulation and the role of classical shadows. **b**, Comparison between different approaches in the number of measurements needed to predict all 4-local Pauli observables in the expansion of $\langle(\hat{H}-\langle\hat{H}\rangle_\theta)^2\rangle_\theta$ with an error equivalent to measuring each Pauli observable at least 100 times. We include a linear-scale plot that compares classical shadows with the original hand-designed measurement scheme in ref. [18] (left) and a log-scale plot that compares with other approaches (right). In the linear-scale plot, $(\times T)$ indicates that the original scheme uses T times the number of measurements compared to classical shadows (derandomized).

outperforms the randomized version by a considerable margin. We then find that the derandomized classical shadow method is significantly more efficient than the other methods we considered, including the hand-crafted method from ref. [18]. Finally, we emphasize that the derandomization procedure is fully automated (see https://github.com/momohuang/predicting-quantum-properties for open-source code) and not problem-specific. It could be used for any prespecified set of local observables.

Outlook

A classical shadow is a succinct classical description of a quantum state, which can be extracted by performing reasonably simple

single-copy measurements on a reasonably small number of copies of the state. We have shown that, given its classical shadow, many properties of a quantum state can be accurately and efficiently predicted with a rigorous performance guarantee. In the case of classical shadows based on random Pauli measurements, our methods are feasible using current quantum platforms, and our numerical experiments indicate that many properties can be predicted more efficiently using classical shadows than by using other methods. We therefore anticipate that classical shadows will be useful in near-term experiments characterizing noise in quantum devices and exploring variational quantum algorithms for optimization, materials science and chemistry. Our results also suggest a variety of avenues for further theoretical exploration. Can the classical shadow of a quantum state be updated efficiently as the state undergoes time evolution governed by a local Hamiltonian? Can we use classical shadows to predict properties of quantum channels rather than states? What are the applications of classical shadows based on other ensembles of unitary transformations, for example ensembles of shallow random quantum circuits? More broadly, by mapping many-particle quantum states to succinct classical data, classical shadows open opportunities for applying classical machine-learning methods to numerous challenging problems in quantum many-body physics[4,28,29], such as the classification of quantum phases of matter and the simulation of strongly correlated quantum phenomena.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41567-020-0932-7.

Received: 20 October 2019; Accepted: 6 May 2020;
Published online: 22 June 2020

References

1. Preskill, J. Quantum computing in the NISQ era and beyond. *Quantum* **2**, 79 (2018).
2. Cramer, M. et al. Efficient quantum state tomography. *Nat. Commun.* **1**, 149 (2010).
3. Carrasquilla, J., Torlai, G., Melko, R. G. & Aolita, L. Reconstructing quantum states with generative models. *Nat. Mach. Intell.* **1**, 155–161 (2019).
4. Torlai, G. et al. Neural-network quantum state tomography. *Nat. Phys.* **14**, 447–450 (2018).
5. Aaronson, S. Shadow tomography of quantum states. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing* (*STOC 2018*) 325–338 (ACM, 2018)
6. Aaronson, S. & Rothblum, G. N. Gentle measurement of quantum states and differential privacy. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing* (*STOC 2019*) 322–333 (ACM, 2019).
7. Guta, M., Kahn, J., Kueng, R. J. & Tropp, J. A. Fast state tomography with optimal error bounds. *J. Phys. A* **53**, 204001 (2020).
8. Gottesman, D. *Stabilizer Codes and Quantum Error Correction* PhD thesis, Caltech (1997).
9. Fano, R. M. *Transmission of Information: A Statistical Theory of Communications* (MIT Press, 1961).
10. Jerrum, M. R., Valiant, L. G. & Vazirani, V. V. Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.* **43**, 169–188 (1986).
11. Nemirovsky, A. S. & Yudin, D. B. *Problem Complexity and Method Efficiency in Optimization* (Wiley-Interscience, 1983).
12. Greenberger, D. M., Horne, M. A. & Zeilinger, A. in *Bell's Theorem, Quantum Theory and Conceptions of the Universe. Fundamental Theories of Physics* Vol. 37 (ed. Kafatos, M.) 69–72 (Springer, 1989).
13. Dennis, E., Kitaev, A., Landahl, A. & Preskill, J. Topological quantum memory.*J. Math. Phys.* **43**, 4452–4505 (2002).
14. Flammia, S. T. & Liu, Y.-K. Direct fidelity estimation from few Pauli measurements. *Phys. Rev. Lett.* **106**, 230501 (2011).
15. Gühne, O. & Tóth, G. Entanglement detection. *Phys. Rep.* **474**, 1–75 (2009).
16. Weilenmann, M., Dive, B., Trillo, D., Aguilar, E. A. & Navascués, M. Entanglement detection beyond measuring fidelities. *Phys. Rev. Lett.* **124**, 200502 (2020).
17. Kandala, A. et al. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature* **549**, 242–246 (2017).
18. Kokail, C. et al. Self-verifying variational quantum simulation of lattice models. *Nature* **569**, 355–360 (2019).
19. Hoeffding, W. in *Breakthroughs in Statistics* 308–334 (Springer, 1992).
20. Brydges, T. et al. Probing Rényi entanglement entropy via randomized measurements. *Science* **364**, 260–263 (2019).
21. Renes, J. M., Blume-Kohout, R., Scott, A. J. & Caves, C. M. Symmetric informationally complete quantum measurements. *J. Math. Phys.* **45**, 2171–2180 (2004).
22. Nandkishore, R. & Huse, D. A. Many-body localization and thermalization in quantum statistical mechanics. *Annu. Rev. Condens. Matter Phys.* **6**, 15–38 (2015).
23. Dasgupta, C. & Ma, S.-k Low-temperature properties of the random Heisenberg antiferromagnetic chain. *Phys. Rev. B* **22**, 1305 (1980).
24. Ma, S.-k, Dasgupta, C. & Hu, C.-k Random antiferromagnetic chain. *Phys. Rev. Lett.* **43**, 1434 (1979).
25. Bonet-Monroig, X., Babbush, R. & O'Brien, T. E. Nearly optimal measurement scheduling for partial tomography of quantum states. Preprint at https://arxiv.org/pdf/1908.05628.pdf (2019).
26. Raghavan, P. Probabilistic construction of deterministic algorithms: approximating packing integer programs.*J. Comput. Syst. Sci.* **37**, 130–143 (1988).
27. Spencer, J. Ten lectures on the probabilistic method. In *CBMS-NSF Regional Conference Series in Applied Mathematics* 2nd edn, Vol. 64 (SIAM, 1994).
28. Carleo, G. & Troyer, M. Solving the quantum many-body problem with artificial neural networks. *Science* **355**, 602–606 (2017).
29. Carrasquilla, J. & Melko, R. G. Machine learning phases of matter. *Nat. Phys.* **13**, 431–434 (2017).
30. Paini, M. & Kalev, A. An approximate description of quantum states. Preprint at https://arxiv.org/pdf/1910.10543.pdf (2019).

Data availability

Source data are available for this paper. All other data that support the plots within this paper and other findings of this study are available from the corresponding author upon reasonable request.

Code availability

Source code for an efficient implementation of the proposed procedure is available at https://github.com/momohuang/predicting-quantum-properties.

Acknowledgements

We thank V. Albert, F. Brandão, M. Endres, I. Roth, J. Tropp, T. Vidick, M. Weilenmann and J. Wright for valuable input and inspiring discussions. L. Aolita and G. Carleo provided helpful advice regarding presentation. Our gratitude extends, in particular, to J. Iverson, who helped us in devising a numerical sampling strategy for toric code ground states. We also thank M. Paini and A. Kalev for informing us about their related work[30], where they discussed succinct classical 'snapshots' of quantum states obtained from randomized local measurements. H.-Y.H. is supported by the Kortschak Scholars Program. R.K. acknowledges funding provided by the Office of Naval Research (award no. N00014-17-1-2146) and the Army Research Office (award no. W911NF121054). J.P. acknowledges funding from ARO-LPS, NSF and DOE. The Institute for Quantum Information and Matter is an NSF Physics Frontiers Center.

Author contributions

H.-Y.H. and R.K. developed the theoretical aspects of this work. H.-Y.H. conducted the numerical experiments and wrote the open-source code. J.P. conceived the applications of classical shadows. H.-Y.H., R.K. and J.P. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41567-020-0932-7.

Correspondence and requests for materials should be addressed to H.-Y.H.

Peer review information *Nature Physics* thanks Yi-Kai Liu and other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

nature physics

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41567-020-0932-7

In the format provided by the authors and unedited.

Predicting many properties of a quantum system from very few measurements

Hsin-Yuan Huang[1,2] ✉, **Richard Kueng**[1,2,3] **and John Preskill**[1,2,4]

[1]Institute for Quantum Information and Matter, California Institute of Technology, Pasadena, CA, USA. [2]Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, USA. [3]Institute for Integrated Circuits, Johannes Kepler University Linz, Linz, Austria. [4]Walter Burke Institute for Theoretical Physics, California Institute of Technology, Pasadena, CA, USA. ✉e-mail: hsinyuan@caltech.edu

Supplementary information: Predicting many properties of a quantum system from very few measurements

Hsin-Yuan Huang,[1,2,*] Richard Kueng,[1,2,3] and John Preskill[1,2,4]

[1] *Institute for Quantum Information and Matter, Caltech, Pasadena, CA, USA*
[2] *Department of Computing and Mathematical Sciences, Caltech, Pasadena, CA, USA*
[3] *Institute for Integrated Circuits, Johannes Kepler University Linz, Austria*
[4] *Walter Burke Institute for Theoretical Physics, Caltech, Pasadena, CA, USA*

(Dated: May 6, 2020)

1. GENERAL FRAMEWORK FOR CONSTRUCTING CLASSICAL SHADOWS

A. Data acquisition and classical shadows

Throughout this work we restrict attention to multi-qubit systems and ρ is a fixed, but unknown, quantum state in $d = 2^n$ dimensions. We present a general-purpose strategy for predicting many properties of this unknown state. To extract meaningful information about ρ, we need to perform a collection of measurements.

Definition S1 (measurement primitive)**.** *We can apply a restricted set of unitary evolutions $\rho \mapsto U\rho U^\dagger$, where U is chosen from an ensemble $\mathcal{U}$. Subsequently, we can measure the rotated state in the computational basis $\{|b\rangle : b \in \{0,1\}^n\}$. Moreover, we assume that this collection is tomographically complete, i.e. for each $\sigma \neq \rho$ there exist $U \in \mathcal{U}$ and b such that $\langle b|U\sigma U^\dagger|b\rangle \neq \langle b|U\rho U^\dagger|b\rangle$.*

Based on this primitive, we repeatedly perform a simple randomized measurement procedure: randomly rotate the state $\rho \mapsto U\rho U^\dagger$ and perform a computational basis measurement. Then, after the measurement, we apply the inverse of U to the resulting computational basis state. This procedure collapses ρ to

$$U^\dagger|\hat{b}\rangle\langle\hat{b}|U \quad \text{where} \quad \Pr\big[\hat{b} = b\big] = \langle b|U\rho U^\dagger|b\rangle,\ b \in \{0,1\}^n \quad \text{(Born's rule)}. \tag{S1}$$

This random snapshot contains valuable information about ρ in expectation:

$$\mathbb{E}\left[U^\dagger|\hat{b}\rangle\langle\hat{b}|U\right] = \mathbb{E}_{U\sim\mathcal{U}} \sum_{b\in\{0,1\}^n} \langle b|U\rho U^\dagger|b\rangle U^\dagger|b\rangle\langle b|U = \mathcal{M}(\rho). \tag{S2}$$

For any unitary ensemble $\mathcal{U}$, this relation describes a quantum channel $\rho \mapsto \mathcal{M}(\rho)$. Tomographic completeness ensures that $\mathcal{M}$ — viewed as a linear map — has a unique inverse $\mathcal{M}^{-1}$ and we set

$$\hat{\rho} = \mathcal{M}^{-1}\left(U^\dagger|\hat{b}\rangle\langle\hat{b}|U\right) \qquad \text{(classical shadow)}. \tag{S3}$$

The classical shadow is a modified post-measurement state that has unit trace, but need not be positive semi-definite. However, it is designed to reproduce the underlying state ρ exactly in expectation: $\mathbb{E}[\hat{\rho}] = \rho$. This classical shadow $\hat{\rho}$ corresponds to the linear inversion (or least squares) estimator of ρ in the single-shot limit. Linear inversion estimators have been used to perform full quantum state tomography [30, 58], where an exponential number of measurements is needed. We wish to show that $\hat{\rho}$ can predict many properties from only very few measurements.

B. Predicting linear functions with classical shadows

Classical shadows are well suited to predict *linear* functions in the unknown state ρ:

$$o_i = \mathrm{tr}\left(O_i\rho\right) \quad 1 \leq i \leq M. \tag{S4}$$

To achieve this goal, we simply replace the (unknown) quantum state ρ by a classical shadow $\hat{\rho}$. Since classical shadows are random, this produces a random variable that yields the correct prediction in expectation:

$$\hat{o}_i = \mathrm{tr}\left(O_i\hat{\rho}\right) \quad \text{obeys} \quad \mathbb{E}\left[\hat{o}\right] = \mathrm{tr}\left(O_i\rho\right). \tag{S5}$$

Fluctuations of $\hat{o}$ around this desired expectation are controlled by the variance.

*Electronic address: hsinyuan@caltech.edu

Lemma S1. *Fix O and set $\hat{o} = \mathrm{tr}(O\hat{\rho})$, where $\hat{\rho}$ is a classical shadow* (S3)*. Then*

$$\mathrm{Var}\left[\hat{o}\right] = \mathbb{E}\left[\left(\hat{o} - \mathbb{E}\left[\hat{o}\right]\right)^2\right] \leq \left\| O - \tfrac{\mathrm{tr}(O)}{2^n}\mathbb{I} \right\|_{\mathrm{shadow}}^2 . \tag{S6}$$

The norm $\|\cdot\|_{\mathrm{shadow}}$ only depends on the measurement primitive:

$$\|O\|_{\mathrm{shadow}} = \max_{\sigma:\ state} \left(\mathbb{E}_{U\sim\mathcal{U}} \sum_{b\in\{0,1\}^n} \langle b|U\sigma U^\dagger|b\rangle \langle b|U\mathcal{M}^{-1}\left(O\right)U^\dagger|b\rangle^2 \right)^{1/2}. \tag{S7}$$

It is easy to check that $\|O\|_{\mathrm{shadow}}$ is nonnegative and homogeneous ($\|0\|_{\mathrm{shadow}} = 0$). After some work, one can verify that this expression also obeys the triangle inequality, and so is indeed a norm.

Proof. Classical shadows have unit trace by construction ($\mathrm{tr}(\hat{\rho}) = 1$). This feature implies that the variance only depends on the traceless part $O_0 = O - \frac{\mathrm{tr}(O)}{2^n}\mathbb{I}$ of O, not O itself:

$$\hat{o} - \mathbb{E}[\hat{o}] = \mathrm{tr}\left(O\hat{\rho}\right) - \mathrm{tr}\left(O\rho\right) = \mathrm{tr}\left(O_0\hat{\rho}\right) - \mathrm{tr}\left(O_0\rho\right). \tag{S8}$$

Moreover, it is easy to check that the inverse of $\mathcal{M}$ (S2) is self-adjoint ($\mathrm{tr}\left(X\mathcal{M}^{-1}(Y)\right) = \mathrm{tr}\left(\mathcal{M}^{-1}(X)Y\right)$ for any pair of matrices X, Y with compatible dimension). These two observations allow us to rewrite the variance in the following fashion:

$$\mathrm{Var}\left[\hat{o}\right] = \mathbb{E}\left[\left(\hat{o} - \mathbb{E}\hat{o}\right)^2\right] = \mathbb{E}\left[\left(\mathrm{tr}(O_0\hat{\rho})\right)^2\right] - \left(\mathrm{tr}\left(O_0\,\mathbb{E}\left[\hat{\rho}\right]\right)\right)^2 = \mathbb{E}\left[\langle\hat{b}|U\mathcal{M}^{-1}(O_0)U^\dagger|\hat{b}\rangle^2\right] - \left(\mathrm{tr}\left(O_0\rho\right)\right)^2. \tag{S9}$$

Classical shadows arise from mixing two types of randomness: (i) a (classical) random choice of unitary $U \sim \mathcal{U}$ and (ii) a random choice of computational basis state $|\hat{b}\rangle$ that is governed by Born's rule (S1). Inserting the average over computational basis states produces a (squared) norm that closely resembles the advertised expression, but does depend on the underlying state:

$$\mathbb{E}\langle\hat{b}|U\mathcal{M}^{-1}(O_0)U^\dagger|\hat{b}\rangle^2 = \mathbb{E}_{U\sim\mathcal{U}} \sum_{b\in\{0,1\}^n} \langle b|U\rho U^\dagger|b\rangle\langle b|U\mathcal{M}^{-1}(O_0)U^\dagger|b\rangle^2. \tag{S10}$$

Maximizing over all possible states σ removes this implicit dependence and produces a universal upper bound on the variance. Ignoring the subtraction of $\left(\mathrm{tr}\left(O_0\rho\right)\right)^2$ (which can only make the bound tighter), we obtain (S6). □

Lemma S1 sets the stage for successful linear function estimation with classical shadows. A single classical shadow (S3) correctly predicts *any* linear function $o_i = \mathrm{tr}(O_i\rho)$ in expectation. Convergence to this desired target can be boosted by forming empirical averages of multiple independent shadow predictions. The *empirical mean* is the canonical example for such a procedure. Construct N independent classical shadows $\hat{\rho}_1, \ldots, \hat{\rho}_N$ and set

$$\hat{o}_i(N,1) = \frac{1}{N}\sum_{j=1}^{N} \mathrm{tr}\left(O_i\hat{\rho}_j\right). \tag{S11}$$

Each summand is an independent random variable with correct expectation and variance bounded by Lemma S1. Convergence to the expectation value $\mathrm{tr}(O_i\rho)$ can be controlled by classical concentration arguments (e.g. Chernoff or Hoeffding inequalities). In order to achieve a failure probability of (at most) δ, the number of samples must scale like $N = \mathrm{Var}\left[\hat{o}_i\right]/(\delta\epsilon^2)$. While the scaling in variance and approximation accuracy ϵ is optimal, the dependence on $1/\delta$ is particularly bad. Unfortunately, this feature of sample mean estimators cannot be avoided without imposing additional assumptions (that do not apply to classical shadows). *Median of means* [36, 47] is a conceptually simple trick that addresses this issue. Instead of using all samples to construct a single empirical mean (S11), construct K independent sample means and form their median:

$$\hat{o}_i(N,K) = \mathrm{median}\left\{\hat{o}_i^{(1)}(N,1), \ldots, \hat{o}_i^{(K)}(N,1)\right\} \quad \text{where} \quad \hat{o}_i^{(k)} = \tfrac{1}{N}\sum_{j=N(k-1)+1}^{Nk} \mathrm{tr}\left(O_i\hat{\rho}_j\right) \tag{S12}$$

for $1 \leq k \leq K$. This estimation technique requires NK samples in total, but it is much more robust with respect to outlier corruption. Indeed, $|\hat{o}(N,K) - \mathrm{tr}(O\rho)| > \epsilon$ if and only if more than half of the empirical means individually deviate by more than ϵ. The probability associated with such an undesirable event decreases exponentially with the number of batches K. This results in an exponential improvement over sample mean estimation in terms of failure probability. The main result of this work capitalizes on this improvement.

Theorem S1. *Fix a measurement primitive $\mathcal{U}$, a collection $O_1, \ldots, O_M$ of $2^n \times 2^n$ Hermitian matrices and accuracy parameters $\epsilon, \delta \in [0,1]$. Set*

$$K = 2\log(2M/\delta) \quad \text{and} \quad N = \frac{34}{\epsilon^2} \max_{1\leq i\leq M} \|O_i - \tfrac{\mathrm{tr}(O_i)}{2^n}\mathbb{I}\|^2_{\mathrm{shadow}}, \tag{S13}$$

where $\|\cdot\|_{\mathrm{shadow}}$ denotes the norm defined in Eq. (S7)*. Then, a collection of NK independent classical shadows allow for accurately predicting all features via median of means prediction* (S12)*:*

$$|\hat{o}_i(N,K) - \mathrm{tr}\left(O_i\rho\right)| \leq \epsilon \quad \text{for all } 1 \leq i \leq M \tag{S14}$$

with probability at least $1-\delta$.

Proof. The claim follows from combining the variance estimates from Lemma S1 with a rigorous performance guarantee for median of means estimation [36, 47]: Let X be a random variable with variance σ^2. Then, K independent sample means of size $N = 34\sigma^2/\epsilon^2$ suffice to construct a median of means estimator $\hat{\mu}(N,K)$ that obeys $\Pr\left[|\hat{\mu}(N,K) - \mathbb{E}\left[X\right]| \geq \epsilon\right] \leq 2\mathrm{e}^{-K/2}$ for all $\epsilon > 0$. The parameters N and K are chosen such that this general statement ensures $\Pr\left[|\hat{o}_i(N,K) - \mathrm{tr}\left(O_i\rho\right)| \geq \epsilon\right] \leq \frac{\delta}{M}$ for all $1 \leq i \leq M$. Apply a union bound over all M failure probabilities to deduce the claim. □

Remark S1 (Constants in Theorem S1)**.** *The numerical constants featuring in N and K result from a conservative (worst case) argument that is designed to be simple, not tight. We expect that the actual constants are* much smaller *in practice.*

Each classical shadow is the result of a single quantum measurement on ρ. Viewed from this angle, Theorem S1 asserts that a total of

$$N_{\mathrm{tot}} = \mathcal{O}\left(\frac{\log(M)}{\epsilon^2} \max_{1\leq i\leq M} \left\|O_i - \tfrac{\mathrm{tr}(O_i)}{2^n}\mathbb{I}\right\|^2_{\mathrm{shadow}}\right) \qquad \text{(sample complexity)} \tag{S15}$$

measurement repetitions suffice to accurately predict a collection of M linear target functions $\mathrm{tr}(O_i\rho)$.

Importantly, this sample complexity only scales logarithmically in the number of target functions M. Moreover, the problem dimension 2^n does not feature explicitly. The sample complexity does, however, depend on the measurement primitive via the norm $\|\cdot\|_{\mathrm{shadow}}$. This term reflects expressiveness and structure of the measurement primitive in question. This subtle point is best illustrated with two concrete examples. We defer technical derivations to subsequent sections and content ourselves with summarizing the important aspects here.

Example 1: Random Clifford measurements Clifford circuits are generated by CNOT, Hadamard and Phase gates and form the group $\mathrm{Cl}(2^n)$. The "random global Clifford basis" measurement primitive — $\mathcal{U} = \mathrm{Cl}(2^n)$ (endowed with uniform weights) — implies the following simple expression for classical shadows and the associated norm $\|\cdot\|_{\mathrm{shadow}}$:

$$\hat{\rho} = (2^n+1)U^\dagger|\hat{b}\rangle\langle\hat{b}|U - \mathbb{I} \quad \text{and} \quad \left\|O - \tfrac{\mathrm{tr}(O)}{2^n}\mathbb{I}\right\|^2_{\mathrm{shadow}} \leq 3\mathrm{tr}(O^2). \tag{S16}$$

We refer to Supplementary Section 5 B for details and proofs. Combined with Eq. (S15), this ensures that $\mathcal{O}(\log(M)\max_i \mathrm{tr}(O_i^2)/\epsilon^2)$ random global Clifford basis measurements suffice to accurately predict M linear functions. This prediction technique is most powerful, when the target functions have constant Hilbert-Schmidt norm. In this case, the sample rate is completely independent of the problem dimension 2^n. Prominent examples include estimating quantum fidelities (with pure states), or entanglement witnesses.

Example 2: Random Pauli measurements Although (global) Clifford circuits are believed to be much more tractable than general quantum circuits, they still feature entangling gates, like CNOT. Such gates are challenging to implement reliably on today's devices. The "random Pauli basis" measurement primitive takes this serious drawback into account and assumes that one is only able to apply single-qubit Clifford gates, i.e. $U = U_1 \otimes \cdots \otimes U_n \sim \mathcal{U} = \mathrm{Cl}(2)^{\otimes n}$ (endowed with uniform weights). This is equivalent to assuming that we can perform arbitrary Pauli (basis) measurements, i.e., measuring each qubit in the X-, Y- and Z-basis, respectively. Such basis measurements decompose nicely into tensor products $(U|\hat{b}\rangle = \bigotimes_{j=1}^n U_j|b_j\rangle$ for $b = (b_1, \ldots, b_n) \in \{0,1\}^n)$ and respect locality. The associated classical shadows and the norm $\|\cdot\|_{\mathrm{shadow}}$ inherit these desirable features:

$$\hat{\rho} = \bigotimes_{j=1}^n \left(3U_j^\dagger|\hat{b}_j\rangle\langle\hat{b}_j|U_j - \mathbb{I}\right) \quad \text{and} \quad \left\|O - \tfrac{\mathrm{tr}(O)}{2^n}\right\|^2_{\mathrm{shadow}} \leq 4^{\mathrm{locality}(O)}\|O\|^2_\infty. \tag{S17}$$

Here, locality(O) counts the number of qubits on which O acts nontrivially. We refer to Supplementary Section 5 C for details and proofs. Combined with Eq. (S15) this ensures that $\mathcal{O}\left(\log(M)4^k/\epsilon^2\right)$ local Clifford (Pauli) basis measurements suffice to predict M bounded observables that are at most k-local. For observables that are the tensor product of k single-qubit observables, the sample complexity can be further improved to $\mathcal{O}\left(\log(M)3^k/\epsilon^2\right)$. This prediction technique is most powerful when the target functions do respect some sort of locality constraint. Prominent examples include k-point correlators, or individual terms in a local Hamiltonian.

Discussion and information-theoretic optimality These two examples complement each other nicely. Random Clifford measurements excel at performing useful subroutines in quantum computing and communication tasks, such as certifying (global) entanglement, which will be feasible using sufficiently advanced hardware. Their practical utility, however, hinges on the ability to execute circuits with many entangling gates. Random Pauli measurements, on the other hand, are much less demanding from a hardware perspective. In today's NISQ era, local Pauli operators can be accurately measured using available hardware platforms. While not well-suited for predicting global features, Pauli measurements excel at making local predictions. Furthermore, for both kinds of randomized measurements, linear prediction based on classical shadows saturates fundamental lower bounds from information theory.

Theorem S2 (random Clifford measurements; informal version)**.** Any *procedure based on a fixed set of single-copy measurements that can predict, with additive error ϵ, M arbitrary linear functions* $\mathrm{tr}(O_i\rho)$*, requires at least* $\Omega(\log(M)\max_i \mathrm{tr}(O_i^2)/\epsilon^2)$ *copies of the state ρ.*

Theorem S3 (random Pauli measurements; informal version)**.** Any *procedure based on a fixed set of single-copy local measurements that can predict, with additive error ϵ, M arbitrary k-local linear functions* $\mathrm{tr}(O_i\rho)$*, requires at least* $\Omega(\log(M)3^k/\epsilon^2)$ *copies of the state ρ.*

We refer to Supplementary Section 7 (Clifford) and 8 (Pauli) for further context, details and proofs. In the random Pauli basis measurement setting, classical shadows provably saturate this lower bound only for tensor product observables. For general k-local observables, there is a small discrepancy between 4^k (upper bound) and 3^k (lower bound).

C. Predicting nonlinear functions with classical shadows

Feature prediction with classical shadows readily extends beyond the linear case. Here, we shall focus on quadratic functions, but the procedure and analysis readily extend to higher order polynomials. Every quadratic function in an unknown state ρ can be recast as a linear function acting on the tensor product $\rho\otimes\rho$:

$$\hat{o}_i = \mathrm{tr}\left(O_i\rho\otimes\rho\right) \quad 1\leq i\leq M. \tag{S18}$$

An immediate generalization of linear feature prediction with classical shadows suggests the following procedure. Take two independent snapshots $\hat{\rho}_1,\hat{\rho}_2$ of the unknown state ρ and set

$$\hat{o}_i = \mathrm{tr}\left(O_i\hat{\rho}_1\otimes\hat{\rho}_2\right) \quad \text{such that} \quad \mathbb{E}\hat{o}_i = \mathrm{tr}\left(O_i\mathbb{E}\hat{\rho}_1\otimes\mathbb{E}\hat{\rho}_2\right) = \mathrm{tr}\left(O_i\rho\otimes\rho\right) = o_i. \tag{S19}$$

This random variable is designed to yield the correct target function in expectation. Similar to linear function prediction we can boost convergence to this desired target by forming empirical averages. To make the best of use of N samples, we average over all $N(N-1)$ (distinct) pairs:

$$\hat{o}_i(N,1) = \frac{1}{N(N-1)}\sum_{j\neq l}\mathrm{tr}\left(O_i\hat{\rho}_j\otimes\hat{\rho}_l\right). \tag{S20}$$

This idea provides a systematic approach for constructing estimators for nonlinear (polynomial) functions. Estimators of this form always yield the desired target in expectation. For context, we point out that the estimator (S20) closely resembles the sample variance, while estimators of higher order polynomials are known as *U-statistics* [33]. Fluctuations of $\hat{o}_i(N,1)$ around its desired expectation are once more controlled by the variance. U-statistics estimators are designed to minimize this variance and therefore considerably boost the rate of convergence.

Lemma S2. *Fix O and a sample size N. Then, the variance of the U-statistics estimator* (S20) *obeys*

$$\mathrm{Var}[\hat{o}(N,1)] \leq \frac{2}{N}\left(\mathrm{Var}[\mathrm{tr}(O\hat{\rho}_1\otimes\rho)] + \mathrm{Var}[\mathrm{tr}(O\rho\otimes\hat{\rho}_1)] + \frac{1}{N}\mathrm{Var}[\mathrm{tr}(O\hat{\rho}_1\otimes\hat{\rho}_2)]\right). \tag{S21}$$

We emphasize that this variance decreases with the number of samples N. This sets the stage for successful quadratic function prediction with classical shadows. Similar to the linear case, we will not use all samples to construct a single U-statistics estimator. Instead, we construct K of them and form their median:

$$\begin{aligned}\hat{o}_i(N,K) =& \text{median}\left\{\hat{o}_i^{(1)}(N,1),\ldots,\hat{o}_i^{(K)}(N,1)\right\}, \quad \text{where}\\ \hat{o}_i^{(k)}(N,1) =& \frac{1}{N(N-1)} \sum_{\substack{j\neq l\\ j,l\in\{N(k-1)+1,\ldots,Nk\}}} \operatorname{tr}\left(O_i\hat{\rho}_j\otimes\hat{\rho}_l\right) \quad \text{for } 1\leq k\leq K. \end{aligned} \tag{S22}$$

This renders the entire estimation procedure more robust to outliers and exponentially suppresses failure probabilities.

Theorem S4. *Fix a measurement primitive $\mathcal{U}$, a collection $O_1,\ldots,O_M$ of (quadratic) target functions and accuracy parameters $\epsilon,\delta\in[0,1]$. Set*

$$\begin{aligned}K =& 2\log(2M/\delta) \quad \textit{and}\\ N =& \frac{34}{\epsilon^2}\max_{1\leq i\leq M} 8\times\max\left(\operatorname{Var}[\operatorname{tr}(O_i\rho\otimes\hat{\rho}_1)], \operatorname{Var}[\operatorname{tr}(O_i\hat{\rho}_1\otimes\rho)], \sqrt{\operatorname{Var}[\operatorname{tr}(O_i\hat{\rho}_1\otimes\hat{\rho}_2)]}\right). \end{aligned} \tag{S23}$$

Then, a collection of NK independent classical shadows allow for accurately predicting all quadratic features via the median of U-statistics estimators (S22)*:*

$$|\hat{o}_i(N,K) - \operatorname{tr}\left(O_i\rho\otimes\rho\right)| \leq \epsilon \quad \textit{for all } 1\leq i\leq M \tag{S24}$$

with probability at least $1-\delta$.

Proof. The proof is similar to the argument for linear prediction. We combine the bound on the variance of U-statistics estimators from Lemma S2 with a rigorous performance guarantee for median estimation [36, 47]. Let Z be a random variable with variance at most $\epsilon^2/34$. Then, setting $\hat{\mu}=\text{median}\{Z_1,\ldots,Z_k\}$ produces an estimator that obeys $\Pr\left[|\hat{\mu}-\mathbb{E}[Z]|\geq\epsilon\right]\leq 2\mathrm{e}^{-K/2}$. The parameter N is chosen ensure that each $\hat{o}_i^{(k)}(N,1)$ has variance at most $\epsilon^2/34$. The parameter K is chosen such that each probability of failure is at most δ/M. The advertised statement then follows from taking a union bound over all M target estimations. □

Remark S2 (Constants in Theorem S4)**.** *The numerical constants featuring in N and K result from a conservative (worst case) argument that is designed to be simple, not tight. We expect that the actual constants are* much smaller *in practice.*

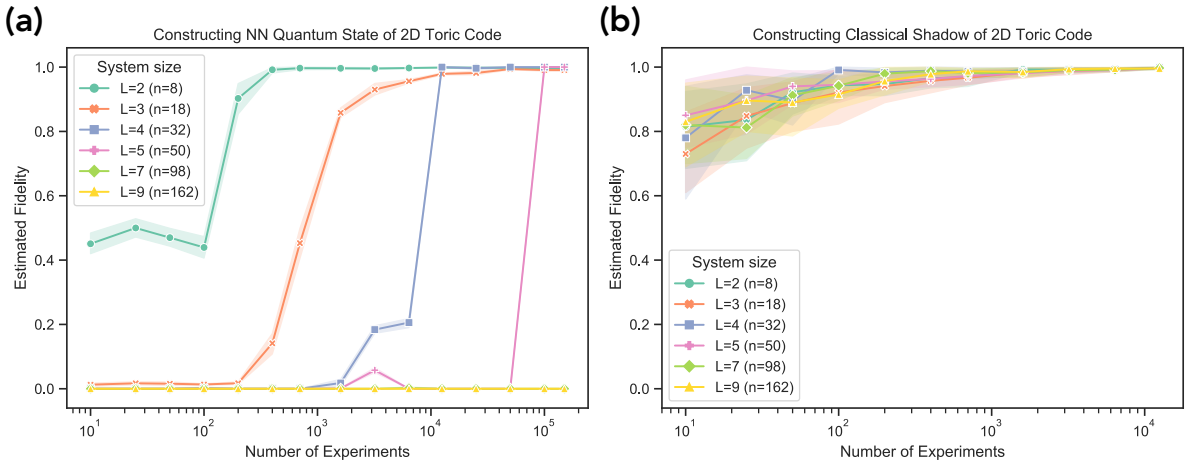
Theorem S4 is a general statement that provides upper bounds for the sample complexity associated with predicting quadratic target functions:

$$N_{\text{tot}} = \mathcal{O}\left(\frac{\log(M)}{\epsilon^2}\max_{1\leq i\leq M}\max\left(\operatorname{Var}[\operatorname{tr}(O_i\rho\otimes\hat{\rho}_1)], \operatorname{Var}[\operatorname{tr}(O_i\hat{\rho}_1\otimes\rho)], \sqrt{\operatorname{Var}[\operatorname{tr}(O_i\hat{\rho}_1\otimes\hat{\rho}_2)]}\right)\right) \tag{S25}$$

independent randomized measurements suffice to accurately predict a collection of M nonlinear target functions $\operatorname{tr}(O_i\rho\otimes\rho)$. This sampling rate once more depends on the measurement primitive and it is instructive to consider concrete examples.

Example 1: Random Pauli measurements We first discuss the practically more relvant example for today's NISQ era: classical shadows constructed from random single-qubit Pauli basis measurements. This measurement primitive remains well-suited for predicting *local* quadratic features $\operatorname{tr}(O\rho\otimes\rho)$. Suppose that O acts nontrivially on k qubits in the first state copy and on k qubits in the second state copy. Thus, when viewed as an observable for a $2n$-qubit system, O is $2k$-local. A technical argument shows that the maximum of the variances in Equation (S25) is bounded by 4^k. We emphasize that this scaling is much better than the naive guess 4^{2k} – one of the key advantages of U-statistics. Hence we only need a total number of $N_{\text{tot}}=\mathcal{O}(\log(M)4^k/\epsilon^2)$ random Pauli basis measurements to predict M quadratic functions $\operatorname{tr}(O_i\rho\otimes\rho)$. An important concrete application of this procedure is the prediction of subsystem Rényi-2 entanglement entropies.

Example 2: Random Clifford measurements Theorem S4 also applies to the global Clifford measurement primitive. There, the maximum of the variances in Equation (S25) can be bounded by $\sqrt{9+6/2^n}\operatorname{tr}(O_i^2)\simeq 3\operatorname{tr}(O_i^2)$. Hence we only need a total number of $N_{\text{tot}}=\mathcal{O}(\log(M)\max_i\operatorname{tr}(O_i^2)/\epsilon^2)$ random Clifford basis measurements to predict M quadratic functions $\operatorname{tr}(O_i\rho\otimes\rho)$. While a clean extension of linear feature prediction with Clifford basis measurements, the applicability of this result seems somewhat limited. Interesting global quadratic features tend to have prohibitively large Hilbert-Schmidt norms. The purity $\operatorname{tr}(\rho^2)$ provides an instructive non-example. It can be written as $\operatorname{tr}\left(S\rho\otimes\rho\right)$, where $S|\psi\rangle\otimes|\phi\rangle=|\phi\rangle\otimes|\psi\rangle$ denotes the swap operator. Alas, $\operatorname{tr}(S^2)=\operatorname{tr}(\mathbb{I})=2^n$ which scales exponentially in the number of qubits. Nonetheless, quadratic feature prediction with Clifford measurements is by no means useless. For instance, it can help provide statistical *a posteriori* guarantees on the quality of linear feature prediction — for example, by estimating sample variances to construct confidence intervals.



Supplementary Figure 1: *Comparison between classical shadow and neural network tomography (NNQST); toric code.* **(a)** *(Left)*: Number of measurements required for neural network tomography to identify a particular toric-code ground state. We use classical fidelity for NNQST, which is an upper bound for quantum fidelity. **(b)** *(Right)*: Performance of classical shadows for the same problem. We use quantum fidelity for classical shadows. The shaded regions are the standard deviation of the estimated fidelity over ten runs.

2. ADDITIONAL NUMERICAL EXPERIMENTS

In this section we report additional numerical experiments that demonstrate the viability of linear feature prediction with classical shadows. We focus on the Clifford basis measurement primitive, *i.e.* applying a random Clifford circuit to ρ and then measuring in the computational basis.

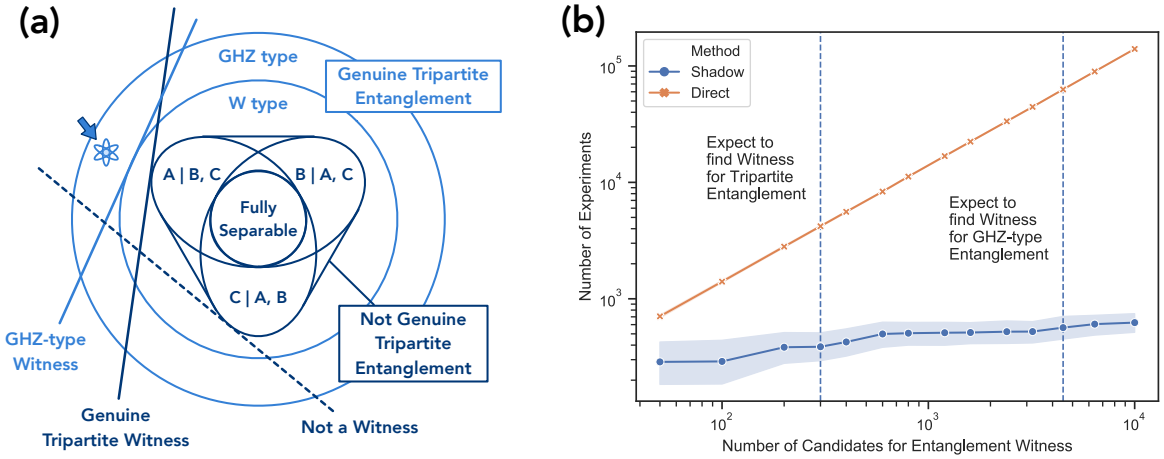
A. Direct fidelity estimation for the toric code ground state

In the main text, we have considered direct fidelity estimation for GHZ states and compared it with neural network quantum state tomography (NNQST). While highly instructive from a theoretical perspective, GHZ states comprised of 100 qubits are very fragile and challenging to implement in practice. To conduct experiments for more physical target states, we consider *Toric code ground states* [18]. Not only are they the most prominent example of a topological quantum error correcting code and thus highly relevant for quantum computing devices. They also correspond to ground states of a Hamiltonian: $H = -\sum_v A_v - \sum_p B_p$, where A_v and B_p denote vertex- and plaquette operators[1]. The ground space of H is four-fold degenerate and we select the superposition of all closed-loop configurations ($|\psi\rangle \propto \sum_{S:\text{ closed loop}} |S\rangle$) as a test state for both classical shadows and NNQST: how many measurement repetitions are required to accurately identify this toric code ground state with high fidelity? The results are shown in Supplementary Figure 1. Neural network tomography based on a deep generative model seems to require a number of samples that scales unfavorably in the system size n (left). In contrast, fidelity estimation with classical shadows is completely independent of the system size. The difficulty of NNQST in learning 2D toric code may be related to some observed failures of deep learning [56] for learning patterns with combinatorial structures. In Supplementary Section 4, we provide further evidence for potential difficulties when using machine learning approaches to reconstruct some simple quantum states due to a well-known computational hardness conjecture.

B. Witnesses for tripartite entanglement

Entanglement is at the heart of virtually all quantum communication and cryptography protocols and an important resource for quantum technologies in general. This renders the task of detecting entanglement

[1] A_v is the product of four Pauli-X operators around a vertex v, while B_p is the product of four Pauli-Z operators around the plaquette p.



Supplementary Figure 2: *Detection of GHZ-type entanglement for 3-qubit states.*
(a) *(Left)*: Schematic illustration of 3-partite entanglement. Entanglement witnesses are linear functions that separate part of one entanglement class from all other classes.
(b) *(Right)*: Number of entanglement witnesses vs. number of experiments required to accurately estimate all of them. The dashed lines represent the expected number of (random) entanglement witnesses required to detect genuine three-partite entanglement and GHZ-type entanglement in a randomly rotated GHZ state. The shaded region is the standard deviation of the required number of experiments over ten independent repetitions of the entire setup.

important both in theory and practice [24, 31]. While bipartite entanglement is comparatively well-understood, multi-partite entanglement has a much more involved structure. Already for $n = 3$ qubits, there is a variety of inequivalent entanglement classes. These include fully-separable, as well as bi-separable states, W-type states and finally GHZ-type states. The relations between these classes are summarized in Supplementary Figure 2 and we refer to [4] for a complete characterization. Despite this increased complexity, entanglement witnesses remain a simple and useful tool for testing which class a certain state ρ belongs to. However, any given entanglement witness only provides a one-sided test – see Supplementary Figure 2 (left) for an illustration – and it is often necessary to compute multiple witnesses for a definitive answer.

Classical shadows based on random Clifford measurements can considerably speed up this search: according to Theorem S1 a classical shadow of moderate size allows for checking an entire list of fixed entanglement witnesses simultaneously. Supplementary Figure 2 (right) underscores the economic advantage of such an approach over measuring the individual witnesses directly. Directly measuring M different entanglement witnesses requires a number of quantum measurements that scales (at least) linearly in M. In contrast, classical shadows get by with $\log(M)$-many measurements only.

More concretely, suppose that the state to be tested is a local, random unitary transformation of the GHZ state. Then, this state is genuinely tripartitely entangled and moreover belongs to the GHZ class. The dashed vertical lines in Supplementary Figure 2 (right) denote the expected number of (randomly selected) witnesses required to detect genuine tripartite entanglement (first) and GHZ-type entanglement (later). From the experiment, we can see that classical shadows achieve these thresholds with an exponentially smaller number of samples than the naive direct method. Finally, classical shadows are based on random Clifford measurements and do not depend on the structure of the concrete witness in question. In contrast, direct estimation crucially depends on the concrete witness in question and may be considerably more difficult to implement.

3. RELATED WORK

General quantum state tomography The task of reconstructing a full classical description — the density matrix ρ — of a d-dimensional quantum system from experimental data is one of the most fundamental problems in quantum statistics, see e.g. [5, 7, 29, 34] and references therein. Sample-optimal protocols, i.e. estimation techniques that get by with a minimal number of measurement repetitions, have only been developed recently. Information-theoretic bounds assert that of order $\mathrm{rank}(\rho)d$ state copies are necessary to fully reconstruct ρ [32]. Constructive protocols [32, 49] saturate this bound, but require entangled circuits and measurements that act on all state copies simultaneously. More tractable single-copy measurement procedures require of order $\mathrm{rank}(\rho)^2 d$ measurements [32]. This more stringent bound is saturated by low rank matrix recovery [22, 42, 43]

and projected least squares estimation [30, 58].

These results highlight an exponential bottleneck for tomography protocols that work in full generality: at least $d = 2^n$ copies of an unknown n-qubit state are necessary. This exponential scaling extends to the computational cost associated with storing and processing the measurement data.

Matrix product state tomography Restricting attention to highly structured subsets of quantum states sometimes allows for overcoming the exponential bottleneck that plagues general tomography. Matrix product state (MPS) tomography [16] is the most prominent example for such an approach. It only requires a polynomial number of samples, provided that the underlying quantum state is well approximated by a MPS with low bond dimension. In quantum many-body physics this assumption is often justifiable [45]. However, MPS representations of general states have exponentially large bond dimension. In this case, MPS tomography offers no advantage over general tomography. Similar ideas could also be extended to multi-scale entangled states (MERA) tomography [44].

Neural network tomography Recently, machine learning has also been applied to the problem of predicting features of a quantum systems. These approaches construct a classical representation of the quantum system by means of a deep neural network that is trained by feeding in quantum measurement outcomes. Compared to MPS tomography, neural network tomography may be more broadly applicable [13, 25, 59]. However, the actual class of systems that can be efficiently represented, reconstructed and manipulated is still not well understood.

Compressed classical description of quantum states To circumvent the exponential scaling in representing quantum states, Gosset and Smolin [26] have proposed a stabilizer sketching approach that compresses a classical description of quantum states to an accurate sketch of subexponential size. This approach bears some similarity with classical shadows based on random Clifford measurements. However, stabilizer sketching requires a fully-characterized classical description of the state as an input. So, it still suffers from an exponential scaling in the resources used in practice. Recently, Paini and Kalev [50] have proposed an approximate classical description of a quantum state that can estimate any observable written as a sum of Pauli operators, with a precision that depends on the number of samples and on a properly-defined semi-norm of the observable. The procedure performs Haar-random single-qubit rotations followed by computational basis measurements on each copy of the system. This is similar to classical shadows based on random Pauli measurements. In our approach, the Haar-random single-qubit rotations [50] are replaced by random single-qubit Clifford rotations, or – equivalently – measuring each qubit in a random Pauli basis. This simplification may be viewed as a partial derandomization and works, because the (single-qubit) Clifford group forms a 3-design [41, 60, 62]. We also employ median-of-means estimation to achieve a stronger concentration to the expected value.

Direct fidelity estimation Direct fidelity estimation is a procedure that allows for predicting a single pure target fidelity $\langle\psi|\rho|\psi\rangle$ up to accuracy ϵ. The best-known technique is based on few Pauli measurements that are selected randomly using importance sampling [17, 23]. The required number of samples depends on the target: it can range from a dimension-independent order of $1/\epsilon^2$ (if $|\psi\rangle$ is a stabilizer state) to roughly $2^n/\epsilon^4$ in the worst case.

Efficient estimation of local observables In quantum many-body physics, many interesting observables can be decomposed into local constituents. This renders the task of accurately predicting many local observables very important — both in theory and practice. A series of recent works [8, 14, 21, 37] propose different measurement strategies to measure many local observables simultaneously. All of them focus on estimating k-local Pauli observables up to accuracy ϵ. This would directly translate to an approximation error $2^k\epsilon$ for general k-local observables. For some measurement schemes, this general error bound seems unavoidable. But, for certain strategies a careful analysis could lead to an improved performance. The two works [8, 14] are based on properly analyzing the commutation relations between the k-local Pauli observables of interest. Subsequently, one can group commuting observables together and measure them all at once. Different from this more standardized strategy, [37] uses entangled Bell-basis measurements, and [21] is based on randomized measurements to efficiently measure local observables. The prior earlier works [8, 14] have worse performance compared to the more recent two [21, 37]. While the latter two procedures are seemingly different from prediction with classical shadows (Pauli measurements), the sample complexities associated with all three approaches are comparable. Derandomizing classical shadows, however, could considerably reduce the number of measurements required. We will address such a substantial and practical improvement in upcoming work.

Shadow tomography Shadow tomography aims at simultaneously estimating the outcome probabilities associated with M 2-outcome measurements up to accuaracy ϵ: $p_i(\rho) = \mathrm{tr}(E_i\rho)$, where each E_i is a positive semidefinite matrix with operator norm at most one [1, 3, 10]. This may be viewed as a generalization of fidelity estimation. The best existing result is due to Aaronson and Rothblum [3]. They showed that $N = \tilde{\mathcal{O}}\left(\log(M)^2\log(d)^2/\epsilon^8\right)$ copies of the unknown state suffice to achieve this task [2]. Broadly speaking,

[2] The scaling symbol $\tilde{\mathcal{O}}$ suppresses logarithmic expressions in other problem-specific parameters.

their protocol is based on performing gentle 2-outcome measurements one-by-one and subsequently (partially) reversing the damage to the quantum state caused by the measurement. This task is achieved by explicit quantum circuits of exponential size that act on all copies of the unknown state simultaneously. This rather intricate procedure bypasses the no-go result advertised in Theorem 2 and results in a sampling rate that is independent of the 2-outcome measurements in question — only their cardinality M matters.

4. DETAILS REGARDING NUMERICAL EXPERIMENTS

A. Predicting quantum fidelities

This numerical experiment considers classical shadows based on random Clifford measurements. We exploit the Gottesman-Knill theorem for efficient classical computations. This well-known result states that Clifford circuits can be simulated efficiently on classical computers; see also [2] for an improved classical algorithm. This has allowed us to address rather large system sizes (more than 160 qubits). To test the performance of feature prediction with classical shadows we first have to simulate the (quantum) data acquisition phase. We do this by repeatedly executing the following (efficient) protocol:

1. Sample a Clifford unitary U from the Clifford group using the algorithm proposed in [39]. This Clifford unitary is parameterized by $(\alpha, \beta, \gamma, \delta, r, s)$ which fully characterize its action on Pauli operators:

$$UP_j^X U^\dagger = (-1)^{r_j} \Pi_{i=1}^n (P_i^X)^{\alpha_{ji}} (P_i^Z)^{\beta_{ji}} \quad \text{and} \quad UP_j^Z U^\dagger = (-1)^{s_j} \Pi_{i=1}^n (P_i^X)^{\gamma_{ji}} (P_i^Z)^{\delta_{ji}} \tag{S26}$$

   for all $j = 1, \ldots, n$. Here, P_j^X, P_j^Z are the Pauli X, Z-operators acting on the j-th qubit, and $\alpha_{ji}, \beta_{ji}, \gamma_{ji}, \delta_{ji}, r_j, s_j \in \{0, 1\}$.

2. Given a unitary U parameterized by $(\alpha, \beta, \gamma, \delta, r, s)$, we can apply U on any stabilizer state by changing the stabilizer generators and the destabilizers as defined in [2].

3. A computational basis measurement can be simulated using the standard algorithm provided in [2].

Although originally designed for pure target states $|\psi_i\rangle\langle\psi_i|$, we can readily extend this strategy to mixed states $\rho = \sum_i p_i |\psi_i\rangle\langle\psi_i|$. Operationally speaking, mixed states arise from sampling from a pure state ensemble. This mixing process can be simulated efficiently on classical machines.

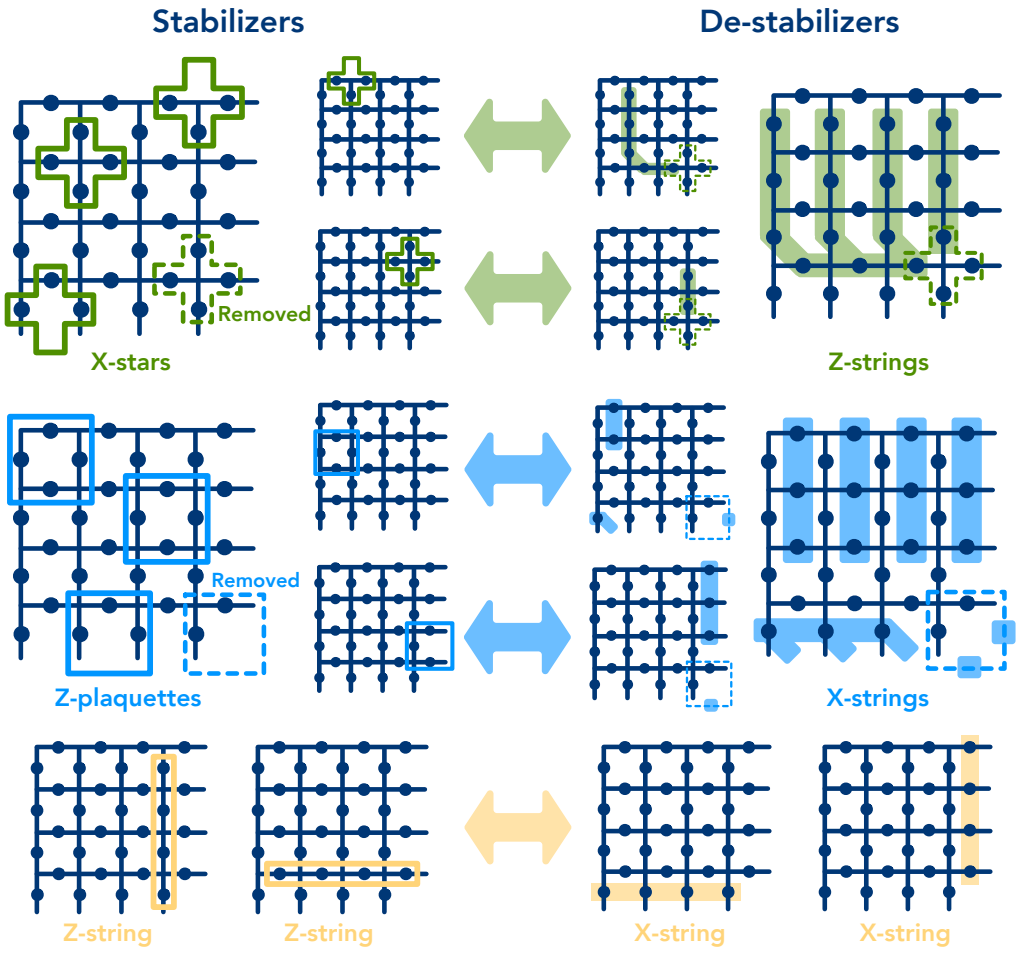
For neural network quantum state tomography, we use the open-source code provided by the authors [13]. The main challenge is generating training data, i.e. simulating measurement outcomes. For pure and noisy GHZ states, we use the tetrahedral POVM [13]. For the toric code ground state, we use the Psi2 POVM (which is a measurement in the computational (Z-) basis). Note that measuring in the Z-basis is not a tomographically complete measurement, but we found machine learning models to perform better using Psi2. This is possibly because the pattern is much more obvious (closed-loop configurations) and the figure of merit used in NNQST is a classical fidelity.

A concrete algorithm for creating training data for pure GHZ states is included in the aforementioned open-source implementation of [13]. It uses matrix product states to simulate quantum measurements efficiently. The training data for noisy GHZ states is a slight modification of the existing code. With probability $1-p$, we sample a measurement outcome from the original state $|\psi_{\mathrm{GHZ}}^+\rangle = \frac{1}{\sqrt{2}}(|0\rangle^{\otimes n} + |1\rangle^{\otimes n})$. And with probability p, we sample a measurement outcome from $|\psi_{\mathrm{GHZ}}^-\rangle = \frac{1}{\sqrt{2}}(|0\rangle^{\otimes n} - |1\rangle^{\otimes n})$ (phase error). Since the figure of merit is the fidelity with the pure GHZ state in both pure and noisy GHZ experiment, we reuse the implementation provided in [13].

Creating training data for toric code is somewhat more involved. The goal is to sample a closed-loop configuration on a 2D torus uniformly at random. This can again be done using classical simulations of stabilizer states [2]. The main technical detail is to create a tableau that contains both the stabilizer and the de-stabilizer for the state in question. The rich structure of the toric code renders this task rather easy. The stabilizers are the X-stars and the Z-plaquettes, with two Z-strings over the two loops of the torus. The de-stabilizer of each stabilizer is a Pauli-string that anticommutes with the stabilizer, but commutes with other stabilizers and other de-stabilizers. The full set of stabilizers and de-stabilizers for the toric code can be seen in Supplementary Figure 3.

B. Potential obstacles for learning certain quantum states

In our numerical studies, we have seen that neural network quantum state tomography based on deep generative models seems to have difficulty learning toric code ground states.



Supplementary Figure 3: Stabilizers and de-stabilizers of the toric code that encodes $|00\rangle$.

Here, we take a closer look at this curious aspect and construct a simple class of quantum states where efficient learning of the quantum state from the measurement data would violate a well-known computational hardness conjecture. First of all, each computational (Z-) basis measurement of the toric code produces a random bit-string. Most bits are sampled uniformly at random from $\{0,1\}$ and the remaining bits are binary functions that only depend on these random bits. Consider a simple class of quantum states that mimic this property. Given $a \in \{0,1\}^{n-1}$ and $f_a(x) = \sum_i a_i x_i \pmod 2$, we define $|a\rangle = \frac{1}{\sqrt{2^{n-1}}} \sum_{x\in\{0,1\}^{n-1}} |x\rangle \otimes |f_a(x)\rangle$. Such states can be created by preparing $|+\rangle$ on the first $n-1$ qubits, $|0\rangle$ on the n-th qubit followed by CNOT gates between i-th qubit and n-th qubit for every $a_i = 1$. Measuring $|a\rangle$ in the computational (Z-) basis is equivalent to sampling the first $n-1$ bits x uniformly at random. The final bit is characterized by the deterministic formula $f_a(x)$. Now, consider a (globally) depolarized version of this pure state:

$$\rho_a = \mathcal{D}_\eta(|a\rangle\langle a|) = (1-\eta)|a\rangle\langle a| + \tfrac{\eta}{2^n}\mathbb{I}^{\otimes n} \quad \text{for some } \eta \in (0,1). \tag{S27}$$

One of the most widely used conjectures for building post-quantum cryptography is the hardness of learning with error (LWE) [54]. LWE considers the task of learning a linear n-ary function f over a finite ring from noisy data samples $(x, f(x)+\eta)$, where x is sampled uniformly at random and η is some independent error. An efficient learning algorithm for LWE will be able to break many post-quantum cryptographic protocals that are believed to be hard even for quantum computers. The simplest example of LWE is called learning parity with error, where $f(x) = \sum_i a_i x_i \pmod 2$ for $x \in \{0,1\}^n$ and some unknown $a \in \{0,1\}^n$. Learning parity with error is also conjectured to be computationally hard [6]. Since learning $|a\rangle$ from computational (Z-) basis measurements on ρ_a is equivalent to learning parity with error, it is unlikely there will be a neural network approach that can learn ρ_a efficiently.

C. Predicting witnesses for tripartite entanglement

This numerical experiment considers classical shadows based on random Clifford measurements. The numerical studies regarding entanglement witnesses are based locally rotated 3-qubit ($n = 3$) GHZ states:

$$|\psi\rangle = U_A \otimes U_B \otimes U_C |\psi^+_{\mathrm{GHZ}}\rangle \quad \text{where } U_A, U_B, U_C \text{ are random single-qubit rotations.} \tag{S28}$$

For $\rho = |\psi\rangle\langle\psi|$, we hope to verify the tripartite entanglement present in the system. To this end, we consider a simple family of entanglement witnesses with compatible structure:

$$O := O(V_A, V_B, V_C) = V_A \otimes V_B \otimes V_C |\psi^+_{\mathrm{GHZ}}\rangle\langle\psi^+_{\mathrm{GHZ}}| V_A^\dagger \otimes V_B^\dagger \otimes V_C^\dagger. \tag{S29}$$

The single-qubit unitaries V_A, V_B, V_C parametrize different witnesses.

A complete characterization of entanglement in three-qubit systems can be found in Supplementary Figure 2. The expectation value of an entanglement witness $O(V_A, V_B, V_C)$ in the tripartite state ρ can certify that ρ belongs to a particular entanglement class. For example, it is known from the analysis in [4] that for any state ρ_s with only bipartite entanglement, $\mathrm{tr}\,(O\rho_s) \leq .5$, while for any state ρ_s with at most W-type entanglement, $\mathrm{tr}\,(O\rho_s) \leq .75$. Therefore verifying that $\mathrm{tr}\,(O\rho) > .5$ certifies that ρ has tripartite entanglement, while $\mathrm{tr}\,(O\rho) > .75$ certifies that ρ has GHZ-type entanglement.

After choosing random unitaries U_A, U_B, U_C to specify the GHZ-type state $|\psi\rangle$, we generate a list of random V_A, V_B, V_C to specify a set of potential entanglement witnesses for $|\psi\rangle$:

$$O_1 = O(V_{A,1}, V_{B,1}, V_{C,1}), \ldots, O_M = O(V_{A,M}, V_{B,M}, V_{C,M}). \tag{S30}$$

If the randomly generated $O_i = O(V_{A,i}, V_{B,i}, V_{C,i})$ satisfies $\mathrm{tr}(O_i\,|\psi\rangle\langle\psi|) > 0.5$, then O_i is an entanglement witness for genuine tripartite entanglement, and if $\mathrm{tr}(O_i\,|\psi\rangle\langle\psi|) > 0.75$, then O_i is a witness for GHZ-type entanglement. We can compute the expected number of random candidates we have to test to find an observable O such that $\mathrm{tr}(O\,|\psi\rangle\langle\psi|) > 0.5$ or $\mathrm{tr}(O\,|\psi\rangle\langle\psi|) > 0.75$; these numbers are indicated as the dashed lines on the right side of Supplementary Figure 2.

Given the list of randomly generated witness candidates $O_1, \ldots, O_M$, we would like to predict $\mathrm{tr}(O_i|\psi\rangle\langle\psi|)$ for all $1 \leq i \leq M$. The naive approach is to directly measure all observables (witnesses). We refer to this as the direct measurement approach. For this approach, we consider the number of total experiments required to estimate every $\mathrm{tr}(O_i|\psi\rangle\langle\psi|)$ up to an error 0.1. Note that the number of required samples may vary from witness to witness — it depends on the variance associated with the estimation. In the worst case, one would need ≈ 100 measurements for each witness candidate.

Instead of this direct measurement approach, one could use classical shadows (Clifford measurements) to predict *all* the observables (witnesses) $O_1, \ldots, O_M$ at once. Because, $\mathrm{tr}(O_i^2) = 1$ for al $1 \leq i \leq M$, the shadow norm obeys $\|O_i\|^2_{\mathrm{shadow}} \leq 3\,\mathrm{tr}\left(O_i^2\right) = 3$, according to the analysis in Supplementary Section 1 B. Hence Theorem 1 shows that classical shadows can predict the expectation values of many candidate witnesses very efficiently.

In the numerical experiment, we gradually increased the number of random Clifford measurements we use to construct classical shadows until the classical shadows could accurately predict all $\mathrm{tr}(O_i\,|\psi\rangle\langle\psi|)$ up to 0.1-error. The results are shown in Supplementary Figure 2. Because the system size is small ($n = 3$ qubits), we simulate the quantum experiments classically by storing and processing all $2^3 = 8$ amplitudes. In practice, one should use statistics, like sample variance estimation or the bootstrap [19], to determine confidence intervals and a posteriori guarantees. Quadratic function prediction with classical shadows (Clifford measurements) can be used to achieve this goal efficiently.

D. Predicting two-point correlation functions

Predicting two-point correlation function could be done efficiently using classical shadows based on random Pauli measurements. To facilitate direct comparison, this numerical experiment is designed to reproduce one of the core examples in in [13]. In particular, we use the same data, downloaded from `https://github.com/carrasqu/POVM_GENMODEL`. The classical shadow (based on random Pauli basis measurements) replaces the original machine learning based approach for predicting local observables. We use multi-core CPU for training and making prediction with the machine learning model. The reported time is the total CPU time. Predicting local observables O using the (Pauli) classical shadow can be done efficiently by creating the reduced density matrix ρ_A, where A is the subsystem O acts on. The reduced density matrix ρ_A can be created by simply neglecting the data for the rest of the system. Importantly, $\mathcal{M}^{-1}(U^\dagger|\hat{b}\rangle\langle\hat{b}|U)$ is never created as an $2^n \times 2^n$ matrix. Taking the inner product of ρ_A with the local observables O yields the desired result.

E. Predicting subsystem Rényi entanglement entropies

We consider classical shadows based on random Pauli measurements for predicting subsystem entanglement entropies. In the first part of the experiment, we consider the ground state of a disordered Heisenberg model. The associated Hamiltonian is $H = \sum_i J_i \langle S_i \cdot S_{i+1} \rangle$, where each J_i is sampled uniformly (and independently) from the unit interval $[0, 1]$. The approximate ground state is found by implementing the recursive procedure from [53]: identify the largest J_i, forming singlet for the connected sites, and reduce the system by removing J_i. We refer to [53] for details. In the experiment, we perform single-shot random Pauli basis measurements on the approximate ground state. I.e. we measure the state in a random Pauli basis only once and then choose a new random basis. However, in physical experiments, it is often easier to repeat a single Pauli basis measurement many times before re-calibrating to measure another Pauli basis. Performing a single random basis measurement for many repetitions can be beneficial experimentally compared to measuring a random basis every single time. Classical shadows (Pauli) are flexible enough to incorporate economic measurement strategies that take this discrepancy into account. We refer to the open source implementation in `https://github.com/momohuang/predicting-quantum-properties` for the exact details.

To obtain a reasonable benchmark, we compare this procedure with the approach proposed by Brydges *et al.* [12]. For a subsystem A comprised of k qubits, the approach proposed in [12] for predicting the Rényi entropy works as follows. First, one samples a random single-qubit unitary rotations independently for all k qubits. Then, one applies the single-qubit unitary rotation to the system and measures the system in the computational basis to obtain a string of binary values $s \in \{0, 1\}^k$. For each random unitary rotation, several repetitions are performed. The precise number of repetitions for a single random basis is a hyper-parameter that has to be optimized. The estimator for the Rényi entropy takes the following form:

$$\mathrm{tr}(\rho_A^2) = 2^k \sum_{s,s' \in \{0,1\}^k} (-2)^{-H(s,s')} \overline{P(s)P(s')}. \tag{S31}$$

The function $H(s, s')$ is the Hamming distance between strings s and s' (i.e, the number of positions at which individual bits are different), while $P(s)$ and $P(s')$ are the probabilities for measuring ρ and obtaining the outcomes s and s', respectively. The probability $P(s)$ is a function that depends on the randomly sampled single-qubit rotation. $\overline{P(s)P(s')}$ is the expectation of $P(s)P(s')$ averaged over the random single-qubit rotations.

The random single-qubit rotations could be taken as single-qubit Haar-random rotations or single-qubit random Clifford rotations. The latter choice is equivalent to random Pauli measurements – the measurement primitive we consider for classical shadows also. For the test cases we considered, using random Pauli measurements yields similar (and sometimes improved) performance compared to single-qubit Haar-random unitary rotation. This allows the approach by [12] and the procedure based on classical shadows to be compared on the same ground. We follow the strategy in [12] to estimate the formula in Eq. (S31). First, we sample N_U random unitary rotations. For each random unitary rotation, we perform N_M repetitions of rotating the system and measuring in the computational basis. The N_M measurement outcomes allow us to construct an empirical distribution for $P(s)$. Thus we could use the N_M measurement outcomes to estimate $2^k \sum_{s,s' \in \{0,1\}^k} (-2)^{-H(s,s')} P(s)P(s')$ for a single random unitary rotation. We then take the average over N_U different random unitary rotations. Choosing a suitable parameter for N_U and N_M is nontrivial. We employ the strategy advocated in [12] for finding the best parameter for N_U and N_M. This strategy is called grid search and is performed by trying many different choices for N_U, N_M and recording the best one.

F. Variational quantum simulation of the lattice Schwinger model

The application for variational quantum simulation uses classical shadows based on random Pauli measurements which is designed to predict a large number of local observables efficiently. It is based on the seminal work presented in [40]. After a Kogut-Susskind encoding to map fermionic configurations to a spin-1/2 lattice with an even number N of lattice sites and a subsequent Jordan-Wigner transform, the Hamiltonian becomes

$$\hat{H} = \underbrace{\frac{w}{2} \sum_{j=1}^{N-1} P_j^X P_{j+1}^X}_{\hat{\Lambda}_X} + \underbrace{\frac{w}{2} \sum_{j=1}^{N-1} P_j^Y P_{j+1}^Y}_{\hat{\Lambda}_Y} + \underbrace{\sum_{j=1}^{N} d_j P_j^z + \sum_{j=1}^{N-2} \sum_{j'=j+1}^{N-1} c_{j,j'} P_j^z P_{j'}^z}_{\hat{\Lambda}_Z}. \tag{S32}$$

Here, P_j^X, P_j^Y, P_j^Z denote Pauli-X, Y, Z operators acting on the j-th qubit ($1 \le j \le N$). This Hamiltonian has very advantageous structure. Each of the three contributions can be estimated by performing a single Pauli basis measurement (measure every qubit in the X basis to determine $\hat{\Lambda}_X$, measure every qubit in the Y basis to

determine $\hat{\Lambda}_Y$ and measure every qubit in the Z basis to determine $\hat{\Lambda}_Z$). The measurement of the Hamiltonian variance $\langle \hat{H}^2 \rangle - \langle \hat{H} \rangle^2$ is more complicated, because $\langle \hat{H}^2 \rangle$ does not decompose nicely. To determine its value, we must first measure $\hat{\Lambda}_X^2$, $\hat{\Lambda}_Y^2$ and $\hat{\Lambda}_Z^2$. This is the easy part, because 3 measurement bases once more suffice. However, in addition, we must also estimate the anti-commutators $\{\hat{\Lambda}_X, \hat{\Lambda}_Y\}, \{\hat{\Lambda}_X, \hat{\Lambda}_Z\}, \{\hat{\Lambda}_Y, \hat{\Lambda}_Z\}$. This may be achieved by measuring the following k-local observables (with k at most 4):

$$\begin{aligned}
&\{\hat{\Lambda}_X, \hat{\Lambda}_Y\} : P_j^X P_{j+1}^X P_{j'}^Y P_{j'+1}^Y, && \forall j, j' \in \{1, N-1\}, \text{ s.t. } j \neq j', j \neq j'+1, j+1 \neq j', \\
&\{\hat{\Lambda}_X, \hat{\Lambda}_Z\} : P_j^X P_{j+1}^X P_{j'}^Z P_{j''}^Z, && \forall j, j', j'' \in \{1, N-1\}, \text{ s.t. } j \neq j', j \neq j'', j+1 \neq j', j+1 \neq j'', j' < j'', \\
&\{\hat{\Lambda}_X, \hat{\Lambda}_Z\} : P_j^X P_{j+1}^X P_{j'}^Z, && \forall j, j' \in \{1, N-1\}, \text{ s.t. } j \neq j', j+1 \neq j', \\
&\{\hat{\Lambda}_Y, \hat{\Lambda}_Z\} : P_j^Y P_{j+1}^Y P_{j'}^Z P_{j''}^Z, && \forall j, j', j'' \in \{1, N-1\}, \text{ s.t. } j \neq j', j \neq j'', j+1 \neq j', j+1 \neq j'', j' < j'', \\
&\{\hat{\Lambda}_Y, \hat{\Lambda}_Z\} : P_j^Y P_{j+1}^Y P_{j'}^Z, && \forall j, j' \in \{1, N-1\}, \text{ s.t. } j \neq j', j+1 \neq j',
\end{aligned} \tag{S33}$$

Although local, estimating all observables of this form is the main bottleneck of the entire procedure. To minimize the number of measurement bases, the original work [40] has performed an analysis of symmetry in the lattice Schwinger model. First, the target Hamiltonian in Equation (S32) satisfies $[\hat{H}, \sum_i P_i^Z] = 0$, which corresponds to a charge conservation symmetry in the scalar fermionic field. [40] further consider a charge symmetry subspace with $\sum_i P_i^Z = 0$, which corresponds to a $\hat{CP}$ symmetry. In this subspace, we have $\langle \{\hat{\Lambda}_X, \hat{\Lambda}_Z\} \rangle = \langle \{\hat{\Lambda}_Y, \hat{\Lambda}_Z\} \rangle$. This ensures that we only have to estimate local observables corresponding to $\{\hat{\Lambda}_X, \hat{\Lambda}_Y\}$ and $\{\hat{\Lambda}_X, \hat{\Lambda}_Z\}$. In the original setup [40], this task was achieved by measuring roughly $2N$ bases in total. We refer to [40, Appendix B and Appendix C] for further details and explanation. We propose to replace this original approach by linear feature prediction with classical shadows (Pauli measurements).

For classical shadows based on random Pauli measurements, every measurement basis is an independent random X, Y, or Z measurement for every qubit. This randomized general purpose procedure does not take into account the fact that we want to measure a specific set of k-local observables given in Equation (S33). The derandomized version of classical shadows is based on the concept of pessimistic estimators [51, 57] (see also [61] for an application with quantum information context). It removes the original randomness by utilizing the knowledge of this specific set of k-local observables. When we throw a dice (or coin) to decide whether we want to measure in either, the $X-$, the $Y-$, or the $Z-$basis, the derandomized version would choose the measurement basis (X, Y, or Z) that would lead to the best expected performance on the set of k-local observables given in Equation (S33). The expected performance is computed based on random Pauli basis measurements and the analysis in Supplementary Section 1. The derandomized version of classical shadows would perform at least as well as the original randomized version. Furthermore, due to the dependence on the specific set of observables for choosing the measurement bases, the derandomized version can exploit advantageous structures in the set of observables we want to measure. As detailed in the main text, classical shadows based on random Pauli measurements provide improvement only for larger system sizes (more than 50 qubits). A derandomized version of classical shadows improves upon the randomized version and leads to a substantial improvement in efficiency and scalability over a wide range of system sizes. As an added benefit, derandomization can be completely automated and does not depend on the concrete set of target observables. We refer to `https://github.com/momohuang/predicting-quantum-properties` for a (roughly linear time) algorithm that derandomizes random Pauli measurements for any collection of target observables with Pauli structure.

5. ADDITIONAL COMPUTATIONS AND PROOFS FOR PREDICTING LINEAR FUNCTIONS

A. Background: Clifford circuits and the stabilizer formalism

Clifford circuits were introduced by Gottesman [27] and form an indispensable tool in quantum information processing. Applications range from quantum error correction [48], to measurement-based quantum computation [11, 52] and randomized benchmarking [20, 38, 46]. For systems comprised of n qubits, the Clifford group is generated by CNOT, Hadamard and phase gates. This results in a finite group of cardinality $2^{\mathcal{O}(n^2)}$ that maps (tensor products of) Pauli matrices to Pauli matrices upon conjugation. This underlying structure allows for efficiently storing and simulating Clifford circuits on classical computers – a result commonly known as Gottesman-Knill theorem. The n-qubit Clifford group $\mathrm{Cl}(2^n)$ also comprises a *unitary 3-design* [41, 60, 62]. Sampling Clifford circuits uniformly at random reproduces the first 3 moments of the full unitary group endowed with the Haar measure. For $k = 1, 2, 3$

$$\mathbb{E}_{U \sim \mathrm{Cl}(2^n)} \left(UXU^\dagger\right)^{\otimes k} = \int_{U(d)} (UAU^\dagger)^{\otimes k} \mathrm{d}\mu_{\mathrm{Haar}}(U) \quad \text{for all } 2^n \times 2^n \text{ matrices } A. \tag{S34}$$

The right hand side of this equation can be evaluated explicitly by using techniques from representation theory, see e.g. [28, Sec. 3.5]. This in turn yields closed-form expressions for Clifford averages of linear and quadratic operator-valued functions. Choose a unit vector $x \in \mathbb{C}^{2^n}$ and let $\mathbb{H}_{2^n}$ denote the space of Hermitian $2^n \times 2^n$ matrices. Then,

$$\mathbb{E}_{U\sim \mathrm{Cl}(2^n)} U^\dagger|x\rangle\langle x|U^\dagger\langle x|UAU^\dagger|x\rangle = \frac{A+\mathrm{tr}(A)\mathbb{I}}{(2^n+1)2^n} = \frac{1}{2^n}\mathcal{D}_{1/(2^n+1)}(A) \text{ for } A \in \mathbb{H}_{2^n}, \tag{S35}$$

$$\mathbb{E}_{U\sim \mathrm{Cl}(2^n)} U^\dagger|x\rangle\langle x|U\langle x|UB_0U^\dagger|x\rangle\langle x|UC_0U^\dagger|x\rangle = \frac{\mathrm{tr}(B_0C_0)\mathbb{I}+B_0C_0+C_0B_0}{(2^n+2)(2^n+1)2^n} \text{ for } B_0, C_0 \in \mathbb{H}_{2^n} \text{ traceless.} \tag{S36}$$

Here, $\mathcal{D}_p(A) = pA + (1-p)\frac{\mathrm{tr}(A)}{2^n}\mathbb{I}$ denotes a n-qubit depolarizing channel with loss parameter p. Linear maps of this form can be readily inverted. In particular,

$$\mathcal{D}^{-1}_{1/(2^n+1)}(A) = (2^n+1)A - \mathrm{tr}(A)\mathbb{I} \quad \text{for any } A \in \mathbb{H}_{2^n}. \tag{S37}$$

These closed-form expressions allow us to develop very concrete strategies and rigorous bounds for classical shadows based on (global and local) Clifford circuits.

B. Performance bound for classical shadows based on random Clifford measurements

Proposition S1. *Adopt a "random Clifford basis" measurement primitive, i.e. each rotation $\rho \mapsto U\rho U^\dagger$ is chosen uniformly from the n qubit Clifford group $\mathrm{Cl}(2^n)$. Then, the associated classical shadow is*

$$\hat{\rho} = (2^n+1)U^\dagger|\hat{b}\rangle\langle\hat{b}|U - \mathbb{I}, \tag{S38}$$

where $\hat{b} \in \{0,1\}^n$ is the observed computational basis measurement outcome (of the rotated state $U\rho U^\dagger$). Moreover, the norm defined in Eq. (S7) *is closely related to the Hilbert-Schmidt norm:*

$$\mathrm{tr}\left(O_0^2\right) \leq \|O_0\|^2_{\mathrm{shadow}} \leq 3\mathrm{tr}\left(O_0^2\right) \quad \textit{for any traceless } O_0 \in \mathbb{H}_{2^n}. \tag{S39}$$

Note that passing from O to its traceless part $O_0 = O - \frac{\mathrm{tr}(O)}{2^n}\mathbb{I}$ is a contraction in Hilbert-Schmidt norm:

$$\mathrm{tr}\left(O_0^2\right) = \mathrm{tr}(O^2) - \frac{\mathrm{tr}(O)^2}{2^n} \leq \mathrm{tr}(O^2). \tag{S40}$$

Hence, we can safely replace the upper bound in Eq. (S39) by $3\mathrm{tr}(O^2)$ — the Hilbert Schmidt norm (squared) of the original observable.

Proof. Eq. (S35) readily provides a closed-form expression for the measurement channel defined in Eq. (S2):

$$\mathcal{M}(\rho) = \sum_{b\in\{0,1\}^n} \mathbb{E}_{U\sim\mathrm{Cl}(2^n)} \langle b|U\rho U^\dagger|b\rangle U^\dagger|b\rangle\langle b|U = \sum_{b\in\{0,1\}^n} \frac{1}{2^n}\mathcal{D}_{1/(2^n+1)}(\rho) = \mathcal{D}_{1/(2^n+1)}(\rho). \tag{S41}$$

This depolarizing channel can be readily inverted, see Eq. (S37). In particular,

$$\hat{\rho} = \mathcal{M}^{-1}\left(U^\dagger|\hat{b}\rangle\langle\hat{b}|U\right) = (2^n+1)U^\dagger|\hat{b}\rangle\langle\hat{b}|U - \mathbb{I} \quad \text{and} \quad \mathcal{M}^{-1}(O_0) = (2^n+1)O_0 \tag{S42}$$

for any traceless matrix $O_0 \in \mathbb{H}_{2^n}$. The latter reformulation considerably simplifies the expression for the norm $\|O_0\|^2_{\mathrm{shadow}}$ defined in Eq. (S7). A slight reformulation allows us to furthermore capitalize on Eq. (S36) to exactly compute this norm for traceless observables:

$$\begin{aligned}\|O_0\|^2_{\mathrm{shadow}} &= \max_{\sigma \text{ state}} \mathrm{tr}\big(\sigma \sum_{b\in\{0,1\}^n} \mathbb{E}_{U\sim\mathrm{Cl}(2^n)} U^\dagger|b\rangle\langle b|U\langle b|U(2^n+1)O_0U^\dagger|b\rangle^2\big) \\ &= \max_{\sigma \text{ state}} \mathrm{tr}\left(\sigma\, \frac{(2^n+1)^2\left(\mathrm{tr}(O_0^2)\mathbb{I}+2O_0^2\right)}{(2^n+2)(2^n+1)2^n}\right) = \frac{2^n+1}{2^n+2}\max_{\sigma \text{ state}}\left(\mathrm{tr}(\sigma)\mathrm{tr}(O_0^2) + 2\mathrm{tr}\left(\sigma O_0^2\right)\right).\end{aligned} \tag{S43}$$

To further simplify this expression, recall $\mathrm{tr}(\sigma) = 1$ and note that $\max_{\sigma \text{ state}} \mathrm{tr}(\sigma O_0^2) = \|O_0^2\|_\infty$, where $\|\cdot\|_\infty$ denotes the spectral norm. The bound Eq. (S39) then foloows from the elementary relation between the spectral and Hilbert-Schmidt norms: $\|O_0^2\|_\infty \leq \mathrm{tr}(O_0^2)$. □

C. Performance bound for classical shadows based on random Pauli measurements

Proposition S2. *Adopt a "random Pauli basis" measurement primitive, i.e. each rotation $\rho \mapsto U\rho U^\dagger$ is a tensor product $U_1 \otimes \cdots \otimes U_n$ of randomly selected single-qubit Clifford gates $U_1, \ldots, U_n \in \mathrm{Cl}(2)$. Then, the associated classical shadow is*

$$\hat{\rho} = \bigotimes_{j=1}^{n} \left(3U_j^\dagger |\hat{b}_j\rangle\langle\hat{b}_j| U_j - \mathbb{I} \right) \quad \text{where} \quad |\hat{b}\rangle = |\hat{b}_1\rangle \otimes \cdots \otimes |\hat{b}_n\rangle \quad \text{and } \hat{b}_1, \ldots, \hat{b}_n \in \{0,1\}. \tag{S44}$$

Moreover, the norm defined in Eq. (S7) *respects locality. Suppose that $O \in \mathbb{H}_2^{\otimes k}$ only acts nontrivially on k-qubits, e.g. $O = \tilde{O} \otimes \mathbb{I}^{\otimes(n-k)}$ with $\tilde{O} \in \mathbb{H}_2^{\otimes k}$. Then $\|O\|_{\mathrm{shadow}} = \|\tilde{O}\|_{\mathrm{shadow}}$, where $\|\tilde{O}\|_{\mathrm{shadow}}$ is the same norm, but for k-qubit systems.*

Proof. Unitary rotation and computational basis measurements factorize completely into tensor products. This insight allows us to decompose the measurement channel $\mathcal{M}$ defined in Eq. (S2) into a tensor product of single-qubit operations. For elementary tensor products $X_1 \otimes \cdots \otimes X_n \in \mathbb{H}_2^{\otimes n}$ we can apply Eq. (S35) separately for each single-qubit action and infer

$$\begin{aligned} \mathcal{M}(X_1 \otimes \cdots \otimes X_n) &= \bigotimes_{j=1}^{n} \Big(\sum_{b_j \in \{0,1\}} \mathbb{E}_{U_j \sim \mathrm{Cl}(2)} U_j^\dagger |b\rangle\langle b| U_j \langle b| U_j X_j U_j^\dagger |b\rangle \Big) \\ &= \bigotimes_{j=1}^{n} \Big(\sum_{b_j \in \{0,1\}} \frac{1}{2} \mathcal{D}_{1/(2+1)}(\rho_j) \Big) = \mathcal{D}_{1/3}^{\otimes n}(X_1 \otimes \cdots \otimes X_n). \end{aligned} \tag{S45}$$

Linear extension to all of $\mathbb{H}_2^{\otimes n}$ yields the following formula for $\mathcal{M}$ and its inverse:

$$\mathcal{M}(X) = \left(\mathcal{D}_{1/3}\right)^{\otimes n}(X) \quad \text{and} \quad \mathcal{M}^{-1}(X) = \left(\mathcal{D}_{1/3}^{-1}\right)^{\otimes n}(X) \quad \text{for all } X \in \mathbb{H}_2^{\otimes n}, \tag{S46}$$

where $\mathcal{D}_{1/3}^{-1}(Y) = 3Y - \mathrm{tr}(Y)\mathbb{I}$ according to Eq. (S37). This formula readily yields a closed-form expression for the classical shadow. Use $U^\dagger|\hat{b}\rangle\langle\hat{b}|U = \bigotimes_{j=1}^{n} U_j|\hat{b}_j\rangle\langle\hat{b}_j|U_j$ to conclude

$$\hat{\rho} = \mathcal{M}^{-1}\left(U^\dagger|\hat{b}\rangle\langle\hat{b}|U\right) = \bigotimes_{j=1}^{n} \mathcal{D}_{1/3}^{-1}\left(U_j^\dagger|\hat{b}_j\rangle\langle\hat{b}_j|U_j\right) = \bigotimes_{j=1}^{n} \left(3U_j^\dagger|\hat{b}_j\rangle\langle\hat{b}_j|U - \mathbb{I}\right). \tag{S47}$$

For the second claim, we exploit a key feature of depolarizing channels and their inverses. The identity matrix is a fix-point, i.e. $\mathcal{D}_{1/3}^{-1}(\mathbb{I}) = \mathbb{I} = \mathcal{D}_{1/3}(\mathbb{I})$. For k-local observables, e.g. $O = \tilde{O} \otimes \mathbb{I}^{\otimes(n-k)}$, this feature ensures

$$\mathcal{M}^{-1}\left(\tilde{O} \otimes \mathbb{I}^{\otimes(n-k)}\right) = \left(\left(\mathcal{D}_{1/3}^{-1}\right)^{\otimes k}(\tilde{O})\right) \otimes \mathbb{I}^{\otimes(n-k)} = \tilde{\mathcal{M}}^{-1}(\tilde{O}) \otimes \mathbb{I}^{\otimes(n-k)}, \tag{S48}$$

where $\tilde{\mathcal{M}}^{-1}(X) = (\mathcal{D}_{1/3}^{-1})^{\otimes k}(X)$ denotes the inverse channel of a k-qubit local Clifford measurement procedure. This observation allows us to compress the norm (S7) to the "active" subset of k qubits. Exploit the tensor product structure $U = U_1 \otimes \cdots \otimes U_n$ with $U_i \sim \mathrm{Cl}(2)$ to conclude

$$\begin{aligned} \left\| \tilde{O} \otimes \mathbb{I}^{\otimes(n-k)} \right\|_{\mathrm{shadow}}^2 &= \max_{\sigma:\,\mathrm{state}} \mathbb{E}_{U \sim \mathrm{Cl}(2)^{\otimes n}} \sum_{b \in \{0,1\}^n} \langle b|U\sigma U^\dagger|b\rangle \langle b|U\mathcal{M}^{-1}(O \otimes \mathbb{I}^{\otimes(n-k)})U^\dagger|b\rangle^2 \\ &= \max_{\sigma:\,\mathrm{state}} \mathbb{E}_{U \sim \mathrm{Cl}(2)^{\otimes k}} \sum_{b \in \{0,1\}^k} \langle b|U\,\mathrm{tr}_{k+1,\ldots,n}(\sigma)U^\dagger|b\rangle \langle b|U\tilde{\mathcal{M}}^{-1}(\tilde{O})U^\dagger|b\rangle^2, \end{aligned} \tag{S49}$$

where $\mathrm{tr}_{k+1,\ldots,n}(\sigma)$ denotes the partial trace over all "inactive" subsystems. Partial traces preserve the space of all quantum states. So maximizing over all partial traces $\mathrm{tr}_{k+1,\ldots,n}(\sigma)$ is equivalent to maximizing over all k-qubit states and we exactly recover the norm $\|\tilde{O}\|_{\mathrm{shadow}}^2$ on k qubits. Finally, it is easy to check that the actual location of the active k-qubit support of O does not affect the argument. □

Recall that the (squared) norm $\|\cdot\|_{\mathrm{shadow}}^2$ is the most important figure of merit for feature prediction with classical shadows. According to Theorem S1, $\max_{1 \le i \le M} \|O_i\|_{\mathrm{shadow}}^2$ determines the number of samples required to accurately predict a collection of linear functions $\mathrm{tr}(O_1\rho), \ldots, \mathrm{tr}(O_M\rho)$. Viewed from this angle, Proposition S2 has profound consequences for predicting (collections of) local observables under the local Clifford

measurement primitive. For each local observable O_i, the norm $\|O_i\|^2_{\text{shadow}}$ collapses to its active support, regardless of its precise location. The size of these supports is governed by the locality alone, not the total number of qubits!

It is instructive to illustrate this point with a simple special case first.

Lemma S3. *Let O be a single k-local Pauli observable, e.g. $O = P_{p_1} \otimes \cdots \otimes P_{p_k} \otimes \mathbb{I}^{\otimes(n-k)}$, where $p_j \in \{X, Y, Z\}$. Then, $\|O\|^2_{\text{shadow}} = 3^k$, for any choice of the k qubits where nontrivial Pauli matrices act. This scaling can be generalized to arbitrary elementary tensor products supported on k qubits,* e.g. *$O = O_1 \otimes \cdots \otimes O_k \otimes \mathbb{I}^{\otimes(n-k)}$.*

Proof. Pauli matrices are traceless and obey, $P_{p_j}^2 = \mathbb{I}$ and $\mathcal{D}_{1/3}^{-1}(P_{p_j}) = 3P_{p_j}$ for each $p_j \in \{X, Y, Z\}$. Proposition S2 and the tensor product structure of the problem then ensure

$$\begin{aligned}
\|O\|^2_{\text{shadow}} =& \|P_{p_1} \otimes \cdots \otimes P_{p_k}\|^2_{\text{shadow}} \\
&= \max_{\sigma\text{: state}} \mathbb{E}_{U\sim \text{Cl}(2)^{\otimes k}} \sum_{b\in\{0,1\}^n} \langle b|U^\dagger \sigma U|b\rangle \langle b|U(\mathcal{D}_{1/3}^{-1})^{\otimes k}(P_1 \otimes \cdots \otimes P_k)U^\dagger|b\rangle^2 \\
&= \max_{\sigma\text{: state}} \text{tr}\Big(\sigma \bigotimes_{j=1}^{k} \big(\sum_{b_j\in\{0,1\}} \mathbb{E}_{U_j\sim \text{Cl}(2)} U^\dagger|b_j\rangle\langle b_j|U\langle b_j|U 3P_j U^\dagger U|b_j\rangle^2\big)\Big) \\
&= \max_{\sigma\text{: state}} \text{tr}\Big(\sigma \bigotimes_{j=1}^{k} \big(9 \sum_{b\in\{0,1\}} \frac{\text{tr}\left(P_j^2\right)\mathbb{I} + 2P_j^2}{(2+2)(2+1)2}\big)\Big) = \max_{\sigma\text{: state}} \text{tr}\Big(\sigma \bigotimes_{j=1}^{k} 3\mathbb{I}\Big) = 3^k,
\end{aligned} \tag{S50}$$

where we have used Eq. (S36) to explicitly evaluate the single qubit Clifford averages.

We leave the extension to more general tensor product observables as an exercise for the dedicated reader. □

The norm expression in Lemma S3 scales exponentially in the locality k, but is independent of the total number of qubits n. The compression property (Proposition S2) suggests that this desirable feature should extend to general k-local observables. And, indeed, it is relatively straightforward to obtain crude upper bounds that scale with 3^{2k}. The additional factor of two, however, effectively doubles the locality parameter and can render conservative feature prediction with classical shadows prohibitively expensive in concrete applications.

The main result of this section considerably improves upon these crude bounds and *almost* reproduces the (tight) scaling associated with k-local Pauli observables.

Proposition S3. *Let O be a k-local observable, e.g. $O = \tilde{O} \otimes \mathbb{I}^{\otimes(n-k)}$ with $\tilde{O} \in \mathbb{H}_2^{\otimes k}$ Then,*

$$\|O\|^2_{\text{shadow}} \leq 4^k \|O\|^2_\infty, \quad \textit{where } \|\cdot\|_\infty \textit{ denotes the spectral/operator norm.} \tag{S51}$$

The same bound holds for the shadow norm of the traceless part of O: $\|O - \frac{\text{tr}(O)}{2^n}\mathbb{I}\|^2_{\text{shadow}} \leq 4^k \|O\|^2_\infty$.

The proof is considerably more technical than the proof of Lemma S3 and relies on the following auxiliary result.

Lemma S4. *Fix two k-qubit Pauli observables $P_{\mathbf{p}} = P_{p_1} \otimes \cdots \otimes P_{p_k}$, $P_{\mathbf{q}} = P_{q_1} \otimes \cdots \otimes P_{q_k}$ with $\mathbf{p}, \mathbf{q} \in \{\mathbb{I}, X, Y, Z\}^k$. Then, the following formula is true for any state σ:*

$$\mathbb{E}_{U\sim \text{Cl}(2)^{\otimes k}} \sum_{b\in\{0,1\}^k} \langle b|U\sigma U^\dagger|b\rangle \langle b|U(\mathcal{D}_{1/3}^{-1})^{\otimes k}(P_{\mathbf{p}})U^\dagger|b\rangle \langle b|U(\mathcal{D}_{1/3}^{-1})^{\otimes k}(P_{\mathbf{q}})U^\dagger|b\rangle = f(\mathbf{p}, \mathbf{q}) \text{tr}\left(\sigma P_{\mathbf{p}} P_{\mathbf{q}}\right), \tag{S52}$$

where $f(\mathbf{p}, \mathbf{q}) = 0$ whenever there exists an index i such that $p_i \neq q_i$ and $p_i, q_i \neq \mathbb{I}$. Otherwise, $f(\mathbf{p}, \mathbf{q}) = 3^s$, where s is the number of non-identity Pauli indices that match ($s = |\{i : p_i = q_i, p_i \neq \mathbb{I}\}|$).

This combinatorial formula follows from a straightforward, but somewhat cumbersome, case-by-case analysis based on the (single-qubit) relations (S35) and (S36). We include a proof at the end of this subsection.

Proof of Proposition S3. Proposition S2 allows us to restrict our attention to the relevant k-qubit region on which $\tilde{O} \in \mathbb{H}_2^{\otimes k}$ acts nontrivially. Next, expand $\tilde{O}$ in the (tensor product) Pauli basis, i.e. $\tilde{O} = \sum_{\mathbf{p}} \alpha_{\mathbf{p}} P_{\mathbf{p}}$ with $\mathbf{p} \in \{\mathbb{I}, X, Y, Z\}^k$. Fix an arbitrary k-qubit state σ and use Lemma S4 to conclude

$$\begin{aligned}
\|\tilde{O}\|^2_{\text{shadow}} &= \max_{\sigma \text{ state}} \mathbb{E}_{U\sim \text{Cl}(2)^{\otimes k}} \sum_{b\in\{0,1\}^k} \langle b|U\sigma U^\dagger|b\rangle \langle b|U(\mathcal{D}_{1/3}^{-1})^{\otimes k}(\tilde{O})U^\dagger|b\rangle^2 \\
&= \max_{\sigma \text{ state}} \sum_{\mathbf{p},\mathbf{q}} \alpha_{\mathbf{p}}\alpha_{\mathbf{q}} \mathbb{E}_{U\sim \text{Cl}(2)^{\otimes k}} \sum_{b\in\{0,1\}^k} \langle b|U\sigma U^\dagger|b\rangle \langle b|U(\mathcal{D}_{1/3}^{-1})^{\otimes k}(P_{\mathbf{p}})U^\dagger|b\rangle \langle b|U(\mathcal{D}_{1/3}^{-1})^{\otimes k}(P_{\mathbf{q}})U^\dagger|b\rangle
\end{aligned}$$

$$= \max_{\sigma \text{ state}} \sum_{\mathbf{p},\mathbf{q}} \alpha_{\mathbf{p}} \alpha_{\mathbf{q}} f(\mathbf{p},\mathbf{q}) \mathrm{tr}\left(\sigma P_{\mathbf{p}} P_{\mathbf{q}}\right) = \max_{\sigma \text{ state}} \mathrm{tr}\left(\sigma \sum_{\mathbf{p},\mathbf{q}} \alpha_{\mathbf{p}} \alpha_{\mathbf{q}} f(\mathbf{p},\mathbf{q}) \mathrm{tr}\left(\sigma P_{\mathbf{p}} P_{\mathbf{q}}\right)\right)$$
$$= \left\| \sum_{\mathbf{p},\mathbf{q}} \alpha_{\mathbf{p}} \alpha_{\mathbf{q}} f(\mathbf{p},\mathbf{q}) \mathrm{tr} P_{\mathbf{p}} P_{\mathbf{q}} \right\|_{\infty}, \tag{S53}$$

where $f(\mathbf{p},\mathbf{q})$ is the combinatorial function defined in Lemma S4. The last equality follows from the dual characterization of the spectral norm: $\|A\|_\infty = \max_{\sigma:\text{ state}} \mathrm{tr}(\sigma A)$ for any positive semidefinite matrix A.

We can further simplify this expression by introducing a partial order on Pauli strings $\mathbf{q}, \mathbf{s} \in \{\mathbb{I}, X, Y, Z\}^n$. We write $\mathbf{q} \rhd \mathbf{s}$ if it is possible to obtain $\mathbf{q}$ from $\mathbf{s}$ by replacing some local non-identity Paulis with $\mathbb{I}$. Moreover, let $|\mathbf{q}| = |\{i :\ q_i \neq \mathbb{I}\}|$ denote the number of non-identity Pauli's in the string $\mathbf{q}$. Then,

$$\left\| \sum_{\mathbf{p},\mathbf{q}} \alpha_{\mathbf{p}} \alpha_{\mathbf{q}} f(\mathbf{p},\mathbf{q}) \mathrm{tr} P_{\mathbf{p}} P_{\mathbf{q}} \right\|_{\infty} = \left\| \tfrac{1}{3^k} \sum_{\mathbf{s} \in \{X,Y,Z\}^k} \left(\sum_{\mathbf{q} \rhd \mathbf{s}} 3^{|\mathbf{q}|} \alpha_{\mathbf{q}} P_{\mathbf{q}} \right)^2 \right\|_{\infty} \leq \frac{1}{3^k} \sum_{\mathbf{s} \in \{X,Y,Z\}^k} \left(\sum_{\mathbf{q} \rhd \mathbf{s}} 3^{|\mathbf{q}|} \alpha_{\mathbf{q}} P_{\mathbf{q}} \right)^2, \tag{S54}$$

where we have used $\|P_{\mathbf{q}}\|_\infty = 1$ for all Pauli strings. Next, note that for fixed $\mathbf{s} \in \{X, Y, Z\}^k$,

$$\sum_{\mathbf{q} \rhd \mathbf{s}} 3^{|\mathbf{q}|} = 3^k + k 3^{k-1} + \binom{k}{2} 3^{k-2} + \cdots + 1 = 4^k. \tag{S55}$$

Together with Cauchy-Schwarz, this numerical insight implies

$$\tfrac{1}{3^k} \sum_{\mathbf{s} \in \{X,Y,Z\}^k} \left(\sum_{\mathbf{q} \rhd \mathbf{s}} 3^{|\mathbf{q}|} |\alpha_{\mathbf{q}}| \right)^2 \leq \tfrac{1}{3^k} \sum_{\mathbf{s} \in \{X,Y,Z\}^k} \left(\sum_{\mathbf{q} \rhd \mathbf{s}} 3^{|\mathbf{q}|} \right) \left(\sum_{\mathbf{q} \rhd \mathbf{s}} 3^{|\mathbf{q}|} |\alpha_{\mathbf{p}}^2 \right) = 4^k \sum_{\mathbf{s} \in \{X,Y,Z\}} \sum_{\mathbf{q} \rhd \mathbf{s}} 3^{|\mathbf{q}|-k} |\alpha_{\mathbf{q}}|^2. \tag{S56}$$

Finally, observe that every $\mathbf{q} \in \{\mathbb{I}, X, Y, Z\}^k$ is dominated by exactly $3^{k-|\mathbf{q}|}$ different strings $\mathbf{s} \in \{X, Y, Z\}^k$. This ensures

$$4^k \sum_{\mathbf{s} \in \{X,Y,Z\}} 3^{|\mathbf{q}|-k} |\alpha_{\mathbf{q}}|^2 = 4^k \sum_{\mathbf{q} \in \{\mathbb{I},X,Y,Z\}} |\alpha_{\mathbf{q}}|^2 = 4^k 2^{-k} \|\tilde{O}\|_2^2, \tag{S57}$$

because Pauli matrices are proportional to an orthonormal basis of $\mathbb{H}_2^{\otimes k}$: $\sum_{\mathbf{q}} |\alpha_{\mathbf{q}}|^2 = \sum_{\mathbf{q}} \left|2^{-k} \mathrm{tr}\left(\sigma_{\mathbf{q}} \tilde{O}\right)\right|^2 = 2^{-k} \|\tilde{O}\|_2^2$. The general claim then follows from the fundamental relation among Schatten norms: $\|\tilde{O}\|_2^2 \leq 2^k \|\tilde{O}\|_\infty^2 = 2^k \|O\|_\infty^2$.

The bound on traceless parts O_0 of observables is nearly analogous, because the transition from O to O_0 respects locality. E.g. $O = \tilde{O} \otimes \mathbb{I}^{\otimes(n-k)}$ obeys $O_0 = \tilde{O}_0 \otimes \mathbb{I}^{\otimes(n-k)}$. To get the same bound, we use that this transition is a contraction in Hilbert-Schmidt norm:

$$\|O_0\|_{\text{shadow}}^2 = \|\tilde{O}_0\|_{\text{shadow}}^2 \leq 4^k 2^{-k} \|\tilde{O}_0\|_2^2 \leq 4^k 2^{-k} \|\tilde{O}\|_2^2 \leq 4^k \|\tilde{O}\|_\infty^2 = \|O\|_\infty^2.$$

□

Proof of Lemma S4. Since Pauli observables decompose nicely into tensor products, this claim readily follows from extending a single-qubit argument. Note that $\mathcal{D}_{1/3}^{-1}(P_p) = 3P_p$ for $p \neq \mathbb{I}$ and $\mathcal{D}_{1/3}^{-1}(\mathbb{I}) = \mathbb{I}$. It is straightforward to evaluate the single-qubit expression for the trivial case $P_p = P_q = \mathbb{I}$. Fix a state σ and compute

$$\mathbb{E}_{U \sim \mathrm{Cl}(2)} \sum_{b \in \{0,1\}} \langle b|U\sigma U^\dagger|b\rangle \langle b|U \mathcal{D}_{1/3}^{-1}(\mathbb{I}) U^\dagger|b\rangle^2 = \mathbb{E}_{U \sim \mathrm{Cl}(2)} \sum_{b \in \{0,1\}} \langle b|U\sigma U^\dagger|b\rangle = \mathbb{E}_{U \sim \mathrm{Cl}(2)} \mathrm{tr}(\sigma) = \mathrm{tr}\left(\sigma \mathbb{I}^2\right). \tag{S58}$$

Next, suppose $P_q = \mathbb{I}$, but $P_p \neq \mathbb{I}$. This single-qubit case is covered by Eq. (S35):

$$\mathbb{E}_{U \sim \mathrm{Cl}(2)} \sum_{b \in \{0,1\}} \langle b|U\sigma U^\dagger|b\rangle \langle b|U \mathcal{D}_{1/3}^{-1}(P_p) U^\dagger|b\rangle \langle b|U \mathcal{D}_{1/3}^{-1} \mathbb{I} U^\dagger|b\rangle$$
$$= \mathrm{tr}\Big(\sigma \sum_{b \in \{0,1\}} U^\dagger|b\rangle\langle b|U \langle b|U 3 P_p U^\dagger|b\rangle\Big) = 3\mathrm{tr}\Big(\sigma \sum_{b \in \{0,1\}} \frac{1}{2} \mathcal{D}_{1/3}(P_p)\Big) = \mathrm{tr}\left(\sigma P_p \mathbb{I}\right), \tag{S59}$$

because $\mathcal{D}_{1/3}(P_p) = \frac{1}{3}P_p$. The case $P_p = \mathbb{I}$ and $P_q \neq \mathbb{I}$ leads to analogous results. Finally, suppose that both $P_p, P_q \neq \mathbb{I}$. By assumption $\mathcal{D}_{1/3}^{-1}(P_p)$, $\mathcal{D}_{1/3}^{-1}(P_q)$ and both matrices are traceless. Hence, we can resort to Eq. (S36) to conclude

$$\begin{aligned}&\mathbb{E}_{U\sim \mathrm{Cl}(2)^{\otimes n}} \sum_{b\in\{0,1\}^k} \langle b|U\sigma U^\dagger|b\rangle\langle b|U(\mathcal{D}_{1/3}^{-1})^{\otimes k}(P_p)U^\dagger|b\rangle\langle b|U(\mathcal{D}_{1/3}^{-1})^{\otimes k}(P_q)U^\dagger|b\rangle \\ =&\mathrm{tr}\Big(\sigma \sum_{b\in\{0,1\}} U^\dagger|b\rangle\langle b|U\langle b|U3P_pU^\dagger|b\rangle\langle b|U3P_qU^\dagger|b\rangle\Big) = 9\mathrm{tr}\Big(\sigma \sum_{b\in\{0,1\}} \frac{\mathrm{tr}(P_pP_q)\mathbb{I} + P_pP_q + P_qP_p}{(2+2)(2+1)2}\Big) \end{aligned} \tag{S60}$$

for any state σ. Pauli matrices are orthogonal ($\mathrm{tr}(P_pP_q) = 2\delta_{p,q}$) and anticommute ($P_pP_q + P_qP_p = 2\delta_{p,q}$). This implies that the above expression vanishes whenever $p \neq q$. If $p = q$ it evaluates to $3\mathrm{tr}(\sigma P_pP_q)$ and we can conclude that the single qubit average always equals

$$f(p,q)\mathrm{tr}\left(\sigma P_pP_q\right) \quad \text{where} \quad f(p,q) = \begin{cases} 1 & \text{if } p = \mathbb{I} \text{ or } q = \mathbb{I}, \\ 3 & \text{if } p = q \neq \mathbb{I}, \\ 0 & \text{else.} \end{cases} \tag{S61}$$

The statement then follows from extending this formula to tensor products of k Pauli matrices. □

6. ADDITIONAL COMPUTATIONS AND PROOFS FOR PREDICTING NONLINEAR FUNCTIONS

We focus on the particularly relevant task of predicting quadratic functions with classical shadows, using

$$\hat{o}(N,1) = \frac{1}{N(N-1)}\sum_{j\neq l}\mathrm{tr}(O\hat{\rho}_i\otimes\hat{\rho}_j) \quad \text{to predict} \quad \mathrm{tr}\left(O\rho\otimes\rho\right) = \mathbb{E}\,\hat{o}(N,1). \tag{S62}$$

A. General variance bound

Lemma S5 (Variance)**.** *The variance associated with the estimator $\hat{O}(N,1)$ obeys*

$$\begin{aligned}\mathrm{Var}[\hat{o}(N,1)] &= \binom{N}{2}^{-1}\Big(2(N-2)\,\mathrm{Var}[\mathrm{tr}(O_s\hat{\rho}_1\otimes\rho)] + \mathrm{Var}[\mathrm{tr}(O_s\hat{\rho}_1\otimes\hat{\rho}_2)]\Big) \\ &\leq \frac{4}{N^2}\,\mathrm{Var}[\mathrm{tr}(O\hat{\rho}_1\otimes\hat{\rho}_2)] + \frac{2}{N}\,\mathrm{Var}[\mathrm{tr}(O\hat{\rho}_1\otimes\rho)] + \frac{2}{N}\,\mathrm{Var}[\mathrm{tr}(O\rho\otimes\hat{\rho}_1)], \end{aligned} \tag{S63}$$

where $O_s = (O + SOS)/2$ is the symmetrized version of O and S denotes the swap operator ($S|\psi\rangle\otimes|\phi\rangle = |\phi\rangle\otimes|\psi\rangle$).

Proof. First, note that $\hat{o}(N,1)$ and the target $\mathrm{tr}(O\rho\otimes\rho)$ are invariant under symmetrization. This ensures

$$\hat{o}(N,1) = \binom{N}{2}\sum_{i<j}\mathrm{tr}\left(O_s\hat{\otimes}\hat{\rho}_j\right) \quad \text{and moreover} \quad \mathrm{tr}\left(O\rho\otimes\rho\right) = \mathrm{tr}\left(O_s\rho\otimes\rho\right). \tag{S64}$$

Thus, we may without loss replace the original observable O by its symmetrized version O_s. Next, we expand the definition of the variance:

$$\begin{aligned}\mathrm{Var}[\hat{o}(N,1)] =&\mathbb{E}\left[\left(\hat{o}(N,1) - \mathrm{tr}(O_s\rho\otimes\rho)\right)^2\right] \\ =&\binom{N}{2}^{-2}\sum_{i<j}\sum_{k<l}\Big(\mathbb{E}\left[\mathrm{tr}(O_s\hat{\rho}_i\otimes\hat{\rho}_j)\,\mathrm{tr}(O_s\hat{\rho}_k\otimes\hat{\rho}_l)\right] - \mathrm{tr}(O_s\rho\otimes\rho)^2\Big) \\ =&\binom{N}{2}^{-2}\sum_{i<j}\mathbb{E}\left[\mathrm{tr}(O_s\hat{\rho}_i\otimes\hat{\rho}_j)^2\right] - \mathrm{tr}(O_s\rho\otimes\rho)^2\Big) \\ &+2\binom{N}{2}^{-2}\sum_{i<j}\sum_{l\neq i,j}\Big(\mathbb{E}\left[\mathrm{tr}(O_s\hat{\rho}_i\otimes\hat{\rho}_j)\,\mathrm{tr}(O_s\hat{\rho}_i\otimes\hat{\rho}_l)\right] - \mathrm{tr}(O_s\rho\otimes\rho)^2\Big)\end{aligned}$$

$$= \binom{N}{2}^{-1} \mathrm{Var}[\mathrm{tr}(O_s \hat{\rho}_1 \otimes \hat{\rho}_2)] + \binom{N}{2}^{-1} 2(N-2)\, \mathrm{Var}[\mathrm{tr}(O_s \hat{\rho}_1 \otimes \rho)]. \tag{S65}$$

We can use the inequality $\mathrm{Var}[(A+B)/2] \leq (\mathrm{Var}[A] + \mathrm{Var}[B])/2$ (for any pair of random variables A, B) to obtain a simplified upper bound:

$$\begin{aligned}
\mathrm{Var}[\hat{o}(N,1)] &= \binom{N}{2}^{-1} \mathrm{Var}[\mathrm{tr}(O_s \hat{\rho}_1 \otimes \hat{\rho}_2)] + \binom{N}{2}^{-1} 2(N-2)\, \mathrm{Var}[\mathrm{tr}(O_s \hat{\rho}_1 \otimes \rho)] \\
&\leq \frac{4}{N^2} \mathrm{Var}[\mathrm{tr}(O_s \hat{\rho}_1 \otimes \hat{\rho}_2)] + \frac{4}{N} \mathrm{Var}[\mathrm{tr}(O_s \hat{\rho}_1 \otimes \rho)] \\
&\leq \frac{4}{N^2} \mathrm{Var}[\mathrm{tr}(O \hat{\rho}_1 \otimes \hat{\rho}_2)] + \frac{2}{N} \mathrm{Var}[\mathrm{tr}(O \hat{\rho}_1 \otimes \rho)] + \frac{2}{N} \mathrm{Var}[\mathrm{tr}(O \rho \otimes \hat{\rho}_1)].
\end{aligned} \tag{S66}$$

□

B. Concrete variance bounds for random Pauli measurements

Proposition S4. *Suppose that O describes a quadratic function $\mathrm{tr}(O\rho \otimes \rho)$ that acts on at most k-qubits in the first system and at most k-qubits in the second system and obeys $\|O\|_\infty \geq 1$. Then,*

$$\max\left(\mathrm{Var}[\mathrm{tr}(O\rho \otimes \hat{\rho}_1)], \mathrm{Var}[\mathrm{tr}(O\hat{\rho}_1 \otimes \rho)], \sqrt{\mathrm{Var}[\mathrm{tr}(O\hat{\rho}_1 \otimes \hat{\rho}_2)]}\right) \leq 4^k \|O\|_\infty^2 . \tag{S67}$$

Proof. Because of the single-qubit tensor product structure in the random Pauli measurement and the inverted quantum channel $\mathcal{M}_P^{-1}$, the tensor product of two snapshots $\hat{\rho}_1 \otimes \hat{\rho}_2$ of the unknown quantum state ρ may be viewed as a single snapshot of the tensor product state $\rho \otimes \rho$:

$$\begin{aligned}
\hat{\rho}_1 \otimes \hat{\rho}_2 &= \bigotimes_{i=1}^{n} \left(\mathcal{M}_1^{-1}(U_1^{(i)}|b_1^{(i)}\rangle\langle b_1^{(i)}|(U_1^{(i)})^\dagger)\right) \bigotimes_{i=1}^{n} \left(\mathcal{M}_1^{-1}(U_2^{(i)}|b_2^{(i)}\rangle\langle b_2^{(i)}|(U_2^{(i)})^\dagger)\right) \\
&= \bigotimes_{i=1}^{2n} \mathcal{M}_1^{-1}(U^{(i)}|b^{(i)}\rangle\langle b^{(i)}|(U^{(i)})^\dagger) =: \hat{\rho}.
\end{aligned} \tag{S68}$$

Hence $\mathrm{tr}(O\hat{\rho}_1 \otimes \hat{\rho}_2) = \mathrm{tr}(O\hat{\rho})$ and, by assumption, O is an observable that acts on $k+k = 2k$ qubits only. The claim then follows from invoking the variance bounds for linear feature prediction presented in Proposition S3. □

C. Concrete variance bounds for random Clifford measurements

In contrast to the Pauli basis setup, variances for quadratic feature prediction with Clifford basis measurements cannot be directly reduced to its linear counterpart. Nonetheless, a more involved direct analysis does produces bounds that do closely resemble the linear base case.

Proposition S5. *Suppose that O describes a quadratic function $\mathrm{tr}(O\rho \otimes \rho)$ and obeys $\mathrm{tr}(O^2) \geq 1$. Then, the variance associated with classical shadow estimation (random Clifford measurements) obeys*

$$\max\left(\mathrm{Var}[\mathrm{tr}(O\rho \otimes \hat{\rho}_1)], \mathrm{Var}[\mathrm{tr}(O\hat{\rho}_1 \otimes \rho)], \sqrt{\mathrm{Var}[\mathrm{tr}(O\hat{\rho}_1 \otimes \hat{\rho}_2)]}\right) \leq \sqrt{9 + 6/2^n}\, \mathrm{tr}(O^2). \tag{S69}$$

The pre-factor $\sqrt{9+6/2^n}$ converges to the constant 3 at an exponential rate in system size.

This claim is based on the following technical Lemma and insights regarding linear feature prediction.

Lemma S6. *Suppose that O describes a quadratic function $\mathrm{tr}(O\rho \otimes \rho)$. Then,*

$$\mathrm{Var}[\mathrm{tr}(O\hat{\rho}_1 \otimes \hat{\rho}_2)] \leq 9\,\mathrm{tr}(O^2) + \frac{6}{2^n}\|O\|_\infty^2. \tag{S70}$$

Proof of Proposition S5. The variance of $\mathrm{tr}(O\rho \otimes \hat{\rho}_1)$ is equivalent to the variance of $\mathrm{tr}(\tilde{O}_\rho \hat{\rho})$, where $\tilde{O}_\rho = \mathrm{tr}_1(\rho \otimes \mathbb{I}O)$ describes a linear function. According to Proposition S1, this variance term obeys

$$\mathrm{Var}\left[\mathrm{tr}\left(O\rho \otimes \hat{\rho}\right)\right] = \mathrm{Var}\left[\mathrm{tr}\left(\tilde{O}_\rho \hat{\rho}_1\right)\right] \leq 3\mathrm{tr}\left(\tilde{O}_\rho^2\right) = \mathrm{tr}\left(\mathrm{tr}_1\left(\rho \otimes \mathbb{I}O\right)^2\right) \leq 3\mathrm{tr}(O^2), \tag{S71}$$

because $\mathrm{tr}(\rho) = 1$ and $\mathrm{tr}(\rho^2) \leq 1$. A similar argument takes care of the second variance contribution $\mathrm{Var}\left[\mathrm{tr}\left(O\hat{\rho}_1 \otimes \rho\right)\right]$. Lemma S6 supplies a bound for the square of the final contribution. By assumption $\sqrt{\mathrm{tr}(O^2)} \leq \mathrm{tr}(O^2)$ and the claim follows. □

The remainder of this section is devoted to proving Lemma S6. Unfortunately, there does not seem to be a direct way to relate this task to variance bounds for linear feature prediction. Instead, we base our analysis on the 3-design property (S36) of Clifford circuits and a reformulation of this feature in terms of permutation operators. This strategy is inspired by the approach developed in [9], but conceptually and technically somewhat simpler. We believe that similar arguments extend to variances associated with higher order polynomials, but do refrain from a detailed analysis. Instead, we carefully outline the main ideas and leave a rigorous extension to future work.

Problem statement and reformulation: We will ignore symmetrization (which can only make the variance smaller) and focus on bounding the variance of $\mathrm{tr}\left(O\hat{\rho}_1 \otimes \hat{\rho}_2\right)$, where each $\hat{\rho}_i$ is an independent classical shadow. To simplify notation, we set $d = 2^n$ and define the following traceless variants of O:

$$\begin{aligned} O_0^{(1)} =& \mathrm{tr}_2(O) - \frac{\mathrm{tr}\left(O\right)}{d}\mathbb{I}, \quad \text{and} \quad O_0^{(2)} = \mathrm{tr}_1(O) - \frac{\mathrm{tr}(O)}{d}\mathbb{I}, \quad \text{as well as} \\ O_0^{(1,2)} =& O - \mathrm{tr}_2(O) \otimes \frac{\mathbb{I}}{d} - \frac{\mathbb{I}}{d} \otimes \mathrm{tr}_1(O) + \mathrm{tr}(O)\frac{\mathbb{I}}{d} \otimes \frac{\mathbb{I}}{d}. \end{aligned} \tag{S72}$$

Here, $\mathrm{tr}_a(O)$ with $a = 1, 2$ denotes the partial trace over the first and second system, respectively. All three operators are traceless (recall $\mathrm{tr}\left(\mathrm{tr}_a(O)\right) = \mathrm{tr}(O)$) and the final (bipartite) operator has the additional property that both partial traces vanish identically: $\mathrm{tr}_a\big(O_0^{(1,2)}\big) = 0$.

Proposition S1 asserts $\hat{\rho}_a = (d+1)U_a^\dagger|\hat{b}_a\rangle\langle\hat{b}_a|U_a - \mathbb{I}$, where each $U_a \in \mathrm{Cl}(d)$ is a random Clifford unitary and $\hat{b}_a \in \{0,1\}^n$ is the outcome of a computational basis measurement. These explicit formulas allow us to decompose the expression of interest in the following fashion:

$$\begin{aligned} \mathrm{tr}\left(O\hat{\rho}_1 \otimes \hat{\rho}_2\right) =& (d+1)^2 \mathrm{tr}\left(O_0^{(1,2)} U_1^\dagger|\hat{b}_1\rangle\langle\hat{b}_1|U_1 \otimes U_2^\dagger|\hat{b}_1\rangle\langle\hat{b}_2|U_2\right) + \frac{\mathrm{tr}(O)^2}{d^2} \\ &+ \frac{d+1}{d}\mathrm{tr}\left(O_0^{(1)} U_1^\dagger|\hat{b}_1\rangle\langle\hat{b}_1|U_1\right) + \frac{d+1}{d}\mathrm{tr}\left(O_0^{(2)} U_2^\dagger|\hat{b}_2\rangle\langle\hat{b}_2|U_2\right). \end{aligned} \tag{S73}$$

The variance corresponds to the expected square of this expression. The second term is constant and does not contribute. We analyze the remaining terms on a case-by case basis.

Linear terms: The third and fourth terms in Eq. (S73) are linear feature functions in one classical shadow only. Their (squared) contribution to the overall variance is characterized by Proposition S1:

$$\mathbb{E}\left[\left(\frac{d+1}{d}\mathrm{tr}\left(O_0^{(a)} U_a^\dagger|\hat{b}_a\rangle\langle\hat{b}_a|U_a\right)\right)^2\right] \leq \frac{3}{d^2}\left\|O_0^{(a)}\right\|_2^2 \quad \text{for } a = 1, 2. \tag{S74}$$

Both bounds can be related to the Hilbert-Schmidt norm (squared) of the original observable:

$$\frac{3}{d^2}\left\|O_0^{(a)}\right\|_2^2 \leq \frac{3}{d^2}\left\|\mathrm{tr}_a(O)\right\|_2^2 \leq 3\|O\|_2^2 = 3\mathrm{tr}\left(O^2\right). \tag{S75}$$

Leading-order term: We need to bound $\mathbb{E}\big[(d+1)^4 \mathrm{tr}\left(O_0^{(1,2)} U_1^\dagger|\hat{b}_1\rangle\langle\hat{b}_1|U_1 \otimes U_2^\dagger|\hat{b}_2\rangle\langle\hat{b}_2|U_2\right)^2\big]$, where $O_0^{(1,2)}$ has the special property that both partial traces vanish identically: $\mathrm{tr}_a\big(O_0^{(1,2)}\big) = 0$ for $a = 1, 2$. Moreover, the Hilbert-Schmidt norm (squared) of this operator factorizes nicely:

$$\left\|O_0^{(1,2)}\right\|_2^2 = \|O\|_2^2 - \frac{1}{d}\left\|O_0^{(1)}\right\|_2^2 - \left\|O_0^{(2)}\right\|_2^2 - \frac{\mathrm{tr}(O)^2}{d^2}. \tag{S76}$$

Not only is this expression bounded by the original Hilbert-Schmidt norm $\|O\|_2^2$. The norms of partial traces also feature explicitly with a minus sign. This will allow us to fully counter-balance the variance contributions (S75) from the linear terms.

Next, we use the 3-design property (S34) of Clifford circuits in dimension $d = 2^n$:

$$\mathbb{E}_{U_a \sim \mathrm{Cl}(d)}\left[\left(U_a^\dagger|b_a\rangle\langle b_a|U_a\right)^{\otimes 3}\right] = \binom{d+2}{3}^{-1} P_{\vee^3}, \tag{S77}$$

where $P_{\vee^3}$ is the projector onto the totally symmetric subspace of $\mathbb{C}^d \otimes \mathbb{C}^d \otimes \mathbb{C}^d$. This formula implies

$$\mathbb{E}\left[(d+1)^4 \mathrm{tr}\left(O_0^{(1,2)} U_1^\dagger |\hat{b}_1\rangle\langle\hat{b}_1| U_1 \otimes U_2^\dagger |\hat{b}_2\rangle\langle\hat{b}_2| U_2\right)^2\right] \leq \mathrm{tr}\left(O_0^{(1,2)} \otimes O_0^{(1,2)} \otimes \rho \otimes \rho \; P_{\vee^3}^{(\mathrm{odd})} \otimes P_{\vee^3}^{(\mathrm{even})}\right), \quad \text{(S78)}$$

where the superscripts "even" and "odd" indicate on which subset of tensor factors the projectors act.

Next, we exploit the fact that symmetric projectors can be decomposed into permutation operators: $(3!)P_{\vee^3} = \sum_{\pi\in S_3} W_\pi$, where S_3 is the group of all six permutations of three elements and the permutation operators act like $W_\pi|\psi_1\rangle \otimes |\psi_2\rangle \otimes |\psi_3\rangle = |\psi_{\pi^{-1}(1)}\rangle \otimes |\psi_{\pi^{-1}(2)}\rangle \otimes |\psi_{\pi^{-1}(3)}\rangle$:

$$\mathrm{tr}\left(O_0^{(1,2)} \otimes O_0^{(1,2)} \otimes \rho \otimes \rho \; P_{\vee^3}^{(\mathrm{odd})} \otimes P_{\vee^3}^{(\mathrm{even})}\right) = \sum_{\pi,\tau\in S_3} \mathrm{tr}\left(O_0^{(1,2)} \otimes O_0^{(1,2)} \otimes \rho \otimes \rho \; W_\pi^{(\mathrm{odd})} \otimes W_\tau^{(\mathrm{even})}\right). \quad \text{(S79)}$$

The specific structure of $O_0^{(1,2)}$ implies that several contributions must vanish. Permutations that have either 1 or 2 as a fix-point lead to a partial trace of $O_0^{(1,2)}$ that evaluates to zero. There are only three permutations that do not have such fix-points: The flip $(1,2,3) \mapsto (2,1,3)$ and the two cycles $(1,2,3) \mapsto (3,1,2)$, $(1,2,3) \mapsto (2,3,1)$. There are in total $9 = 3^2$ potential combinations of such permutations. Each of them results in a trace expression that can be upper-bounded by Hilbert-Schmidt norms. For instance the pair flip and flip produces

$$\mathrm{tr}\left(O_0^{(1,2)} O_0^{(1,2)}\right) \mathrm{tr}(\rho)^2 = \left\|O_0^{(1,2)}\right\|_2^2. \quad \text{(S80)}$$

All other 8 contributions can also be bounded by this expression and we conclude

$$\mathbb{E}\left[(d+1)^4 \mathrm{tr}\left(O_0^{(1,2)} U_1^\dagger |\hat{b}_1\rangle\langle\hat{b}_1| U_1 \otimes U_2^\dagger |\hat{b}_2\rangle\langle\hat{b}_2| U_2\right)^2\right] \leq 9\left\|O_0^{(1,2)}\right\|_2^2 \quad \text{(S81)}$$

Bounds on cross-terms: Cross-terms are considerably easier to evaluate, because one (or both) random matrices only feature linearly. We can use $\mathbb{E}\left[U_a^\dagger |\hat{b}_a\rangle\langle\hat{b}_a| U_a\right] = \mathcal{D}_{1/(d+1)}(\rho) = \frac{\rho+\mathbb{I}}{d+1}$ to effectively get rid of the linear contribution. For instance,

$$\left(\frac{d+1}{d}\right)^2 \mathbb{E}\left[\prod_{a=1,2} \mathrm{tr}\left(O_0^{(1)} U_a^\dagger |\hat{b}_a\rangle\langle\hat{b}_a| U_a\right)\right] = \frac{1}{d^2}\mathrm{tr}\left(O_0^{(1)}\rho\right)\mathrm{tr}\left(O_0^{(2)}\rho\right) \leq \frac{1}{2d^2}\left(\|O_0^{(1)}\|_\infty^2 + \|O_0^{(2)}\|_\infty^2\right), \quad \text{(S82)}$$

where $\|\cdot\|_\infty$ denotes the operator norm. Cross terms that do feature the leading order term require slightly more work, but can be addressed in a similar fashion. Using linearity in one snapshot reduces the expression to an expectation of a quadratic function in one snapshot only. The remaining computation is similar to the proof of Proposition S1 and yields

$$\frac{(d+1)^3}{d}\mathbb{E}\left[\mathrm{tr}\left(O_0^{(1,2)} U_1^\dagger |\hat{b}_1\rangle\langle\hat{b}_1| U_1 \otimes U_2^\dagger |\hat{b}_2\rangle\langle\hat{b}_2| U_2\right)\mathrm{tr}\left(O_0^{(a)} U_a^\dagger |\hat{b}_a\rangle\langle\hat{b}_a| U_a\right)\right] \leq \frac{3}{2d^2}\left(\|\tilde{O}_\rho^{(a)}\|_2^2 + \|O_0^{(a)}\|_2^2\right), \quad \text{(S83)}$$

for $a = 1, 2$, as well as $\tilde{O}_\rho^{(1)} = \mathrm{tr}_2\left(\mathbb{I} \otimes \rho O\right)$ and $\tilde{O}_\rho^{(2)} = \mathrm{tr}_1\left(\rho \otimes \mathbb{I} O\right)$, respectively.

Full variance bound: We are now ready to combine all individual bounds to control the full variance:

$$\begin{aligned}
\mathrm{Var}\left[\hat{o}\right] \leq & \mathbb{E}\left((d+1)^2 \mathrm{tr}\left(O_0^{(1,2)} U_1^\dagger |\hat{b}_1\rangle\langle\hat{b}_1| U_1 \otimes U_2^\dagger |\hat{b}_2\rangle\langle\hat{b}_2| U_2\right) + \sum_{a=1,2} \frac{d+1}{d}\mathrm{tr}\left(O_0^{(a)} U_a^\dagger |\hat{b}_a\rangle\langle\hat{b}_a| U_a\right)\right)^2 \\
\leq & 9\|O_0^{(1,2)}\|_2^2 + \frac{6}{2d^2}\left(\|\mathrm{tr}_2\left(\mathbb{I} \otimes \rho O\right)\|_2^2 + \|O_0^{(1)}\|_2^2\right) + \frac{6}{2d^2}\left(\|\mathrm{tr}_1\left(\rho \otimes \mathbb{I} O\right)\|_2^2\right) \\
& + \frac{3}{d^2}\|O_0^{(1)}\|_2^2 + \frac{3}{d^2}\|O_0^{(2)}\|_2^2 + \frac{1}{2d^2}\left(\|O_0^{(1)}\|_\infty^2 + \|O_0^{(2)}\|_\infty^2\right). \quad \text{(S84)}
\end{aligned}$$

Standard norm inequalities, as well as the explicit expression for $\|O_0^{(1,2)}\|_2^2$ allow for counter-balancing some of the sub-leading terms and we conclude

$$\mathrm{Var}\left[\hat{o}\right] \leq 9\|O_0\|_2^2 + \frac{3}{d^2}\left(\|\mathrm{tr}_2\left(\mathbb{I} \otimes \rho O\right)\|_2^2 + \|\mathrm{tr}_1\left(\rho \otimes \mathbb{I} O\right)\|_2^2\right) \leq 9\|O_0\|_2^2 + \frac{6}{d}\|O\|_\infty^2. \quad \text{(S85)}$$

7. INFORMATION-THEORETIC LOWER BOUND WITH SCALING IN HILBERT-SCHMIDT NORM

Before stating the content of the statement, we need to introduce some additional notation. In quantum mechanics, the most general notion of a quantum measurement is a POVM (positive operator-valued measure). A d-dimensional POVM F consists of a collection $F_1, \ldots, F_N$ of positive semidefinite matrices that sum up to the identity matrix: $\langle x|F_i|x\rangle \geq 0$ for all $x \in \mathbb{C}^d$ and $\sum_i F_i = \mathbb{I}$. The index i is associated with different potential measurement outcomes and Born's rule asserts $\Pr[i|\rho] = \mathrm{tr}(F_i\rho)$ for all $1 \leq i \leq M$ and any d-dimensional quantum state ρ. We present a simplified version of the proof by consider the relevant case where $M \leq \exp(2^n/32)$. The full proof can be found in [35].

A. Detailed statement and proof idea

Theorem S5 (Detailed restatement of Theorem 2 for Hilbert-Schmidt norm)**.** *Fix a sequence of POVMs $F^{(1)}, \ldots, F^{(N)}$. Suppose that given any M features $0 \preceq O_1, O_2, \ldots, O_M \preceq I$ with $\max_i \left(\|O_i\|_2^2 \right) \leq B$, there exists a machine (with arbitrary runtime as long as it always terminates) that can use the measurement outcomes of $F^{(1)}, \ldots, F^{(N)}$ on N copies of an unknown d-dimensional quantum state ρ to ϵ-accurately predict $\mathrm{tr}(O_1\rho), \ldots, \mathrm{tr}(O_M\rho)$ with high probability. Assuming $M \leq \exp(d/32)$, then necessarily*

$$N \geq \Omega\left(\frac{B \log(M)}{\epsilon^2} \right). \tag{S86}$$

It is worthwhile to put this statement into context and discuss consequences, as well as limitations. Theorem 1 (Clifford measurements) equips classical shadows with a *universal* convergence guarantee: (order) $\log(M) \max_i \mathrm{tr}(O_i^2)/\epsilon^2$ single-copy measurements suffice to accurately predict *any* collection of M target functions in *any* state. Theorem S5 implies that there are cases where this number of measurements is unavoidable. This highlights that the sample complexity of feature prediction with classical shadows is optimal in the worst case – a feature also known as minimax optimality.

Minimax optimality, however, does not rule out potential for further improvement in certain best-case scenarios. Advantageous structure in ρ or the O_i's (or both) can facilitate the design of more efficient prediction techniques. Prominent examples include matrix product state tomography (MPST) [16, 45] and neural network tomography (NNQST) [13]. Such tailored approaches, however, hinge on additional assumptions about the states to be measured or the properties to be predicted.[3]

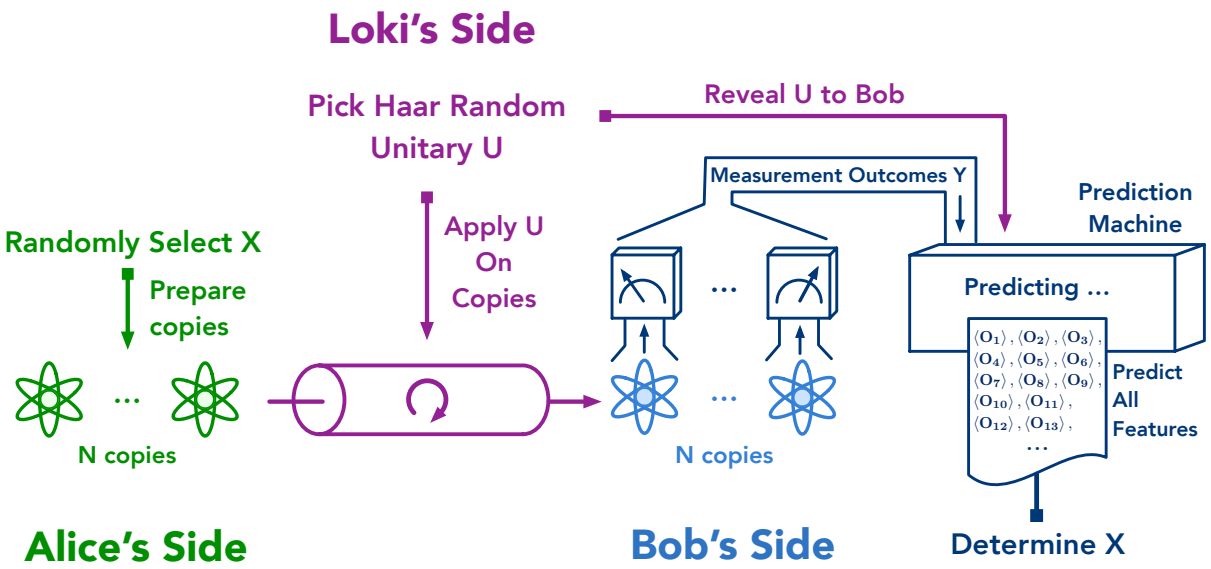
Finally, we emphasize that Theorem 2 only applies to single-copy measurements. Another way to bypass this lower bound is to use joint quantum measurements that act on all copies of the quantum state ρ simultaneously. Although very challenging to implement, such procedures can get by with substantially fewer state copies while still being universal. Shadow tomography [1, 3] is a prominent example.

Proof idea: We adapt a versatile proof technique for establishing information-theoretic lower bounds on tomographic procedures that is originally due to Flammia *et al.* [22]; see also [32, 55] for adaptations and refinements. The key idea is to consider a communication task in which Alice chooses a quantum state from among an alphabet of possible states and then sends copies of her chosen state to Bob, who measures all the copies hoping to extract a classical message from Alice. If we choose Alice's alphabet suitably, then by learning many properties of Alice's state Bob will be able to identify the state, hence decoding Alice's message. Information-theoretical lower bounds on the number of copies Bob needs to decode the message can therefore be translated into lower bounds on how many copies Bob needs to learn the properties.

To be more specific, suppose Alice chooses her state from an ensemble of M possible n-qubit signal states $\{\rho_1, \rho_2, \ldots \rho_M\}$ and suppose there are M linear operators $\{O_1, O_2, \ldots O_M\}$, each with $\mathrm{tr}\left(O_i^2\right) \leq B$, such that learning the expectation values of all the operators $\{O_i\}$ up to an additive error ϵ suffices to determine ρ_i uniquely. Suppose furthermore that if Bob receives N copies of *any* n-qubit state, and measures them one at a time, he is able to learn all of the properties $\{O_i\}$ with an additive error no larger than ϵ with high success probability. This provides Bob with a method for identifying the state ρ_i with high probability. Therefore, if Alice chooses her signal state uniformly at random from among the M possible states, by performing the

---

[3] Although tractable in theory, MPST becomes prohibitively expensive if ρ is not well-approximated by a MPS with small bond dimension. Likewise, NNQST seems to struggle to identify quantum states with intricate combinatorial structure, such as toric code ground states. We refer to the other supplementary sections for numerical (Supplementary Section 2 A) and theoretical (Supplementary Section 4 B) support of this claim.



Supplementary Figure 4: *Illustration of the communication protocol behind Theorem S5 and Theorem S6.* Two parties (Alice and Bob) devise a protocol that allows them to communicate classical bit strings: Alice encodes a bit string X in a quantum state and sends N independent copies of the state to Bob. Bob performs quantum measurements and uses a black box device (e.g. classical shadows) to decode Alice's original message. An unpredictable trickster (Loki) tampers with this procedure by randomly rotating Alice's quantum states en route to Bob. Loki reveals his actions only after Bob has completed the measurement stage of his protocol.

appropriate single-copy measurements Bob can acquire $\log_2 M$ bits of information about Alice's message. A lower bound on how many copies Bob needs to gain $\log_2 M$ bits of information about Alice's state, then, becomes a lower bound on how many copies Bob needs to learn the M properties $\{O_i\}$. To get the best possible lower bound, we choose Alice's signal ensemble $\{\rho_i\}$ so that it is as hard as possible for Bob to distinguish the signals using properties with $\mathrm{tr}\left(O_i^2\right) \leq B$.

So far, this lower bound on N would apply even if Bob has complete knowledge of Alice's signal states and the properties he should learn to distinguish them. We can derive a stronger lower bound on N by invoking a powerful feature of classical shadows — that Bob must make his measurements *before* he finds out which properties he must learn. To obtain this stronger bound, we introduce into the communication scenario a third party, named Loki[4], who tampers with the signal states. Loki chooses a Haar-random n-qubit unitary U, and replaces all N copies of Alice's signal state ρ_i by the rotated states $U\rho_i U^\dagger$ before presenting the states to Bob (Loki's mischief).

If Bob knew Loki's unitary U, he could modify his measurement procedure to learn the rotated properties $\{UO_iU^\dagger\}$. These rotated properties are just as effective for distinguishing the rotated states as the unrotated properties were effective for distinguishing the unrotated states. However, Loki keeps U secret, so Bob is forced to perform his measurements on the rotated states without knowing U. Only after Bob's data acquisition phase is completed does Loki confide in Bob and provide him with a full classical description of the unitary he applied earlier (Loki's redemption). This three-party scenario is illustrated in Supplementary Figure 4.

Suppose, though, that using the classical shadow based on his measurements, Bob can predict *any* M properties (with additive error bounded by ϵ and with high success probability), provided that the Hilbert-Schmidt norm is no larger than $\sqrt{B}$ for each property. Then he is just as well equipped to learn $\{UO_iU^\dagger\}$ as $\{O_i\}$, and can therefore decode Alice's message successfully once Loki reveals U. It must be, then, that Bob's measurement outcomes provide $\log_2 M$ bits of information about Alice's prepared state, when U is known. This is the idea we use to derive the stronger upper bound on N, and hence prove Theorem S5.

We emphasize again that quantum feature prediction with classical shadows can cope with Loki's mischief, by merely rotating the features Bob predicts, because the predicted features need not be known at the time Bob measures. The lower bound in Theorem S5 does not apply to the task of learning features that are already

[4] In Norse mythology, Loki is infamous for mischief and trickery. However, not entirely malicious, he often shows up in the nick of time to remedy the dire consequences of his actions.

known in advance. We also emphasize again that Theorem S5 assumes that the copies of the state are measured individually. It does not apply to protocols where collective measurements are applied across many copies.

B. Description of the communication protocol

We show how Alice can communicate any integer in $\{1, \ldots, M\}$ to Bob. Alice and Bob first agree on a codebook for encoding any integer selected from $\{1, \ldots, M\}$ in a d-dimensional quantum state. We denote these codebook states by $\rho_1, \ldots, \rho_M$. Alice and Bob also agree on a set of linear features $O_1, \ldots, O_M$ that satisfies

$$\mathrm{tr}(O_i \rho_i) \geq \max_{j \neq i} \mathrm{tr}(O_j \rho_i) + 3\epsilon. \tag{S87}$$

Therefore, if each feature can be predicted with additive error ϵ, these features can be used to identify the state ρ_i. The communication protocol between Alice and Bob is now apparent:

1. Alice randomly selects an integer X from $\{1, \ldots, M\}$.
2. Alice prepares N copies of the code-state ρ_X associated to X and sends them to Bob.
3. Bob performs POVMs $F^{(i)}$ on individual states and receives a string of measurement outcomes Y.
4. Bob inputs Y into the feature prediction machine to estimate $\mathrm{tr}(O_1 \rho_X), \ldots, \mathrm{tr}(O_M \rho_X)$.
5. Bob finds $\overline{X}$ that has the largest $\mathrm{tr}(O_{\overline{X}} \rho_X)$.

The working assumption is that the feature prediction machine can estimate $\mathrm{tr}(O_1 \rho_X), \ldots, \mathrm{tr}(O_M \rho_X)$ within ϵ-error and high success probability. This in turn ensures that this plain communication protocol is mostly successful, i.e. $\overline{X} = X$ with high probability. In words: Alice can transmit information to Bob, when no adversary is present.

We now show how they can still communicate safely in the presence of an adversary (Loki) who randomly rotates the transmitted code states en route: $\rho_X \mapsto U \rho_X U^\dagger$ and U is a Haar-random unitary.

This random rotation affects the measurement outcome statistics associated with the fixed POVMs $F^{(1)}, \ldots, F^{(N)}$. Each element of $Y = \left[Y^{(1)}, \ldots, Y^{(N)}\right]$ is now a random variable that depends on both X and U. After Bob has performed the quantum measurements to obtain Y, the adversary confesses to Bob and reveals the random unitary U. While Bob no longer has any copies of ρ_X, he can still incorporate precise knowledge of U by instructing the machine to predict linear features $UO_1U^\dagger, \ldots, UO_MU^\dagger$, instead of the original $O_1, \ldots, O_M$. This reverses the effect of the original unitary transformation, because $\mathrm{tr}(UO_iU^\dagger U \rho_X U^\dagger) = \mathrm{tr}(O_i \rho_X)$. This modification renders the original communication protocol stable with respect to Loki's actions. Alice can still send any integer in $\{1, \ldots, M\}$ to Bob with high probability.

C. Information-theoretic analysis

The following arguments use properties of Shannon entropy and mutual information which can be found in standard textbooks on information theory, such as [15].

The communication protocol is guaranteed to work with high probability, ensuring that Bob's recovered message $\bar{X}$ equals Alice's input X with high probability. Moreover, we assume that Alice selects her message uniformly at random. Fano's inequality then implies

$$I(X : \overline{X}) = H(X) - H(X|\overline{X}) \geq \Omega(\log(M)), \tag{S88}$$

where $I(X : \overline{X})$ is the mutual information, and $H(X)$ is the Shannon entropy. By assumption, Loki chooses the unitary roatation U uniformly at random, regardless of the message X. This implies $I(X : U) = 0$ and, in turn

$$I(X : \overline{X}) \leq I(X : \overline{X}, U) = I(X : U) + I(X : \overline{X}|U) = I(X : \overline{X}|U). \tag{S89}$$

For fixed U, $\overline{X}$ is the output of the machine that only takes into account the measurement outcomes Y. The data processing inequality then yields

$$I(X : Y|U) \geq I(X : \overline{X}|U) \geq I(X : \overline{X}) \geq \Omega(\log(M)). \tag{S90}$$

Recall that Y is the measurement outcome of the N POVMs $F_1, \ldots, F_N$. We denote the measurement outcome of F_k as Y_k. Because $Y_1, \ldots, Y_N$ are random variables that depend on X and U,

$$\begin{aligned} I(X:Y|U) &= H(Y_1,\ldots,Y_N|U) - H(Y_1,\ldots,Y_N|X,U) \\ &\leq H(Y_1|U) + \ldots + H(Y_N|U) - H(Y_1,\ldots,Y_N|X,U) \\ &= \sum_{k=1}^{N} \Big(H(Y_k|U) - H(Y_k|X,U) \Big) = \sum_{k=1}^{N} I(X:F_k \text{ on } U\rho_X U^\dagger|U). \end{aligned} \tag{S91}$$

The second to last equality uses the fact that when X, U are fixed, $Y_1, \ldots, Y_N$ are independent. To obtain the best lower bound, we should choose Alice's signal states $\{\rho_i\}$ such that $I(X : F_k \text{ on } U\rho_X U^\dagger|U)$ is as small as possible. In Sec. 7 D, we will see that, no matter how Bob chooses his measurements $\{F_1, F_2, \ldots, F_N\}$, there are signal states satisfying (S87) such that

$$I(X:F_k \text{ on } U\rho_X U^\dagger|U) \leq \frac{36\epsilon^2}{B}, \forall k. \tag{S92}$$

Assuming that this relation holds, we have established a connection between M and N: $\Omega(\log(M)) \leq I(X : Y|U) \leq 36N\epsilon^2/B$ and, therefore, $N \geq \Omega\Big(B\log(M)/\epsilon^2\Big)$. This establishes the claim in Theorem S5.

D. Detailed construction of quantum encoding and linear prediction decoding

We now construct a codebook $\rho_1, \ldots, \rho_M$ and linear features $0 \preceq O_1, O_2, \ldots, O_M \preceq \mathbb{I}$ with $\max_i \|O_i\|_2^2 \leq B$ that obey two key properties:

1. the code states $\rho_1, \ldots, \rho_M$ obey the requirement displayed in Eq. (S92).
2. the linear features $O_1, \ldots, O_M$ are capable of identifying a unique code state:

$$\operatorname{tr}(O_i\rho_i) \geq \max_{j\neq i} \operatorname{tr}(O_j\rho_i) + 3\epsilon \quad \text{for all} \quad 1 \leq i \leq M. \tag{S93}$$

The second condition requires each ρ_i to be distinguishable from $\rho_1, \ldots, \rho_M$ via linear features O_i. The first condition, on the contrary, requires ρ_X to convey as little information about X as possible. The general idea would then be to create distinguishable quantum states that are, at the same time, very similar to each other.

In order to achieve these two goals, we choose M rank-$B/4$ subspace projectors $\Pi_1, \ldots, \Pi_M$ that obey $\operatorname{tr}(\Pi_i\Pi_j)/r < 1/2$ for all $i \neq j$. The probabilistic method asserts that such a projector configuration exists; see Lemma S7 below. Now, we set

$$\rho_i = (1-3\epsilon)\frac{\mathbb{I}}{d} + 3\epsilon\frac{4\Pi_i}{B}, \quad \text{and} \quad O_i = 2\Pi_i, \quad \text{for all} \quad 1 \leq i \leq M. \tag{S94}$$

It is easy to check that this construction meets the requirement displayed in Eq. (S93). The other condition – Eq. (S92) is verified in Lemma S8 below.

Lemma S7. *If $M \leq \exp(rd/32)$ and $d \geq 4r$, then $\exists M$ rank-r subspace projectors $\Pi_1, \ldots, \Pi_M$ such that*

$$\operatorname{tr}(\Pi_i\Pi_j)/r < 1/2, \forall i \neq j. \tag{S95}$$

Proof. We find the subspace projectors using a probabilistic argument. We randomly choose M rank-r subspaces according to the unitarily invariant measure in the Hilbert space, the Grassmannian, and bound the probability that the randomly chosen subspaces do not satisfy the condition. For a pair of fixed $i \neq j$, we have

$$\Pr\left[\frac{1}{r}\operatorname{tr}(\Pi_i\Pi_j) \geq \frac{1}{2}\right] \leq \exp\left(-r^2 f\left(\frac{d}{2r}-1\right)\right) < \exp\left(-\frac{rd}{16}\right), \tag{S96}$$

where we make use of [32, Lemma 6] in the first inequality and $f(z) = z - \log(1+z) > z/4$ for all $z \geq 1$ in the second inequality. A union bound then asserts

$$\Pr\left[\exists i \neq j, \frac{1}{r}\operatorname{tr}(\Pi_i\Pi_j) \geq \frac{1}{2}\right] < M^2 \exp\left(-\frac{rd}{16}\right) \leq 1. \tag{S97}$$

Because the probability is less than one, there must exist $\Pi_1, \ldots, \Pi_M$ that satisfy the desired property. □

Lemma S8. *Consider a set of d-dimensional quantum states $\{\rho_1, \ldots, \rho_M\}$ such that $\rho_i = (1-\alpha)\frac{\mathbb{I}}{d} + \alpha\frac{\Pi_i}{r}$, where Π_i is a rank-r subspace projector. Consider U sampled from Haar measure, and X sampled from $\{1, \ldots, M\}$ uniformly at random. Consider any POVM measurement F. Then the information gain regarding X, conditioned on U, obtained from the measurement F performed on the state $U\rho_X U^\dagger$ satisfies*

$$I(X : F \text{ on } U\rho_X U^\dagger | U) \leq \frac{\alpha^2}{r}. \tag{S98}$$

Note that we can obtain the statement (S92) by choosing $\alpha = 3\epsilon$ and $r = B/4$, hence completing the proof of Theorem S5.

Proof. First of all, let us decompose all POVM elements $\{F_1, \ldots, F_l\}$ to rank-1 elements $F' = \left\{ w_i d \left|v_i\right\rangle \left\langle v_i\right| \right\}_{i=1}^{l'}$, where $l \leq l'$. We can perform measurement F by performing measurement with F': when we measure a rank-1 element, we return the original POVM element the rank-1 element belongs to. Using data processing inequality, we have $I(X : F \text{ on } U\rho_X U^\dagger|U) \leq I(X : \tilde{F} \text{ on } U\rho_X U^\dagger|U)$. From now on, we can consider the POVM $\vec{F}$ to be $\left\{ w_i d \left|v_i\right\rangle \left\langle v_i\right| \right\}_{i=1}^{l}$. Normalization demands

$$\operatorname{tr}\Big(\sum_i w_i d \left|v_i\right\rangle \left\langle v_i\right| \Big) = \operatorname{tr}(\mathbb{I}) = d \quad \text{and therefore} \quad \sum_i w_i = 1. \tag{S99}$$

Let us define the probability vector $\vec{p} = \operatorname{tr}(U\rho_1 U^\dagger \vec{F})$, so $p_i = w_i d \left\langle v_i\right| U\rho_1 U^\dagger \left|v_i\right\rangle$. And the expression we hope to bound satisfies $I(X : F \text{ on } U\rho_X U^\dagger|U) = I(X, U : F \text{ on } U\rho_X U^\dagger) - I(U : F \text{ on } U\rho_X U^\dagger) \leq I(X, U : F \text{ on } U\rho_X U^\dagger)$ using the chain rule and the nonnegativity of mutual information. We now bound

$$\begin{aligned}
I(X, U : F \text{ on } U\rho_X U^\dagger) =& H\Big(\sum_{X=1}^{M} \frac{1}{M} \mathbb{E}_U[\operatorname{tr}(U\rho_X U^\dagger \vec{F})] \Big) - \sum_{X=1}^{M} \frac{1}{M} \mathbb{E}_U \left[H\Big(\operatorname{tr}(U\rho_X U^\dagger \vec{F}) \Big) \right] \\
=& H\Big(\operatorname{tr}(\mathbb{E}_U[U\rho_1 U^\dagger] \vec{F}) \Big) - \mathbb{E}_U \left[H\Big(\operatorname{tr}(U\rho_1 U^\dagger \vec{F}) \Big) \right] \\
=& \sum_i -(\mathbb{E}_U\, p_i) \log(\mathbb{E}_U\, p_i) + \mathbb{E}_U[p_i \log p_i] \\
\leq& \sum_i -(\mathbb{E}_U\, p_i) \log(\mathbb{E}_U\, p_i) + \mathbb{E}_U \left[p_i \log(\mathbb{E}_U\, p_i) + p_i \frac{p_i - \mathbb{E}_U\, p_i}{\mathbb{E}_U\, p_i} \right] \\
=& \sum_i \frac{\mathbb{E}_U[p_i^2] - \mathbb{E}_U[p_i]^2}{\mathbb{E}_U[p_i]}.
\end{aligned} \tag{S100}$$

The second equality uses the fact that $\mathbb{E}_U\, f(U\rho_X U^\dagger) = \ E_U f(U\rho_1 U^\dagger), \forall X$ which follows from the fact that $\forall X, \exists U_X, \rho_X = U_X \rho_1 U_X^\dagger$. The inequality uses the fact that $\log(x)$ is concave, so $\log(x) \leq \log(y) + \frac{x-y}{y}$. Using properties of Haar random unitary $d \times d$ matrices, we conclude

$$\mathbb{E}_U[p_i] = w_i, \ \mathbb{E}_U[p_i^2] = w_i^2 \frac{d}{(d+1)} \left(1 + \frac{1}{d} + \alpha^2 \Big(\frac{1}{r} - \frac{1}{d} \Big) \right). \tag{S101}$$

Therefore we have

$$\frac{\mathbb{E}_U[p_i^2] - \mathbb{E}_U[p_i]^2}{\mathbb{E}_U[p_i]} = w_i \alpha^2 \frac{d}{d+1} \Big(\frac{1}{r} - \frac{1}{d} \Big) \leq \frac{w_i \alpha^2}{r}, \tag{S102}$$

which establishes the claim:

$$I(X : F \text{ on } U\rho_X U^\dagger | U) \leq \sum_i \frac{\mathbb{E}_U[p_i^2] - \mathbb{E}_U[p_i]^2}{\mathbb{E}_U[p_i]} \leq \frac{\alpha^2}{r}. \tag{S103}$$

□

8. INFORMATION-THEORETIC BOUNDS ON PREDICTING LOCAL OBSERVABLES

In Theorem S5, we have shown that if a procedure can predict arbitrary observables with $\operatorname{tr}(O_i^2) \leq B$, then it must use at least $\Omega(B \log(M)/\epsilon^2)$ single-copy measurements (as long as M is not extraordinarily large). A

similar argument can be used to show that if a procedure can predict arbitrary k-local observables, then it requires at least $\Omega(2^k \log(M)/\epsilon^2)$ single-copy measurements (when M is not too large). This is because if we focus on a k-qubit subsystem, then the guarantee allows us to predict arbitrary observables $0 \preceq O_i \preceq \mathbb{I}$ with $\mathrm{tr}(O_i^2) \leq 2^k$. In the following theorem, we show a stronger lower bound by focusing on local measurements. A local measurement is a POVM $\{w_i d\,|v_i\rangle\langle v_i|\}_i$ where $|v_i\rangle = |v_i^{(1)}\rangle \otimes \ldots \otimes |v_i^{(n)}\rangle$, $\sum_i w_i = 1$, and $d = 2^n$. This is the same as not performing any entangling gates when implementing the measurement. (Random) Pauli basis measurements are a prominent example.

Theorem S6 (Detailed restatement of Theorem 2 for exponential scaling in locality)**.** *Fix a sequence of local measurements $F_1, \ldots, F_N$ on n-qubit system, i.e., $F_j = \{w_{j,i} d\,|v_{j,i}\rangle\langle v_{j,i}|\}_i$ where $|v_{j,i}\rangle = |v_{j,i}^{(1)}\rangle \otimes \ldots \otimes |v_{j,i}^{(n)}\rangle$, $\sum_i w_{j,i} = 1$, and $d = 2^n$. Suppose that given any M k-local observables $-\mathbb{I} \preceq O_1, O_2, \ldots, O_M \preceq \mathbb{I}$, there exists a machine (with arbitrary runtime as long as it always terminates) that can use the measurement outcomes of $F_1, \ldots, F_N$ on N copies of an unknown quantum state ρ to ϵ-accurately predict $\mathrm{tr}(O_1\rho), \ldots, \mathrm{tr}(O_M\rho)$ with high probability. Assuming $M \leq 3^k \binom{n}{k}$, then necessarily*

$$N \geq \Omega\left(\frac{3^k \log(M)}{\epsilon^2}\right). \tag{S104}$$

Proof. The proof uses a quantum communication protocol between Alice and Bob, with Loki interfering in the middle. Alice would encode some classical information in the quantum state and send to Bob. Bob would then use the prediction procedure to decode the encoded classical information. In the middle, Loki will alter the quantum state by applying a random unitary. Loki would then reveal the random unitary to Bob after Bob performed quantum measurements on the quantum states. An illustration of the communication protocol can be found in Supplementary Figure 4. The quantum state Alice encodes, the unitary applied by Loki, and the features predicted by Bob is considerably simplified in this result compared to the previous proof.

We define $\rho_i = (\mathbb{I} + 3\epsilon P_i)/2^n, \forall i = 1, \ldots, M$. P_i is the i-th Pauli observable acting on k qubits in the n-qubit system. Any ordering of the Pauli observables is fine. Note that there are at most $3^k \binom{n}{k}$ such Pauli observables. This is the reason why we assume $M \leq 3^k \binom{n}{k}$. The corresponding linear functions chosen by Bob are $O_i = P_i, \forall i = 1, \ldots, M$. This guarantees the following relation:

$$\mathrm{tr}(O_i \rho_j) = 3\epsilon\delta_{ij} \quad \text{for all } 1 \leq i, j \leq M, \tag{S105}$$

where δ_{ij} is the Kronecker-delta ($\delta ij = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise). The random unitary applied by Loki consists of random single-qubit unitary rotations, i.e. $U = U^{(1)} \otimes \ldots \otimes U^{(n)}$. The complete communication protocol works as follows.

1. Alice randomly selects an integer X from $\{1, \ldots, M\}$.
2. Alice prepares N copies of the code-state ρ_X according associated to X and sends them to Bob.
3. Loki intercepts the N copies, samples a random unitary $U = U^{(1)} \otimes \ldots \otimes U^{(n)}$, applies U on all copies of $\rho_X \to U\rho_X U^\dagger$, and sends to Bob.
4. Bob performs local measurements F_j on individual states and receives a string of measurement outcomes Y.
5. Loki reveals the random unitary U to Bob. Now Bob would have to predict the expectation value of $UO_1U^\dagger, \ldots, UO_MU^\dagger$ instead of the original $O_1, \ldots, O_M$.
6. Since $UO_1U^\dagger, \ldots, UO_MU^\dagger$ are still k-local observables, Bob can input Y into the feature prediction machine to estimate $\langle UO_iU^\dagger\rangle_{U\rho_X U^\dagger} = \mathrm{tr}(O_i\rho_X), \forall i = 1, \ldots, M$.
7. Bob finds $\overline{X} \in \{1, \ldots, M\}$ that has the largest $\mathrm{tr}(O_{\overline{X}}\rho_X)$.

Because $\mathrm{tr}(O_i\rho_X)$ are predicted to ϵ additive error, and $\mathrm{tr}(O_i\rho_X) = 3\epsilon\delta_{iX}$, if the prediction procedure works as guaranteed, Bob's decoded information $\hat{X}$ would be equal to Alice's encoded information X with high probability. Moreover, we assume that Alice selects her message uniformly at random. Fano's inequality then implies

$$I(X : \overline{X}) = H(X) - H(X|\overline{X}) \geq \Omega(\log(M)), \tag{S106}$$

where $I(X : \overline{X})$ is the mutual information, and $H(X)$ is the Shannon entropy. By assumption, Loki chooses the random unitary U regardless of the message X. This implies $I(X : U) = 0$ and, in turn

$$I(X : \overline{X}) \leq I(X : \overline{X}, U) = I(X : U) + I(X : \overline{X}|U) = I(X : \overline{X}|U). \tag{S107}$$

For fixed U, $\overline{X}$ is the output of the machine that only takes into account the measurement outcomes Y. The data processing inequality then implies

$$I(X:Y|U) \geq I(X:\overline{X}|U) \geq I(X:\overline{X}) \geq \Omega(\log(M)). \tag{S108}$$

Recall that Y is the measurement outcome of the N POVMs $F_1, \ldots, F_N$. We denote the measurement outcome of F_j as Y_j. Because $Y_1, \ldots, Y_N$ are random variables that depend on X and U,

$$\begin{aligned} I(X:Y|U) &= H(Y_1,\ldots,Y_N|U) - H(Y_1,\ldots,Y_N|X,U) \\ &\leq H(Y_1|U) + \ldots + H(Y_N|U) - H(Y_1,\ldots,Y_N|X,U) \\ &= \sum_{j=1}^{N} \Big(H(Y_j|U) - H(Y_j|X,U) \Big) = \sum_{j=1}^{N} I(X:F_j \text{ on } U\rho_X U^\dagger|U). \end{aligned} \tag{S109}$$

The second to last equality uses the fact that when X, U are fixed, $Y_1, \ldots, Y_N$ are independent. This part of the derivation is exactly the same as in Supplementary Section 7 C. All that is left is to properly upper bound $I(X:F_j \text{ on } U\rho_X U^\dagger|U)$. First, by definition,

$$\begin{aligned} I(X:F_j \text{ on } U\rho_X U^\dagger|U) &= \mathbb{E}_U \left[H(F_j \text{ on } U\rho_X U^\dagger) - H(X, F_j \text{ on } U\rho_X U^\dagger) \right] \\ &= \mathbb{E}_U \left[H\left(\mathbb{E}_X \operatorname{tr}(U\rho_X U^\dagger \vec{F}_j) \right) - \mathbb{E}_X H\left(\operatorname{tr}(U\rho_X U^\dagger \vec{F}_j) \right) \right] \\ &\leq H\left(\mathbb{E}_X \mathbb{E}_U \operatorname{tr}(U\rho_X U^\dagger \vec{F}_j) \right) - \mathbb{E}_X \mathbb{E}_U H\left(\operatorname{tr}(U\rho_X U^\dagger \vec{F}_j) \right). \end{aligned} \tag{S110}$$

The last inequality exploits concavity of the Shannon entropy $H(\cdot)$. By assumption, the F_j's must be local measurements, i.e. $F_j = \{w_{j,i} d\, |v_{k,i}\rangle\langle v_{k,i}|\}_i$ where $|v_{k,i}\rangle = |v_{k,i}^{(1)}\rangle \otimes \ldots \otimes |v_{k,i}^{(n)}\rangle$, $\sum_i w_i = 1$, and $d = 2^n$. We define the probability of measuring i-th outcome using POVM F_j as

$$p_{j,i} = w_{j,i} d \, \langle v_{j,i}| \, U\rho_X U^\dagger \, |v_{j,i}\rangle \,, \tag{S111}$$

which is a random number depending on X and U. Using Equation (S110) and the definition of $H(\cdot)$, we have

$$\begin{aligned} I(X:F_j \text{ on } U\rho_X U^\dagger|U) &\leq H\left(\mathbb{E}_X \mathbb{E}_U \operatorname{tr}(U\rho_X U^\dagger \vec{F}^{(k)}) \right) - \mathbb{E}_X \mathbb{E}_U H\left(\operatorname{tr}(U\rho_X U^\dagger \vec{F}^{(k)}) \right) \\ &= \sum_i \left(\mathbb{E}_{X,U}[p_{j,i}\log(p_{j,i})] - \mathbb{E}_{X,U}[p_{j,i}] \log(\mathbb{E}_{X,U}[p_{j,i}]) \right) \\ &\leq \sum_i -(\mathbb{E}_{X,U}\, p_{j,i}) \log(\mathbb{E}_{X,U}\, p_{j,i}) + \mathbb{E}_{X,U}\left[p_{j,i} \log(\mathbb{E}_{X,U}\, p_{j,i}) + p_{j,i} \frac{p_{j,i} - \mathbb{E}_{X,U}\, p_{j,i}}{\mathbb{E}_{X,U}\, p_{j,i}} \right] \\ &= \sum_i \frac{\mathbb{E}_{X,U}[p_{j,i}^2] - \mathbb{E}_{X,U}[p_{j,i}]^2}{\mathbb{E}_{X,U}[p_{j,i}]}. \end{aligned} \tag{S112}$$

The second inequality uses the fact that $\log(x)$ is concave, so $\log(x) \leq \log(y) + \frac{x-y}{y}$. We now compute $\mathbb{E}_{X,U}[p_{j,i}]$ and $\mathbb{E}_{X,U}[p_{j,i}^2]$ by using the following relation for single-qubit random unitary:

$$\mathbb{E}_{U^{(j)}}\left[U^{(j)} \, |v_{k,i}^{(j)}\rangle\langle v_{k,i}^{(j)}| \, (U^{(j)})^\dagger \right] = \frac{\mathbb{I}^{(j)}}{2}, \quad \mathbb{E}_{U^{(j)}}\left[\left(U^{(j)} \, |v_{k,i}^{(j)}\rangle\langle v_{k,i}^{(j)}| \, (U^{(j)})^\dagger \right)^{\otimes 2} \right] = \frac{\mathbb{I}^{(j)} \otimes \mathbb{I}^{(j)} + S^{(j)}}{3}, \tag{S113}$$

where j refers to the j-th qubit, and S is the two qubit swap operator ($|\psi\rangle \otimes |\phi\rangle = |\phi\rangle \otimes |\psi\rangle$). Recall the definition of $p_{j,i}$ in Equation (S111). Together with the above relation, we have

$$\begin{aligned} \mathbb{E}_{X,U}[p_{j,i}] &= \mathbb{E}_X\left[w_{j,i} d \operatorname{tr}\left(\rho_X \frac{\mathbb{I}}{2^n} \right) \right] = \mathbb{E}_X\left[w_{j,i} 2^n \operatorname{tr}\left(\frac{\mathbb{I} + 3\epsilon P_X}{2^n} \frac{\mathbb{I}}{2^n} \right) \right] = w_{j,i} \quad \text{and} \\ \mathbb{E}_{X,U}[p_{j,i}^2] &= \mathbb{E}_X\left[w_{j,i}^2 d^2 \operatorname{tr}\left(\rho_X^{\otimes 2} \bigotimes_{j=1}^{n} \left(\frac{\mathbb{I}^{(j)} \otimes \mathbb{I}^{(j)} + S^{(j)}}{3} \right) \right) \right] = w_{j,i}^2 \left(1 + \frac{9\epsilon^2}{3^k} \right). \end{aligned} \tag{S114}$$

Putting this computation into Inequality (S112), we have obtained

$$I(X:F_j \text{ on } U\rho_X U^\dagger|U) \leq \sum_i w_{j,i} \frac{9\epsilon^2}{3^k} = \frac{9\epsilon^2}{3^k}. \tag{S115}$$

Combining the above result with Inequality (S108) and (S109), we have

$$\frac{9N\epsilon^2}{3^k} \geq I(X:Y|U) \geq \Omega(\log(M)) \quad \text{which implies} \quad N \geq \Omega\left(\frac{3^k \log(M)}{\epsilon^2}\right). \tag{S116}$$

□

[1] S. Aaronson. Shadow tomography of quantum states. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, pages 325–338, New York, NY, USA, 2018. ACM.

[2] S. Aaronson and D. Gottesman. Improved simulation of stabilizer circuits. *Phys. Rev. A*, 70:052328, Nov 2004.

[3] S. Aaronson and G. N. Rothblum. Gentle measurement of quantum states and differential privacy. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2019, page 322–333, New York, NY, USA, 2019. Association for Computing Machinery.

[4] A. Acín, D. Bruß, M. Lewenstein, and A. Sanpera. Classification of mixed three-qubit states. *Phys. Rev. Lett.*, 87:040401, Jul 2001.

[5] K. Banaszek, M. Cramer, and D. Gross. Focus on quantum tomography. *New J. Phys.*, 15(12):125020, Dec 2013.

[6] A. Blum, A. Kalai, and H. Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *J. ACM*, 50(4):506–519, July 2003.

[7] R. Blume-Kohout. Optimal, reliable estimation of quantum states. *New J. Phys.*, 12(4):043034, Apr 2010.

[8] X. Bonet-Monroig, R. Babbush, and T. E. O'Brien. Nearly optimal measurement scheduling for partial tomography of quantum states. *arXiv preprint arXiv:1908.05628*, 2019.

[9] F. G. Brandão, W. Chemissany, N. Hunter-Jones, R. Kueng, and J. Preskill. Models of quantum complexity growth. *arXiv preprint arXiv:1912.04297*, 2019.

[10] F. G. Brandão, A. Kalev, T. Li, C. Y.-Y. Lin, K. M. Svore, and X. Wu. Quantum SDP solvers: Large speed-ups, optimality, and applications to quantum learning. In *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

[11] H. J. Briegel, D. E. Browne, W. Dür, R. Raussendorf, and M. Van den Nest. Measurement-based quantum computation. *Nat. Phys.*, 5:19–26, Jan 2009.

[12] T. Brydges, A. Elben, P. Jurcevic, B. Vermersch, C. Maier, B. P. Lanyon, P. Zoller, R. Blatt, and C. F. Roos. Probing Rényi entanglement entropy via randomized measurements. *Science*, 364(6437):260–263, 2019.

[13] J. Carrasquilla, G. Torlai, R. G. Melko, and L. Aolita. Reconstructing quantum states with generative models. *Nat. Mach. Intell.*, 1(3):155–161, 2019.

[14] J. Cotler and F. Wilczek. Quantum overlapping tomography. *Phys. Rev. Lett.*, 124:100401, Mar 2020.

[15] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2006.

[16] M. Cramer, M. B. Plenio, S. T. Flammia, R. Somma, D. Gross, S. D. Bartlett, O. Landon-Cardinal, D. Poulin, and Y.-K. Liu. Efficient quantum state tomography. *Nat. Commun.*, 1:149, 2010.

[17] M. P. da Silva, O. Landon-Cardinal, and D. Poulin. Practical characterization of quantum devices without tomography. *Phys. Rev. Lett.*, 107(21):210404, 2011.

[18] E. Dennis, A. Kitaev, A. Landahl, and J. Preskill. Topological quantum memory. *Journal of Mathematical Physics*, 43(9):4452–4505, 2002.

[19] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, New York, 1993.

[20] J. Emerson, R. Alicki, and K. Życzkowski. Scalable noise estimation with random unitary operators. *J. Opt. B Quantum Semiclass. Opt.*, 7(10):S347–S352, 2005.

[21] T. J. Evans, R. Harper, and S. T. Flammia. Scalable Bayesian Hamiltonian learning. *arXiv preprint arXiv:1912.07636*, 2019.

[22] S. T. Flammia, D. Gross, Y.-K. Liu, and J. Eisert. Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators. *New J. Phys.*, 14(9):095022, 2012.

[23] S. T. Flammia and Y.-K. Liu. Direct fidelity estimation from few Pauli measurements. *Phys. Rev. Lett.*, 106:230501, Jun 2011.

[24] N. Friis, G. Vitagliano, M. Malik, and M. Huber. Entanglement certification from theory to experiment. *Nat. Rev. Phys.*, 1(1):72–87, 2019.

[25] X. Gao and L.-M. Duan. Efficient representation of quantum many-body states with deep neural networks. *Nat. Commun.*, 8(1):662, 2017.

[26] D. Gosset and J. Smolin. A Compressed Classical Description of Quantum States. In *14th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2019)*, volume 135 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 8:1–8:9, Dagstuhl, Germany, 2019. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[27] D. Gottesman. *Stabilizer codes and quantum error correction. Caltech Ph. D.* PhD thesis, Thesis, eprint: quant-ph/9705052, 1997.

[28] D. Gross, F. Krahmer, and R. Kueng. A partial derandomization of PhaseLift using spherical designs. *J. Fourier Anal. Appl.*, 21(2):229–266, 2015.

[29] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert. Quantum state tomography via compressed sensing. *Phys. Rev. Lett.*, 105:150401, Oct 2010.

[30] M. Guta, J. Kahn, R. J. Kueng, and J. A. Tropp. Fast state tomography with optimal error bounds. *J. Phys. A*, 2020.

[31] O. Gühne and G. Tóth. Entanglement detection. *Phys. Rep.*, 474(1):1 – 75, 2009.

[32] J. Haah, A. W. Harrow, Z. Ji, X. Wu, and N. Yu. Sample-optimal tomography of quantum states. *IEEE T. Inform. Theory*, 63(9):5628–5641, 2017.

[33] W. Hoeffding. A class of statistics with asymptotically normal distribution. In *Breakthroughs in Statistics*, pages 308–334. Springer, 1992.

[34] Z. Hradil. Quantum-state estimation. *Phys. Rev. A*, 55:R1561–R1564, Mar 1997.

[35] H.-Y. Huang and R. Kueng. Predicting features of quantum systems using classical shadows. *arXiv preprint arXiv:1908.08909*, 2019.

[36] M. R. Jerrum, L. G. Valiant, and V. V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.*, 43(2-3):169–188, 1986.

[37] Z. Jiang, A. Kalev, W. Mruczkiewicz, and H. Neven. Optimal fermion-to-qubit mapping via ternary trees with applications to reduced quantum states learning. *arXiv preprint arXiv:1910.10746*, 2019.

[38] 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. *Phys. Rev. A*, 77:012307, Jan 2008.

[39] R. Koenig and J. A. Smolin. How to efficiently select an arbitrary Clifford group element. *J. Math. Phys.*, 55(12):122202, 12, 2014.

[40] C. Kokail, C. Maier, R. van Bijnen, T. Brydges, M. K. Joshi, P. Jurcevic, C. A. Muschik, P. Silvi, R. Blatt, C. F. Roos, et al. Self-verifying variational quantum simulation of lattice models. *Nature*, 569(7756):355–360, 2019.

[41] R. Kueng and D. Gross. Qubit stabilizer states are complex projective 3-designs. *arXiv preprint arXiv:1510.02767*, 2015.

[42] R. Kueng, H. Rauhut, and U. Terstiege. Low rank matrix recovery from rank one measurements. *Appl. Comput. Harmon. Anal.*, 42(1):88–116, 2017.

[43] R. Kueng, H. Zhu, and D. Gross. Low rank matrix recovery from Clifford orbits. *arXiv preprint arXiv:1610.08070*, 2016.

[44] O. Landon-Cardinal and D. Poulin. Practical learning method for multi-scale entangled states. *New J. of Phys.*, 14(8):085004, 2012.

[45] B. P. Lanyon, C. Maier, M. Holzäpfel, T. Baumgratz, C. Hempel, P. Jurcevic, I. Dhand, A. S. Buyskikh, A. J. Daley, M. Cramer, M. B. Plenio, R. Blatt, and C. F. Roos. Efficient tomography of a quantum many-body system. *Nat. Phys.*, 13:1158 EP –, Sep 2017.

[46] E. Magesan, J. M. Gambetta, and J. Emerson. Scalable and robust randomized benchmarking of quantum processes. *Phys. Rev. Lett.*, 106:180504, May 2011.

[47] A. S. Nemirovsky and D. B. a. Yudin. *Problem complexity and method efficiency in optimization.* A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

[48] M. A. Nielsen and I. L. Chuang. *Quantum computation and quantum information.* Cambridge University Press, Cambridge, 2000.

[49] R. O'Donnell and J. Wright. Efficient quantum tomography. In *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*, STOC '16, pages 899–912, New York, NY, USA, 2016. ACM.

[50] M. Paini and A. Kalev. An approximate description of quantum states. *arXiv preprint arXiv:1910.10543*, 2019.

[51] P. Raghavan. Probabilistic construction of deterministic algorithms: approximating packing integer programs. *Journal of Computer and System Sciences*, 37(2):130–143, 1988.

[52] R. Raussendorf and H. J. Briegel. A one-way quantum computer. *Phys. Rev. Lett.*, 86:5188–5191, May 2001.

[53] G. Refael and E. Altman. Strong disorder renormalization group primer and the superfluid–insulator transition. *C. R. Phys.*, 14(8):725–739, 2013.

[54] O. Regev. On lattices, learning with errors, random linear codes, and cryptography. *J. ACM*, 56(6), Sept. 2009.

[55] I. Roth, R. Kueng, S. Kimmel, Y.-K. Liu, D. Gross, J. Eisert, and M. Kliesch. Recovering quantum gates from few average gate fidelities. *Phys. Rev. Lett.*, 121:170502, Oct 2018.

[56] S. Shalev-Shwartz, O. Shamir, and S. Shammah. Failures of gradient-based deep learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3067–3075, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[57] J. Spencer. *Ten lectures on the probabilistic method*, volume 64 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 1994.

[58] T. Sugiyama, P. S. Turner, and M. Murao. Precision-guaranteed quantum tomography. *Phys. Rev. Lett.*, 111:160406, Oct 2013.

[59] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko, and G. Carleo. Neural-network quantum state tomography. *Nat. Phys.*, 14(5):447–450, 2018.

[60] Z. Webb. The clifford group forms a unitary 3-design. *Quantum Information & Computation*, 16(15-16):1379–1400, 2016.

[61] A. Wigderson and D. Xiao. Derandomizing the Ahlswede-Winter matrix-valued Chernoff bound using pessimistic estimators, and applications. *Theory Comput.*, 4:53–76, 2008.

[62] H. Zhu. Multiqubit Clifford groups are unitary 3-designs. *Phys. Rev. A*, 96:062336, Dec 2017.

Chapter 4

Incompressibility of generic quantum circuits

or: Models of quantum complexity growth

Abstract

The concept of quantum complexity has far-reaching implications spanning theoretical computer science, quantum many-body physics, and high-energy physics. The quantum complexity of a unitary transformation or quantum state is defined as the size of the shortest quantum computation that executes the unitary or prepares the state. It is reasonable to expect that the complexity of a quantum state governed by a chaotic many-body Hamiltonian grows linearly with time for a time that is exponential in the system size; however, because it is hard to rule out a shortcut that improves the efficiency of a computation, it is notoriously difficult to derive lower bounds on quantum complexity for particular unitaries or states without making additional assumptions. To go further, one may study more generic models of complexity growth. We provide a rigorous connection between complexity growth and unitary k-designs, ensembles that capture the randomness of the unitary group. This connection allows us to leverage existing results about design growth to draw conclusions about the growth of complexity. We prove that local random quantum circuits generate unitary transformations whose complexity grows linearly for a long time, mirroring the behavior one expects in chaotic quantum systems and verifying conjectures by Brown and Susskind. Moreover, our results apply under a strong definition of quantum complexity based on optimal distinguishing measurements.

Authors

Fernando G.S.L. Brandão, Wissam Chemissany, Nicholas Hunter-Jones, Richard Kueng, John Preskill.

Journal

PRX Quantum, **2**:030316 (2021) [editor's suggestion].

Confirmation of declaration of author contributions (Fernando G.S.L. Brandão)

Publication:

F.G.S.L. Brandão, W. Chemissany, N. Hunter-Jones, R. Kueng, J. Preskill, Models of quantum complexity growth, *PRX Quantum* **2**:030316 (2021) [editor's suggestion]

Declaration of author contributions:

Richard Kueng and Nicholas Hunter-Jones developed the theoretical aspects of this work. Fernando G.S.L. Brandão, Wissam Chemissany and John Preskill provided guidance and put the results into a broader context. All authors wrote the manuscript.

Confirmation by co-author:

I confirm this declaration of author contributions, as well as my co-authorship.

Fernando G.S.L. Brandão

Confirmation of declaration of author contributions (Wissam Chemissany)

Publication:

F.G.S.L. Brandão, W. Chemissany, N. Hunter-Jones, R. Kueng, J. Preskill, Models of quantum complexity growth, *PRX Quantum* **2**:030316 (2021) [editor's suggestion]

Declaration of author contributions:

Richard Kueng and Nicholas Hunter-Jones developed the theoretical aspects of this work. Fernando G.S.L. Brandão, Wissam Chemissany and John Preskill provided guidance and put the results into a broader context. All authors wrote the manuscript.

Confirmation by co-author:

I confirm this declaration of author contributions, as well as my co-authorship.

Wissam Chemissany

Wissam Chemissany

Confirmation of declaration of author contributions (Nicholas Hunter-Jones)

Publication:

F.G.S.L. Brandão, W. Chemissany, N. Hunter-Jones, R. Kueng, J. Preskill, Models of quantum complexity growth, *PRX Quantum* **2**:030316 (2021) [editor's suggestion]

Declaration of author contributions:

Richard Kueng and Nicholas Hunter-Jones developed the theoretical aspects of this work. Fernando G.S.L. Brandão, Wissam Chemissany and John Preskill provided guidance and put the results into a broader context. All authors wrote the manuscript.

Confirmation by co-author:

I confirm this declaration of author contributions, as well as my co-authorship.

Nicholas Hunter-Jones

Confirmation of declaration of author contributions (John Preskill)

Publication:

F.G.S.L. Brandão, W. Chemissany, N. Hunter-Jones, R. Kueng, J. Preskill, Models of quantum complexity growth, *PRX Quantum* **2**:030316 (2021) [editor's suggestion]

Declaration of author contributions:

Richard Kueng and Nicholas Hunter-Jones developed the theoretical aspects of this work. Fernando G.S.L. Brandão, Wissam Chemissany and John Preskill provided guidance and put the results into a broader context. All authors wrote the manuscript.

Confirmation by co-author:

I confirm this declaration of author contributions, as well as my co-authorship.

John Preskill

Editors' Suggestion

Models of Quantum Complexity Growth

Fernando G.S.L. Brandão,[1,2,3,4] Wissam Chemissany,[1] Nicholas Hunter-Jones,[1,5,*] Richard Kueng,[1,3,6,†] and John Preskill[1,2,3,4]

[1] *Institute for Quantum Information and Matter, Caltech, Pasadena, California, USA*
[2] *AWS Center for Quantum Computing, Pasadena, California, USA*
[3] *Department of Computing and Mathematical Sciences, Caltech, Pasadena, California, USA*
[4] *Walter Burke Institute for Theoretical Physics, Caltech, Pasadena, California, USA*
[5] *Perimeter Institute for Theoretical Physics, Waterloo, Ontario, Canada*
[6] *Institute for Integrated Circuits, Johannes Kepler University Linz, Austria*

(Received 13 January 2021; accepted 21 May 2021; published 29 July 2021)

The concept of quantum complexity has far-reaching implications spanning theoretical computer science, quantum many-body physics, and high-energy physics. The quantum complexity of a unitary transformation or quantum state is defined as the size of the shortest quantum computation that executes the unitary or prepares the state. It is reasonable to expect that the complexity of a quantum state governed by a chaotic many-body Hamiltonian grows linearly with time for a time that is exponential in the system size; however, because it is hard to rule out a shortcut that improves the efficiency of a computation, it is notoriously difficult to derive lower bounds on quantum complexity for particular unitaries or states without making additional assumptions. To go further, one may study more generic models of complexity growth. We provide a rigorous connection between complexity growth and unitary k-designs, ensembles that capture the randomness of the unitary group. This connection allows us to leverage existing results about design growth to draw conclusions about the growth of complexity. We prove that local random quantum circuits generate unitary transformations whose complexity grows linearly for a long time, mirroring the behavior one expects in chaotic quantum systems and verifying conjectures by Brown and Susskind. Moreover, our results apply under a strong definition of quantum complexity based on optimal distinguishing measurements.

DOI: 10.1103/PRXQuantum.2.030316

I. MOTIVATION AND OVERVIEW

The *complexity* of a computation is a measure of the resources needed to perform the computation. In a classical model of computation, the complexity of a Boolean function may be defined as the minimal number of elementary steps needed to evaluate the function. The precise number of steps needed depends on how the model is chosen, but this notion of complexity provides a useful way to quantify the hardness of a computational problem because how the number of steps scales with the size of the input to the problem has only weak dependence on the choice of model. By broad consensus, a computational task is considered to be feasible if its complexity grows no faster than a power of the input size, and intractable otherwise; using this criterion, all classical models of computation agree about which problems are (classically) "easy" and which ones are "hard."

Likewise, we may separate computational tasks into those that are easy or hard for a quantum computer. The circuit model of quantum computation provides a convenient way to quantify quantum complexity—namely, the quantum complexity of a Boolean function is the minimal size of a quantum circuit, which computes the function and outputs the right answer with high success probability. Here by the size of the circuit we mean the number of quantum gates in the circuit. These gates are chosen from a universal set of gates, where each gate in the set is a unitary transformation acting on a constant number of qubits or qudits. Though there are many ways to choose the universal gate set, any set of universal gates can simulate another accurately and efficiently, so that circuit size provides a useful model-independent measure of complexity.

*nickrhj@pitp.ca
†richard.kueng@jku.at

From a physicist's perspective, a quantum computation is a process governed by a local time-dependent Hamiltonian, and an intractable computation is a process that requires a time, which grows superpolynomially with the system size. Such intractable processes are not expected to be observed in nature.

Furthermore, in quantum physics, in contrast to classical digital computation, there is a meaningful notion of complexity not only for processes, but also for quantum states. Starting from a state in which all the bits are set to 0, any string of n classical bits can be prepared by flipping at most n bits. But the time needed to prepare a pure n-qubit quantum state, starting from a product state, even if we are permitted to use any time-dependent Hamiltonian, which is a sum of terms with constant weight and bounded norm, can be exponential in n. In fact, because the volume of the n-qubit Hilbert space is *doubly exponential* in n, while the number of quantum circuits with T gates is merely exponential in T, *most* n-qubit pure quantum states have exponentially large complexity. That is, for a typical pure state in the n-qubit Hilbert space, the time needed to prepare the state with some small constant error δ, starting from a product state, grows exponentially with n. Thus, nearly all quantum states of any macroscopic system will forever be far beyond the grasp of the quantum engineers [1].

While the complexity of quantum *circuits* has long been a foundational concept in quantum information theory [2], appreciation that quantum *state* complexity is an important concept has blossomed relatively recently. For example, the complexity of ground-state wave functions may be used to classify topological phases of matter at zero temperature [3]. Furthermore, a chaotic quantum Hamiltonian H can be usefully characterized by saying that evolution governed by H over a long time period generates highly complex states. A particularly intriguing proposal is that, in the context of the anti-de Sitter/conformal field theory (AdS/CFT) correspondence, the complexity of a quantum state of the boundary theory corresponds to the volume in the bulk geometry, which is hidden behind the event horizon of a black hole [4–7].

When we say a quantum state is highly complex, we mean there is no easy way to prepare the state, but how can we be sure? Perhaps we were not clever enough to think of an ingenious shortcut that prepares the state efficiently. It is not possible in practice to enumerate all the quantum circuits that approximate a specified state to find one of minimal size. For that reason, it is quite difficult to obtain a useful lower bound on the complexity of the quantum state prepared by a specified many-body Hamiltonian in a specified time. It is reasonable to expect that, for a chaotic Hamiltonian H and an unentangled initial state, the complexity grows linearly in time for an exponentially long time, but we do not have the tools to prove it from first principles for any particular H.

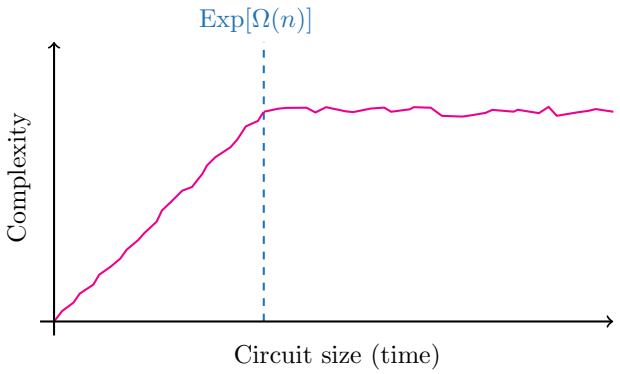


FIG. 1. *Expected complexity growth in random circuits.* Conjecture 1 states that, for random quantum circuits acting on n qubits, the circuit complexity grows linearly with circuit size (time) until it saturates at a value exponentially large in n. Our work provides rigorous evidence supporting this picture for quantum systems with sufficiently large local dimension; see Corollary 5.

One possible approach is to rely on highly plausible complexity theory assumptions to derive nontrivial conclusions about the complexity of states generated by particular circuits or Hamiltonians [8–10]. Another is to consider ensembles of circuits, and to derive lower bounds on complexity, which hold with high probability when samples are selected from these ensembles. We follow the latter approach here, drawing inspiration from recent work by Susskind [8] and Brown and Susskind [7]. These authors state a conjecture about the complexity growth of geometrically local random quantum circuits (see Fig. 1).

Conjecture 1 (Brown and Susskind [7]; Susskind [8]): *Most local random circuits of size T have a complexity that scales linearly in T for an exponentially long time.*

Our goal is to strengthen the evidence supporting this conjecture.

Brown and Susskind provided evidence for this scaling law by means of a simple counting argument; see also Ref. [11]. For a fixed finite set of universal quantum gates, consider the ensemble of all circuits with size T. By definition, this ensemble accurately approximates (to within a specified error δ) all unitary transformations with complexity T or less. Furthermore, the number of circuits increases exponentially with T, and, because the unitary group has a very large volume, it seems reasonable to assume that "collisions" between circuits are rare unless T is very large; that is, that the number of distinct unitary transformations realized by this ensemble (where "distinct" means more than distance δ apart) is comparable to the number of circuits. This means that the number of circuits with size T' is too small to account for more than a small fraction of the unitary transformations realized by

circuits of size T if T' is much smaller than T. In other words, most random circuits with size T have complexity at least T', where T' is comparable to T.

This argument hinges on a crucial assumption, which sounds plausible but is hard to prove: *collisions between circuits of subexponential size are rare*. Collisions certainly occur for any circuit size T, and necessarily become common for circuits of exponential size, where T is comparable to the Hilbert-space dimension so that the exponential of T is comparable to the Hilbert-space volume. Thus an analytic treatment of complexity growth seems like a daunting combinatorial task.

The work [12] provides some rigorous support for Conjecture 1. There, the authors show that local random circuits can "fool" short-measurement procedures. That is, a typical quantum state prepared by a local random circuit of size polynomial in n, acting on an initial product state, cannot be distinguished from a maximally mixed state by any procedure that is much simpler than running the circuit backwards and verifying that the initial product state is recovered. Although not stated in this fashion, the results from Ref. [12] imply that, with high probability, a local random circuit of size T has complexity $\Omega(T^{1/11})$. While this argument rigorously proves a weakened version of Conjecture 1, there are still issues we wish to address:

(i) *Restricted notion of complexity:* The authors implicitly define complexity as the capability of fooling short-measurement protocols. While this operational notion of complexity is well motivated, the actual measurement procedures considered are quite restrictive. In particular, they do not take into account ancilla-assisted measurements—a mainstay of modern quantum information.

(ii) *Collisions are not treated explicitly:* The ensemble of local random circuits of size T defines a probability distribution on the n-qubit unitaries. If we are only interested in specifying unitary transformations up to some specified error δ, collisions occur, so that some unitaries are more likely than others. The arguments in Ref. [12] show that the unitaries sampled from this distribution typically have complexity $\Omega(T^{1/11})$, but do not rule out the possibility that the distribution is highly nonuniform. It is at least a logical possibility, compatible with the findings of Ref. [12], that the ensemble contains only a small number of unitaries, which have high complexity, all of which occur with relatively high probability. To conclude that the ensemble contains many high-complexity unitaries, we need to know more about the properties of the probability distribution governing the ensemble.

(iii) *Polynomial relation between circuit size and complexity:* The relation between circuit size T and expected minimal complexity $T^{1/11}$ is polynomial, not (yet) linear as required by Conjecture 1.

In this work we make progress toward a rigorous proof of Conjecture 1 by developing a general framework that addresses some of the shortcomings of the previously known rigorous evidence in favor of the conjecture [12]. In particular, we define and use a *strong* notion of complexity, which captures the difficulty of distinguishing a given circuit from the most useless possible quantum channel: the completely depolarizing channel $\mathcal{D}(\rho) = [\mathrm{Tr}(\rho)/d]\mathbb{I}$ that maps any state to the maximally mixed state.

Definition 1 (Strong complexity: Informal definition): *The complexity of a quantum circuit U is the minimal circuit size required to implement an ancilla-assisted measurement that is capable of distinguishing $\rho \mapsto U\rho U^\dagger$ from the completely depolarizing channel $\rho \mapsto (1/d)\mathbb{I}$.*

We refer to Sec. II A for a more detailed motivation and a precise statement of this definition. For now, we emphasize that this strong definition of complexity implies other (weaker) definitions, such as the minimal circuit size required to approximate U.

Our first main contribution is a rigorous connection between complexity growth and the notion of *approximate unitary k-designs* [13,14]. We use the notation $\{p_i, U_i\}$ for an ensemble of unitary transformations in which the unitary U_i occurs with probability p_i. A unitary k-design is an ensemble with strong pseudorandom properties; an approximate k-design accurately approximates the first k-moments of the Haar measure on the unitary group. Hence a k-design with large k behaves essentially like a Haar-random ensemble of unitaries, while a small-k-design can be highly structured. For instance, the n-qubit Pauli group forms a 1-design, while the n-qubit Clifford group is a 3-design [15–17]. The design order k allows us to interpolate between these two very different regimes. Intuitively, we would expect that the complexity of a k-design grows with k. Our first technical contribution makes this intuition precise: a linear growth in design implies a linear growth in (strong) complexity.

Theorem 2 (Informal statement): *Let $\{p_i, U_i\}$ be an approximate unitary k-design. Then, a randomly selected (according to the weights) element is very likely to have strong circuit complexity approximately equal to k.*

We refer to Theorem 9 for a more detailed, quantitative statement. This result strengthens the assertions in Ref. [12] by allowing ancilla-assisted measurement procedures. To do so we prove novel bounds on Haar moments, see Sec. II D for details. Our second technical contribution

shows that the k-design property alone severely limits the likelihood of collisions.

Lemma 3: *Let $\{p_i, U_i\}$ be an approximate k-design. Then, the associated weight distribution cannot be too spiky:* $\max_i p_i \lesssim k! d^{-2k}$.

This result formalizes the intuitive idea that giving unusually high weight to some unitaries cannot be compatible with the k-design property, but we are not aware of any precise statements along these lines in the existing literature. Importantly, because Lemma 3 establishes that the distribution is nearly flat, knowing that sampling from a k-design yields a high-complexity unitary with high probability (as stated in Theorem 2) allows us to infer that there must be many distinct high-complexity unitaries in the ensemble. Here our reasoning is based on an approximate version of Laplace's definition of probability: if events are assigned nearly equal probabilities, then the probability of property X is approximately the number of events with property X divided by the total number of events. Together, Theorem 2 and Lemma 3 imply the following corollary.

Corollary 4: *Any approximate k-design contains exponentially many (in k) unitaries that have circuit complexity* $\Omega(k)$.

While Corollary 4 does not by itself strongly constrain how these high-complexity unitary transformations are distributed geometrically within the n-qudit unitary group, we are also able to prove a stronger result: *An approximate k-design contains exponentially many (in k) high-complexity unitaries whose pairwise distance* (i.e., the distance between any pair of unitaries) *is almost maximal in the diamond norm.* This stronger statement rules out the possibility that most of the high-complexity unitaries reside inside a few tightly packed clusters within $U(d)$.

Approximate unitary k-designs are a central concept in quantum information, where their pseudorandom properties have found extensive application across subfields, e.g., state distinguishability [18], decoupling [19], state tomography [20,21], randomized benchmarking [22], equilibration [12] (and references therein), information scrambling [11,23], and many more. As a result, several probabilistic constructions are known. Applying Corollary 4 to any of these constructions establishes a rigorous model for quantum complexity growth. In particular, the following.

(a) *Local random quantum circuits with polynomial design growth:* Ref. [12] proves that the set of all geometrically local circuits of size $T = O(n^2 k^{11})$ forms an approximate unitary k-design [24]. Corollary 4 therefore implies that local circuits of size T contain at least $\exp[\Omega(T^{1/11})]$ elements with strong complexity $\Omega(T^{1/11})$.

(b) *Stochastic quantum Hamiltonians with polynomial design growth:* One can study the growth of complexity in continuous-time models of chaotic dynamics, rather than the discrete-time dynamics embodied by random circuits [25–27]. Stochastic Hamiltonian dynamics, in which a local Hamiltonian fluctuates as a function of time, has been shown to realize approximate k-designs [26] with a relationship between k and the evolution time similar to what was established in Ref. [12] for local random circuits. Further progress achieved in Ref. [27] shows that, for a particular class of stochastic Hamiltonians, evolution time linear in k suffices to generate approximate k-designs for $k = o(\sqrt{n})$. Corollary 4 therefore implies that with high probability the complexity grows linearly in time, at least for a while.

(c) *Local random circuits with linear design growth:* Recently, the results of Ref. [12] were improved using an exact mapping from random circuits to the statistical mechanics of a lattice model [28], showing that local circuits of size $T = O(n^2 k)$ form approximate k-designs in the limit of large local dimension (Hilbert space dimension $d = q^n$ with q large). The q dependence was subsequently improved in Ref. [29] by studying the spectral gap of the moment operator for random quantum circuits. Combined with Corollary 4 this establishes a *linear* relation between circuit size and complexity. Thus we can prove the following statement analogous to Conjecture 1.

Corollary 5: *The set of all local circuits of size T contains at least* $\exp[\Omega(T)]$ *elements with strong complexity* $\Omega(T)$*, provided that the local dimension is sufficiently large:* $q \geq \Omega(k^2)$.

More precise statements of our main results, and a more detailed comparison to previous work, can be found in Sec. II.

II. QUANTUM COMPLEXITY AND UNITARY DESIGNS

A. Operational definitions of complexity

1. State complexity

We consider systems comprised of n qudits with local dimension q: $d = q^n$. Existing works on complexity typically start with identifying a class of states that are *useful* starting states for quantum computations. In this work we take a reverse approach and start with identifying a *useless*

state. The maximally mixed state

$$\rho_0 = \frac{\mathbb{I}}{d}, \tag{1}$$

is unique in the sense that it is invariant under arbitrary unitary evolutions, including any quantum circuit. Intuitively, useful starting states should be as far away from this useless state as possible. If we use trace distance, this intuition is true to some extent. Any pure state $|\psi\rangle\langle\psi|$ obeys

$$\frac{1}{2}\,\||\psi\rangle\langle\psi| - \rho_0\|_1 = 1 - \frac{1}{d}. \tag{2}$$

But this is clearly too coarse for distinguishing the usefulness of different pure states. In order to achieve such a task, we recall the operational interpretation of the trace distance. It corresponds the optimal bias achievable in distinguishing the state $|\psi\rangle\langle\psi|$ from ρ_0 using a single measurement [30,31]. We refer to Appendix B 1b for a more detailed exposition. The optimal measurement achieving this bias is $M = |\psi\rangle\langle\psi|$ and *does* depend on the state in question. Such a measurement may be challenging to implement for states that we would intuitively assign a high complexity to (such as random states) and very easy for states that we consider useful (such as computational basis states). We can interpolate between these extreme cases by limiting the resources available to implement distinguishing measurements. Let $\mathbb{H}_d$ denote the space of $d \times d$ Hermitian matrices. For fixed $r \in \mathbb{N}$, we consider the class of measurements $\mathsf{M}_r(d) \subset \mathbb{H}_d$ that can be implemented by combining (at most) r 2-local gates from a fixed, universal gate set $\mathsf{G} \subset U(4)$. We refer to Appendix B 2 for further details and justification. The maximal bias achievable for quantum states (QS) with such a restricted set of measurements is the solution to the following optimization problem:

$$\begin{aligned}\beta^{\sharp}_{\mathrm{QS}}(r, |\psi\rangle) = &\text{maximize} \quad |\mathrm{Tr}\left[M\left(|\psi\rangle\langle\psi| - \rho_0\right)\right]| \\ &\text{subject to} \quad M \in \mathsf{M}_r(d).\end{aligned} \tag{3}$$

We may decompose the true optimal measurement as $|\psi\rangle\langle\psi| = U|0\rangle\langle 0|U^\dagger$ for some $U \in U(d)$. The unitary U may be approximated to arbitrary precision by 2-local circuits chosen from a universal gate set [32]. This ensures

$$\beta^{\sharp}_{\mathrm{QS}}(r, |\psi\rangle) \to \frac{1}{2}\||\psi\rangle\langle\psi| - \rho_0\|_1 = 1 - \frac{1}{d} \quad \text{as } r \to \infty. \tag{4}$$

For simple states, like computational basis states, this convergence happens rapidly, while generic states require exponentially large circuit sizes. This observation is the motivation for the following definition of complexity.

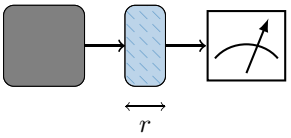


FIG. 2. *Pictographic illustration of strong state complexity (Definition 2).* A blackbox either outputs a (known) pure state $\rho = |\psi\rangle\langle\psi|$, or the maximally mixed state $\rho_0 = (1/d)\mathbb{I}$. The task is to correctly guess which one it produced by applying a preprocessing circuit V (blue line pattern) of limited size r and performing a simple measurement (right). We say that $|\psi\rangle$ has *strong state complexity* less than r if the probability of correctly distinguishing both possibilities is close to optimal.

Definition 2 (Strong state complexity): *Fix $r \in \mathbb{N}$ and $\delta \in (0, 1)$. We say that a pure state $|\psi\rangle$ has strong δ-state complexity at most r if and only if*

$$\beta^{\sharp}_{QS}(r, |\psi\rangle) \geq 1 - \frac{1}{d} - \delta, \tag{5}$$

which we denote as $\mathcal{C}_\delta(|\psi\rangle) \leq r$.

This definition has a ready operational interpretation that is illustrated in Fig. 2. The following result relates it to more traditional definitions.

Lemma 6: *Suppose that $|\psi\rangle \in \mathbb{C}^d$ obeys $\mathcal{C}_\delta(|\psi\rangle) \geq r + 1$ for some $\delta \in (0, 1)$ and $r \in \mathbb{N}$. Then,*

$$\min_{size(V)\leq r} \frac{1}{2}\left\||\psi\rangle\langle\psi| - V|0\rangle\langle 0|V^\dagger\right\|_1 > \sqrt{\delta}, \tag{6}$$

i.e., it is impossible to accurately produce $|\psi\rangle$ with fewer than r elementary gates.

The converse is false in general. To see this, select a generic state $|\widetilde{\psi}\rangle$ on $(n-1)$ qudits and set $|\psi\rangle = |0\rangle \otimes |\widetilde{\psi}\rangle$. Then, the quantity in Eq. (6) is dominated by the (traditional) complexity of $|\widetilde{\psi}\rangle$, which may be very high. Nonetheless, the simple distinguishing measurement $M = |0\rangle\langle 0| \otimes \mathbb{I}$ (ignore everything but the first qudit) achieves

$$\begin{aligned}\mathrm{Tr}\left[M(|\psi\rangle\langle\psi| - \rho_0)\right] &= \mathrm{Tr}\left[|0\rangle\langle 0|\left(|0\rangle\langle 0| - \frac{1}{q}\mathbb{I}\right)\right] \\ &= 1 - \frac{1}{q},\end{aligned} \tag{7}$$

which is high, especially for large local dimension q. This example highlights that Definition 2 is indeed more stringent than traditional definitions of state complexity.

Proof of Lemma 6. By contraposition. Let $\mathsf{G}_r \subset U(d)$ denote the class of unitary circuits that are comprised of at most r 2-local gates chosen from a universal gate

set G. Suppose there exists a size-r circuit $V \in \mathsf{G}_r$ such that $\frac{1}{2}\left\||\psi\rangle\langle\psi| - V|0\rangle\langle 0|V^\dagger\right\|_1 \leq \sqrt{\delta}$. The state difference in question has rank two, which allows for explicitly computing the trace distance: $\frac{1}{2}\||\psi\rangle\langle\psi| - V|0\rangle\langle 0|V^\dagger\|_1 = \sqrt{1-|\langle 0|V^\dagger|\psi\rangle|^2}$. The assumption is therefore equivalent to $|\langle 0|V^\dagger|\psi\rangle|^2 \geq 1-\delta$ and we conclude

$$\begin{aligned}\beta^\sharp_{\mathrm{QS}}(r,|\psi\rangle) &\geq \mathrm{Tr}\left[V|0\rangle\langle 0|V^\dagger(|\psi\rangle\langle\psi| - \rho_0)\right] \\ &= |\langle 0|V^\dagger|\psi\rangle|^2 - \frac{1}{d} \geq 1 - \frac{1}{d} - \delta, \end{aligned} \tag{8}$$

because $V|0\rangle\langle 0|V^\dagger \in \mathsf{M}_r$. This in turn implies $\mathcal{C}_\delta(|\psi\rangle) \leq r$ and the claim follows. ■

2. Unitary complexity

We define the complexity of unitary channels $\mathcal{U}(\rho) = U\rho U^\dagger$ in a fashion similar to state complexity. We start with identifying the completely depolarizing channel as the most *useless* channel conceivable:

$$\mathcal{D}(\rho) = \rho_0 = \frac{\mathbb{I}}{d} \quad \text{for all states } \rho. \tag{9}$$

The *diamond distance* between $\mathcal{D}$ and any unitary channel is close to maximal:

$$\frac{1}{2}\|\mathcal{U} - \mathcal{D}\|_\diamond = 1 - \frac{1}{d^2}. \tag{10}$$

As detailed in Appendix B 1c, the diamond distance also has an appealing operational definition [33]. It corresponds to the maximal bias achievable for the task of distinguishing $\mathcal{U}$ from $\mathcal{D}$ with a single channel use. The optimal strategy may involve a quantum memory. Choose a state in the doubled Hilbert space $|\phi\rangle\langle\phi|$, with $|\phi\rangle \in \mathbb{C}^d \otimes \mathbb{C}^d$ and input one half into the unknown channel, while the other half remains unchanged in the quantum memory. Subsequently, perform a two-outcome measurement on the output state to distinguish both channels.

An optimal strategy for distinguishing $\mathcal{U}$ from $\mathcal{D}$ corresponds to choosing a maximally entangled (Bell) state $|\Omega\rangle \in \mathbb{C}^d \otimes \mathbb{C}^d$ as input and measuring $M = (U \otimes \mathbb{I})|\Omega\rangle\langle\Omega|(U^\dagger \otimes \mathbb{I})$. Equivalently, choose $(U^\dagger \otimes \mathbb{I})|\Omega\rangle$ as input and measure $M = |\Omega\rangle\langle\Omega|$ on the output. Similar to the state complexity argument, the optimal input state, or the optimal outcome measurement (or both) depend on the unitary $U \in U(d)$ describing the channel $\mathcal{U}$. This may be challenging to implement, especially if U corresponds to a complicated circuit. We restrict apparatus power by bounding the total circuit sizes that are allowed to implement such a measurement procedure. Let $\mathsf{G}_{r'} \subset U(d^2)$ be the set of all unitary circuits on $2n$ qudits (register+memory) that are comprised of at most r' elementary gates. Likewise, let $\mathsf{M}_{r''} \subset \mathbb{H}_d^{\otimes 2}$ denote the class of all two-outcome measurements on $2n$ qudits that require circuit size at most r'' to implement. The optimal bias for quantum channels (QC) achievable under such restrictions is

$$\begin{aligned}\beta^\sharp_{\mathrm{QC}}(r,U) = \text{maximize} \quad & \big|\mathrm{Tr}\big\{M\big[(\mathcal{U}\otimes\mathcal{I})(|\phi\rangle\langle\phi|) \\ & - (\mathcal{D}\otimes\mathcal{I})(|\phi\rangle\langle\phi|)\big]\big\}\big| \\ \text{subject to} \quad & M \in \mathsf{M}_{r'},\ |\phi\rangle = V|0\rangle, \\ & V \in \mathsf{G}_{r''},\ r = r' + r'', \end{aligned} \tag{11}$$

where the identity channel $\mathcal{I} : \mathbb{H}_d \to \mathbb{H}_d$ indicates that the memory is left unchanged. As r increases, more complicated measurements and state preparations become possible. At some point this will include ever more precise approximations of the optimal measurement [32]:

$$\beta^\sharp_{\mathrm{QC}}(r,U) \longrightarrow \frac{1}{2}\|\mathcal{U} - \mathcal{D}\|_\diamond = 1 - \frac{1}{d^2} \quad \text{as } r \to +\infty. \tag{12}$$

Similar to the state case, the rate of convergence does depend on the complexity of the unknown unitary U. This is the basis for our operational definition of unitary complexity.

Definition 3 (Strong unitary complexity): *Fix $r \in \mathbb{N}$ and $\delta \in (0,1)$. We say that a unitary $U \in U(d)$ has strong δ-unitary complexity at most r if and only if*

$$\beta^\sharp_{QC}(r,U) \geq 1 - \frac{1}{d^2} - \delta, \tag{13}$$

which we denote as $\mathcal{C}_\delta(U) \leq r$.

The operational motivation for this definition is sketched in Fig. 3. Strong unitary complexity (Definition 3) is more stringent than traditional definitions that use approximation errors in some norm. But the comparison between the two is not quite as straightforward as in the state complexity case. This is because, the optimal strategy for distinguishing $\mathcal{U}$ from $\mathcal{D}$ involves a maximally entangled (Bell) input state $|\Omega\rangle\langle\Omega|$, as well as a corresponding two-outcome measurement. In the following statement, we explicitly allow such input states and measurements in the distinguishability protocol. Although mild—relatively short circuits allow for transforming computational basis states into Bell states [34]—this assumption does further increase the power of the measurements we are allowed to make. Our main technical results, most notably Theorem 9, do take this into account and apply to this slightly stronger notion of strong unitary complexity.

Lemma 7: *Consider a setup that contains maximally entangled inputs and measurements and suppose that*

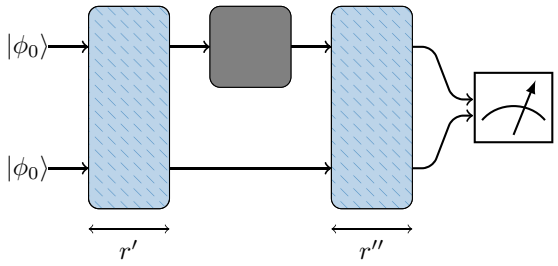


FIG. 3. *Pictographic illustration of strong unitary complexity (Definition 3).* A blackbox (center) takes quantum states as inputs and applies either a unitary channel $\mathcal{U}(\rho) = U\rho U^\dagger$, or the depolarizing channel $\mathcal{D}(\rho) = \rho_0 = \mathbb{I}/d$. The task is to correctly guess which evolution occurred. The rules of the game allow short pre- and postprocessing circuits (blue line patterns) that may involve a quantum memory. The final guess must be based on a simple measurement (right). We say that U has complexity less than $r = r' + r''$ if the probability of correctly distinguishing both options is close to optimal.

$U \in U(d)$ obeys $\mathcal{C}_\delta(U) \geq r + 1$ for some $\delta \in (0, 1)$, $r \in \mathbb{N}$. Then,

$$\min_{size(V)\leq r} \frac{1}{2} \|\mathcal{U} - \mathcal{V}\|_\diamond > \sqrt{\delta}, \tag{14}$$

i.e., it is impossible to accurately approximate U by circuits comprised of fewer than r elementary gates.

Again, the converse relation is false in general.

Proof of Lemma 7. By contraposition. Assume there exists $V \in U(d)$ with $\text{size}(V) \leq r$ such that $\frac{1}{2}\|\mathcal{U} - \mathcal{V}\|_\diamond \leq \sqrt{\delta}$. Then,

$$\begin{aligned}\sqrt{\delta} &\geq \frac{1}{2}\|\mathcal{U} - \mathcal{V}\|_\diamond \geq \frac{1}{2}\left\|(U\otimes\mathbb{I})|\Omega\rangle\langle\Omega|(U^\dagger\otimes\mathbb{I})\right. \\ &\quad \left. - (V\otimes\mathbb{I})|\Omega\rangle\langle\Omega|(V^\dagger\otimes\mathbb{I})\right\|_1 \\ &= \sqrt{1 - |\langle\Omega|V^\dagger U\otimes\mathbb{I}|\Omega\rangle|^2}, \end{aligned} \tag{15}$$

as the second expression involves a trace distance of two pure states, which can be computed explicitly. Next, note that $M = (V\otimes\mathbb{I})|\Omega\rangle\langle\Omega|(V^\dagger\otimes\mathbb{I})$ is a legitimate distinguishing measurement, because $\text{size}(V) \leq r$ and we explicitly include the Bell measurement. Likewise, the input state $|\Omega\rangle\langle\Omega|$ is also allowed and produces a maximally mixed state when completely depolarized: $\mathcal{D}\otimes\mathcal{I}(|\Omega\rangle\langle\Omega|) = \rho_0^{\otimes 2}$ (this is why we need Bell states) ensures

$$\begin{aligned}\beta^{\sharp}_{\text{QC}}(r, U) &\geq \text{Tr}\Big\{(V\otimes\mathbb{I})|\Omega\rangle\langle\Omega|(V^\dagger\otimes\mathbb{I}) \\ &\quad \times \left[(U\otimes\mathbb{I})|\Omega\rangle\langle\Omega|(U^\dagger\otimes\mathbb{I}) - \rho_0^{\otimes 2}\right]\Big\} \\ &= \left|\langle\Omega|V^\dagger U\otimes\mathbb{I}|\Omega\rangle\right|^2 - \langle\Omega|V^\dagger\rho_0 V\otimes\rho_0|\Omega\rangle \\ &\geq 1 - \delta^2 - \frac{1}{d^2}. \end{aligned} \tag{16}$$

■

B. Approximate unitary designs

The concept of *unitary k-designs* [13,14] provides an interpolation between two extreme cases: (i) small collections of highly structured unitaries that form the basic building blocks of quantum-computing devices (e.g., local Pauli gates, or elements of the Clifford group). (ii) generic (Haar random) unitaries that lack any sort of structure and require circuits of exponential size to approximate.

Roughly speaking, an ensemble $\mathcal{E} = \{p_i, U_i\}$ of unitaries is a unitary k-design if it exactly reproduces the first k moments of the Haar measure over the unitary group. More precisely, given the twirling channels $\mathcal{T}^{(k)}_U(X) = \int dU U^{\otimes k} X (U^\dagger)^{\otimes k}$ and $\mathcal{T}^{(k)}_\mathcal{E}(X) = \sum_i p_i U_i^{\otimes k} X (U_i^\dagger)^{\otimes k}$, an ensemble $\mathcal{E}$ is a unitary design with order k if

$$\mathcal{T}^{(k)}_\mathcal{E}(X) = \mathcal{T}^{(k)}_U(X), \tag{17}$$

for all X in the k-fold tensor product. Although seemingly abstract, this notion captures important physical concepts. 1-designs are in one-to-one correspondence with unitary operator frames, while 2-designs sufficiently capture the notion of *scrambling* [11,23].

Unitary k-designs are known to exist for any dimension d and any order k. Nevertheless, explicit constructions are notoriously difficult to find. This challenge can be overcome by relaxing the notion of a k-design. Indeed, for most applications it is sufficient to ensure that Eq. (17) is only approximately true, see Definition 4 in the Appendix for a precise statement. Several conventions for choosing an appropriate distance measure $\|\cdot\|$ have been put forth, but here we opt for the diamond distance $\|\cdot\|_\diamond$, which quantifies the distinguishability of two ensembles.

In contrast to exact k-designs, several explicit constructions for approximate k-designs have been established [12, 26–28,35,36], including local random circuits and various Brownian circuits and stochastic quantum Hamiltonians. These constructions allow us to relate abstract insights about complexity growth in designs to concrete random circuit models.

C. Complexity by design

This section presents our main technical contributions.

1. State complexity growth

Theorem 8: *Consider the set of (pure) states in $d = q^n$ dimensions that results from applying all unitaries associated with an ϵ-approximate $2k$-design to a fixed (but*

arbitrary) starting state $|\psi_0\rangle$. *Then, this set contains at least*

$$\binom{d+k-1}{k}\left[\frac{1}{1+\epsilon}-2d(n+1)^r|\mathsf{G}|^r\left(\frac{16k^2}{d(1-\delta)^2}\right)^k\right],$$

distinct states that obey $\mathcal{C}_\delta(|\psi\rangle)\geq r+1$ *each. Qualitatively, this number is of order* $(d/k)^k$ *as long as r obeys*

$$r\lesssim\frac{k[n-2\log(k)]}{\log(n)}.$$

Because of collisions, the emphasis on distinct is justified; two or more distinct unitaries can lead to the same final state.

2. Unitary complexity growth

Theorem 9: *A discrete approximate* $2k$*-design in* $d=q^n$ *dimensions contains at least*

$$\frac{d^{2k}}{k!}\left[\frac{1}{1+\epsilon}-3d^2n^{2r}|\mathsf{G}|^r\left(\frac{1024k^4}{d(1-\delta)^2}\right)^k\right],$$

distinct unitaries that obey $\mathcal{C}_\delta(U)\geq r+1$ *each. Qualitatively, this number is of order* $(d^2/k)^k$ *as long as r obeys*

$$r\lesssim\frac{k[n-4\log(k)]}{\log(n)}.$$

D. Moment bounds

Both Theorems 8 and 9 follow from an initial probabilistic statement combined with relatively straightforward counting arguments. These probabilistic statements highlight that it is very unlikely to distinguish random k-design elements from their average with a fixed measurement procedure. Markov's inequality—$\Pr[S\geq\tau]=\Pr\left[S^k\geq\tau^k\right]\leq\mathbb{E}[S^k]/\tau^k$ for non-negative random variables S—reduces this probabilistic assertion to a question about moment growth. The larger the moments we can control, the stronger this assertion becomes. Designs appropriately capture this feature: a k-design accurately approximates Haar-random moments up to order k. This is why designs with growing k become increasingly complex.

For state complexity, the associated Haar-moment computation is relatively straightforward:

$$\mathbb{E}_{|\psi\rangle}\left(\left\{\mathrm{Tr}(M|\psi\rangle\langle\psi|)-\mathbb{E}_{|\psi\rangle}\left[\mathrm{Tr}(M|\psi\rangle\langle\psi|)\right]\right\}^k\right)\leq\left(\frac{k^2}{d}\right)^{k/2},\tag{18}$$

for any fixed measurement M, see e.g., Corollary 24 below.

However, such simple moments do not cover strong unitary complexity. Quantum channels allow for more sophisticated measurement procedures that render the associated Haar-moment computations nontrivial. Our main technical contribution is a novel bound that addresses this setting.

Theorem 10: *Fix a bipartite input state* $|\phi\rangle\in\mathbb{C}^d\otimes\mathbb{C}^d$ *and measurement* M *of compatible dimension. For* U *chosen uniformly from the Haar measure, we have*

$$\mathbb{E}_U\left[\left(Tr\left[M\left(U\otimes\mathbb{I}\right)|\phi\rangle\langle\phi|\left(U^\dagger\otimes\mathbb{I}\right)\right]-\mathbb{E}_U\left\{Tr\left[M\left(U\otimes\mathbb{I}\right)|\phi\rangle\langle\phi|\left(U^\dagger\otimes\mathbb{I}\right)\right]\right\}\right)^k\right]\leq\frac{C_k(k!)^2}{d^{k/2}},$$

where $C_k=[1/(k+1)]\binom{2k}{k}<4^k/k$ *denotes the* k*th Catalan number.*

This bound is considerably more general than existing ones in the literature. Reference [12], for instance, utilizes Eq. (18) only. We establish this result by combining Schur-Weyl duality [37,38] with Weingarten calculus [39,40] and auxiliary arguments from tensor network theory [41,42] and convex optimization [43,44]. We believe that the dimensional scaling in the final bound is tight, but there may be room for further improving the k-dependent constants. In particular, we do not know if the Catalan number is necessary, or merely an artifact of our proof technique.

E. Relation to previous work

Quantum complexity has recently become a popular subject in high-energy physics. A considerable amount of attention has been devoted to understanding the complexity accumulated after an exponentially long time. Works by Susskind and Aaronson [4,8,9] point to an intriguing connection to theoretical computer science: unless PSPACE $\subseteq$ BQP/poly (a hypothetical relation between different computational complexity classes that is widely believed to be false), the circuit complexity of certain Hamiltonian evolutions $U=\exp(-iHt)$ achieves superpolynomial values for exponentially long time scales t. In a similar vein, Bohdanowicz and Brandão [10] constructed a family of Hamiltonians that provably achieves superpolynomial complexity in exponential time, unless PSPACE = BQP.

These arguments address late-time complexity and therefore do not yield insights regarding early-time complexity growth. In this regard, relations between complexity growth and approximate k-designs have recently been pointed out in Refs. [11,45]. Specifically, Ref. [11] defined a notion of the complexity of generating an ensemble of unitaries and gave a lower bound on the ensemble complexity in terms of the distance to forming a unitary design. They also argued that the lower bound of the complexity of a k-design is linear in k. Our arguments and results may be regarded as a substantial refinement of these ideas.

The notion of strong complexity put forward in our work has its conceptual roots in quantum information. Encompassing this mindset is the statement from Ref. [46]: "most states are too entangled to be useful as computational resources." At the core of this argument is the following observation. Haar-random pure states are so highly entangled that *local* measurements yield almost uniformly random outcomes. In turn, any quantum device that relies on local measurements and uses known, but Haar-random, states could be efficiently simulated by tossing classical coins! This prevents any genuine quantum advantage for computation.

Strong state complexity (Definition 2) may be thought as a formal version of this observation. Measuring the maximally mixed state ρ_0 always results in a uniform outcome distribution. Moreover, Ref. [46] makes essential use of the fact that the measurements are constrained to be "simple" (in their case: local measurements augmented by classical postprocessing). The core of their argument may be summarized as follows: low complexity measurements do not allow for distinguishing a Haar-random state from the maximally mixed state. We present a variant of this argument in Appendix A 1 below.

While Ref. [46] considers only Haar-random pure states, similar arguments have been established for states that are less generic, see e.g., Ref. [12, Section 3]. Although not stated at this level of generality, Ref. [12, Corollary 10] effectively points out that states generated by approximate k-designs fool short quantum circuits: with high probability they cannot be distinguished from the maximally mixed state by means of any measurement with small circuit size. They also extend this result to circuits [12, Corollary 11]. With high probability, a randomly selected (according to the weights) k-design element cannot be approximated by any short-sized circuit V in the sense that $\|U - V\|_\infty$ is small.

The second main result of our work, Theorem 9, improves upon this result in two ways. Firstly, the strong unitary complexity (Definition 3) is more stringent than their more traditional definition. While Theorem 9 does imply [12, Corollary 11], the converse is not necessarily true.

Secondly, and more importantly, both Corollaries 10 and 11 in Ref. [12] are probabilistic. While this is enough to deduce average-case behavior, a strong quantitative statement about the number of k-design elements with high circuit complexity remains beyond the scope of these assertions. A worst-case caricature may help to illustrate this subtlety. Suppose that the weights accompanying a unitary k-design are extremely spiky. A single high-complexity unitary, say $U_1 \in U(d)$ is accompanied by an exceedingly large weight $p_1 \simeq 1$, while all other design unitaries U_i have low complexity and almost vanishing weights $p_i \simeq 0$. Such a weight distribution would not contradict the assertion of Ref. [12, Corollary 11]. The single high-complexity circuit occurs with high probability (over the weights). Nonetheless, the hypothetical k-design contains only a single high-complexity element.

Here we overcome this issue by explicitly ruling out the possibility of such extreme cases ever occurring. The definition of an approximate k-design alone implies that the weights cannot be too spiky, see Lemma 3. This bound on the weights allows us to convert probabilistic (average case) statements into quantitative ones. Not only does the average circuit complexity grow linearly with the order k of an approximate design, the absolute number of distinct circuits that have high complexity must also grow *exponentially* with k.

Interest in state complexity has been stimulated by its potential role in quantum gravity and the AdS/CFT correspondence; see Sec. IV for further discussion. Recently, the relevance to holographic duality of *computational pseudorandomness* has been emphasized. Specifically, the authors of Ref. [47] argue that one can construct two *mixed* quantum states on the boundary (A and B) such that both A and B can be efficiently prepared, yet A and B cannot be distinguished from maximally mixed states by polynomial-size quantum circuits. Furthermore, the corresponding bulk states (A' and B') *can* be distinguished efficiently from one another. This observation indicates that the holographic dictionary, which relates bulk and boundary states must have high computational complexity.

We stress that this concept of *pseudorandom quantum states*, which can be efficiently prepared yet cannot be distinguished from random by computationally bounded observers, is applicable to mixed states, or ensembles of pure states, but not to individual pure quantum states. If a particular pure state can be prepared efficiently by a quantum circuit, that state can always be distinguished efficiently from a maximally mixed state by running the circuit backwards. An ensemble of pure states can be pseudorandom only if it contains superpolynomially many pure states, where the observer who draws a sample from the ensemble and attempts to distinguish this sampled state from a maximally mixed state has no information about which sample was chosen. In contrast, in our definition of complexity for pure states, the observer is permitted to use a different distinguishing circuit for each possible pure state. On the other hand, the existence of pseudorandom quantum states [48] indicates that, for mixed states, our definition of state complexity, namely the computational cost of *distinguishing* the state from a maximally mixed state, can differ substantially from another natural definition, the computational cost of *preparing* the state.

III. COMPLEXITY GROWTH IN RANDOM CIRCUITS

The rigorous statements put forward in Theorems 8 and 9 gain additional meaning when applied to concrete

examples. The literature contains several proofs of design growth in random circuits. Combining these with our rigorous insights yields a number of concrete models for complexity growth.

A. Local random circuits

For concreteness, we focus here on systems comprised of n qubits, i.e., $q = 2$ and $d = 2^n$. Let $\mathsf{G} \subset U(4)$ be a (finite) universal gate set containing inverses, i.e., $g^{-1} = g^\dagger \in \mathsf{G}$ whenever $g \in \mathsf{G}$. We can generate G-*local random circuits* by sequentially applying a random gate $g \in \mathsf{G}$ to a randomly selected pair of neighboring qubits. Repeating this procedure independently for T steps results in random circuits of size T. We refer to the application of each gate as a time step, such that size T circuits are of depth T and have thus evolved to time T. Intuitively, the larger T, the more random the circuit becomes. A seminal result by Brandão, Harrow, and Horodecki confirms this intuition in a precise sense.

Theorem 11 (Corollary 7 in Ref. [12]): *Fix $d = 2^n$, $\epsilon > 0$, $k \leq \sqrt{d}$, and let $\mathsf{G} \subset U(4)$ be a universal gate set containing inverses [49]. Then, the set of all G-local random circuits of size T forms an ϵ-approximate k-design if*

$$T \geq Cn\lceil \log_2(k)\rceil^2 k^{9.5}\left[nk + \log\left(1/\epsilon\right)\right], \tag{19}$$

where $C > 0$ is a (large) constant, which depends on G.

We emphasize that the weights associated with each unitary in this ensemble are defined implicitly by this random procedure. Several different T-sized circuits may give rise to the same final unitary, say U_1, while another one, say U_2, may exclusively be obtained from a single circuit geometry. The weights associated with U_1 and U_2 take into account this behavior, i.e., $p_1 \geq 2p_2$ for our example. However, the fact that the entire ensemble still forms an approximate k-design limits potential fluctuations. This in turn imposes lower bounds on the minimal number of distinct unitaries and severely limits the potential for collisions. It cannot be too likely that two or more different random circuits coincide. These features were conjectured by Brown and Susskind [7, Sec. 6.5] who, in turn, base their counting argument that relates circuit size and complexity on an extreme version of this conjecture: *collisions do not occur at all*. One of the main results of this work is rigorous proof for a weaker version of their conjectured relation between circuit size and complexity. It is an immediate consequence of Theorems 9 and 11.

Corollary 12 (Polynomial relation between circuit size and circuit complexity for local random circuits): *Fix $\delta \in (0,1)$, $r \leq 2^{n/2}$ and set $T \geq Cn^2\left[\log_2(n)r/n\right]^{11}$. Then, the set of all G-local circuits of size T contains at least $\tilde{C}n^r$ unitaries that obey $\mathcal{C}_\delta(U) > r$. Here, $C, \tilde{C} > 0$ are constants that implicitly depend on δ and G.*

This result establishes a polynomial relation between the size T of G-local circuits and the strong δ-unitary complexity that may be achieved in such a model [50]. The relation $T \simeq r^{11}$ is a consequence of Theorem 11, which features a similar relation between the degree $2k$ of an approximate $2k$-design and the circuit size T required to implement it. This relation between complexity and circuit size can certainly be improved, which we soon discuss, but there are fundamental limits: a lower bound on the design depth for random circuits is known. A converse result (Proposition 8 in Ref. [12]) states that for $\epsilon \leq 1/4$ and $k \leq d^{1/2}$, the size of random circuits on n qudits must be at least

$$T \geq \frac{2kn\log q}{q^4\log k} \quad \text{to form an } \epsilon\text{-approximate } k\text{-design.} \tag{20}$$

See Appendix C 10 for a rederivation of this claim.

B. Relating two conjectures

Fix $q = 2$, $d = 2^n$ (n qubits) and suppose that the aforementioned lower bound were not only necessary, but also (approximately) sufficient: G-local circuits of size $T \simeq 2nk/\log_2(n)$ generate (sufficiently accurate) approximate $2k$-designs. Under this assumption, G-local random circuits of size T contain at least $d^{2k}/(k!)^2$ elements with circuit complexity $r \simeq T$. If we also assume $k \leq \sqrt{d}$ [$\log_2(k) \leq n/2$], then this bound can be simplified further as

$$\begin{aligned}\frac{d^{2k}}{(k!)^2} &= 2^{2nk - 2\log_2(k!)} \gtrsim 2^{2k[n-\log_2(k)]} \\ &\gtrsim 2^{nk} \simeq 2^{\log_2(n)T} \geq 2^T.\end{aligned} \tag{21}$$

This is essentially Conjecture 1: the set of all G-local circuits of size T contains an exponentially growing set of elements with complexity $r \simeq T$. This observation provides a relation between Conjecture 1 (linear growth in complexity) to an existing conjecture in quantum information [12].

Conjecture 13 (Linear growth in design): G-*local circuits on n qubits of size $T = O(n^2k)$ form approximate k-designs.*

To achieve a linear growth in complexity it suffices to have a linear growth in design.

C. Linear growth in design for local random circuits at large local dimension

We again consider a $1d$ system comprised of n qudits of local dimension q, with total dimension $d = q^n$, and

evolve the system by a random circuit consisting of local 2-site unitaries drawn Haar randomly from $U(q^2)$. The results of Ref. [12] also ensure that such random circuits form approximate k-designs when the size is $O(n^2k^{11})$. Although Conjecture 13, a linear design growth in G-local random circuits with local qubits, is currently out of reach, progress was made recently in Ref. [28], improving the k dependence for Haar-local random circuits in the limit of large local dimension and giving a linear growth in the circuit size to form a unitary k-design.

Theorem 14 ([28]): *Random quantum circuits on n qudits of local dimension q form approximate unitary k-designs when the circuit size is $T = O(n^2k)$ for some $q > q_0$, where q_0 depends on the size of the circuit [51].*

The approach of Ref. [28] was to consider the frame potential, capturing the 2-norm distance to forming an approximate design, and make use of an exact statistical mechanical mapping [52,53] in order to write the frame potential as the partition function of a triangular lattice model. The contributions to the partition function can be interpreted as domain walls in the lattice model. In the limit of large q, Ref. [28] showed that only a simple sector of domain walls contribute, allowing for the calculation of the k-design circuit size. More precisely, by computing the single domain-wall terms and showing that the multidomain wall terms contribute at subleading order in $1/q$, it was proved that local random circuits exhibit a linear growth in design for some $q > q_0$, where q_0 depends on the circuit size T and moment k.

Theorem 14 and Corollary 12 allow us to establish Conjecture 1 for local random circuits with Haar-random 2-site unitaries in the limit of large q.

Corollary 15 (Linear complexity growth): *Given the set of local random circuits of size T at large q, most circuits have strong complexity $\Omega(T)$, i.e., growing linearly in T for a long time.*

Although Theorem 9 still applies for local Haar random quantum circuits, giving a lower bound on the number of distinct unitaries with high complexity, its meaning becomes less clear when we have a continuous ensemble. We can consider an ensemble of finite cardinality by constructing an ε-covering of the set of random circuits. We review the notion of an ε-covering in Appendix C 10 and give a bound on the cardinality of a covering for local random circuits. Constructing a coarse net then shows that exponentially many random quantum circuits, with Haar-random 2-site unitaries, have high complexity.

Recently, an improvement was made in the q dependence of Theorem 14. By studying the spectral gap of the moment operator for random quantum circuits, and using Knabe bounds to bound the spectral gap, it was proven in Ref. [29] that one requires only the local dimension to be $q \geq \Omega(k^2)$ to form unitary designs. While that work explicitly studied circuits with Haar-random 2-local gates, the seminal result in Ref. [54] that the spectral gap is k independent for any set of universal gates G (containing inverses and comprised of algebraic entries), guarantees that the circuit size required to form a k-design for G-local circuits changes only by a constant. This allows us to extend the result to random quantum circuits instead comprised of 2-local gates randomly chosen from G.

Theorem 16 ([29]): *G-local random quantum circuits on n qudits of local dimension q form approximate unitary k-designs for $T \geq O(n^2k)$ when $q \geq 6k^2$.*

Therefore, Theorem 16 and Corollary 12 immediately establish Conjecture 1 for G-local random quantum circuits for $q \geq 6k^2$.

Lastly, we emphasize that we do not prove linear complexity growth up to time scales of order d. While taking a large enough q will ensure linear design growth for times exponential in n, such a limit still pushes the true exponential time scales of interest, $t \sim d = q^n$, out of reach. Proving an optimal design growth for local random circuits away from the large q limit would allow us to better probe late-time complexity.

D. Stochastic quantum Hamiltonians

There also exist continuous-time models of chaotic dynamics, analogous to random circuits, which scramble in $O(\log n)$ time [25]. In a similar spirit to models of random walks on the unitary group, one can define a one-parameter family of Hamiltonians, which generate a time-dependent unitary evolution. The Hamiltonian on n qubits at a time step s is given by a sum of random all-to-all 2-body interactions, meaning we sum over all possible 1- and 2-local interactions with independently chosen Gaussian random couplings

$$H_s = \sum_{i<j} \sum_{\alpha,\beta} J_{s,i,j,\alpha,\beta} S_i^\alpha S_j^\beta, \tag{22}$$

where S_i^α is a Pauli operator acting on site i with $\alpha = \{0, 1, 2, 3\}$. The couplings are each drawn independently from a Gaussian distribution with zero mean and variance σ^2. Not only are the couplings random in space, but are further chosen randomly at each time step s. The time evolution to time t is then given by

$$U_t = \prod_{s=1}^{t} e^{-iH_s \delta t}, \tag{23}$$

where we consider the continuum limit $\delta t \to 0$ with the variance of the couplings scaling as $\sigma^2 = J/\delta t$ so that the

interactions strength increases proportionally to the inverse time step and where J is a constant.

It was shown in Ref. [26], using similar techniques to Ref. [12], that these stochastic quantum Hamiltonians (also called Brownian circuits) form k-designs in polynomial time.

Theorem 17 (Corollary 10 in Ref. [26]): *For $d = 2^n$ and $\epsilon > 0$, the ensemble of time evolutions by stochastic Hamiltonians in Eq. (22), forms an ϵ-approximate k-design for times*

$$t \geq C\lceil \log_2(k)\rceil^2 k^{9.5}[nk + \log(1/\epsilon)], \tag{24}$$

where $C > 0$ is a constant.

For the Brownian circuit models, the constant prefactor C depends on the local dimension, here chosen to be 2, but also on the interaction strength of the couplings J, $C \sim 1/J$, meaning if the interactions are stronger then the depth required to form a design decreases accordingly.

We can again use the polynomial relation between complexity and design to discuss complexity growth. Theorems 9 and 17 together give that Brownian circuits have a complexity growing polynomially in time as $\Omega(t^{1/11})$.

E. Nearly time-independent Hamiltonian dynamics

There is another random quantum circuitlike construction of a time-dependent Hamiltonian with varying couplings over discrete time steps. This "nearly time-independent" model of Ref. [27] forms k-designs in a circuit size $O(n^2k)$, for moments up to $k = o(\sqrt{n})$, achieving the nearly optimal lower bound with a linear growth in design for a short time.

Consider a $1d$ system of n qudits, with $d = q^n$, and define a time-dependent set of random couplings

$$\mathcal{J}(t,g) = \left\{\lambda/(\lfloor t/2\rfloor + 1), \lambda \in [-g/2, g/2]\right\}, \tag{25}$$

where λ is drawn uniformly at random from the interval. We now generate two ensembles of Hamiltonians with time-dependent couplings

$$\begin{aligned} \mathcal{E}_Z(t) &= \left\{ -\sum_{j<k} h_{jk} Z_j Z_k - \sum_j b_j Z_j \right\}, \\ \mathcal{E}_X(t) &= \left\{ -\sum_{j<k} h_{jk} X_j X_k - \sum_j b_j X_j \right\}, \end{aligned} \tag{26}$$

with $h_{jk} \in \mathcal{J}(t,h)$ and $b_j \in \mathcal{J}(t,b)$, and where $h = \lfloor t/2\rfloor/2$ and $b = \lfloor t/2\rfloor + 1/2$. We then define the time evolution of our system: we evolve by an X-type Hamiltonian $H_X \sim \mathcal{E}_X$ at even time steps and a Z-type Hamiltonian $H_Z \sim \mathcal{E}_Z$ at odd time steps. Then the unitary time evolutions form an ϵ-approximate k-design for $k = o(n^{1/2})$, after T time steps, where

$$T \geq [k + 1/2 + (1/n)\log_2(1/\epsilon)], \tag{27}$$

where each time step involves $O(n^2)$ gates.

This construction forms unitary k-designs almost linearly in time, with the caveat that the time scale is limited to approximately $\sqrt{n}$. Thus we get a linear growth in design at early times, but not exponentially in n. Consequently, this implies a linear growth in complexity at (very) early times.

F. Comment on time-independence

We discuss a few explicit models of complexity growth in systems that are random in both space and time. As we emphasize, one of our results is that a polynomial growth in design implies a polynomial growth in complexity (Corollary 4). Thus, the random circuit and Brownian circuit models, which form approximate k-designs in poly(k) depth, are also explicit examples of systems with a long-time polynomial growth in complexity.

But for an ensemble of time evolutions to form a k-design, randomness in time is likely essential. For instance, consider an ensemble of time evolutions generated by an ensemble of Hamiltonians, $\mathcal{E}_t = \{e^{-iHt},\ H \in \mathcal{E}_H\}$, where $\mathcal{E}_H$ could be a disordered spin system or any ensemble of random Hermitian matrices. The rigid structure of eigenvalues then prohibits the late-time Haar randomness.

Interestingly, the Gaussian unitary ensemble (GUE), an ensemble of $d \times d$ random Hermitian matrices with a unitarily-invariant measure, does come close to an approximate k-design in 2-norm for moments $k \ll d$ at a specific time scale $t \sim \sqrt{d}$ [45]. But at later times, the 2-norm distance between the ensemble of unitaries generated by GUE Hamiltonians and the Haar ensemble becomes large. More generally, one expects that any ensemble of unitary evolutions generated by time-independent Hamiltonians will not form a k-design at late times. A general argument for this is as follows [11], under the exponential map $\lambda \to e^{i\lambda t}$, the eigenvalues of a Hamiltonian are distributed as time-evolving phases on the unit circle. In the limit $t \to \infty$, the phases become uncorrelated and uniformly distributed, unlike the correlated and logarithmically repelling eigenvalues of Haar-random unitaries. Thus, to understand the complexity growth of (ensembles of) time-independent Hamiltonian evolution, we would need to look beyond their design properties for an alternative approach, for instance, by studying the approximate invariance of the ensemble [45,55].

IV. COMPLEXITY IN HOLOGRAPHIC SYSTEMS

Much of the recent interest in quantum complexity in the high-energy literature has centered on the conjectured relation between complexity growth and the long-time growth of black-hole interiors [4,5,56]. More specifically in the context of the AdS/CFT correspondence, the region behind the horizon of an eternal AdS-Schwarzschild black hole grows linearly in time for an exponential time ($t \sim e^n$). The holographic picture is a two-sided geometry connected by a wormhole, where the throat of the wormhole is growing in time. The claim is that the quantum complexity of the dual CFT state is the long-time linearly increasing quantity, which captures the wormhole growth. There have been a number of proposals for what bulk quantity actually computes the complexity, including the volume and action of the AdS wormhole. The complexity/volume conjecture states that the computational complexity of the boundary state is equal to the volume of the wormhole. More precisely, the complexity of time-evolved thermofield double state of the two boundary CFTs is equal to spatial volume behind the horizon of the two-sided geometry on a maximal time slice [5]. The "complexity equals action" conjecture posits that the action computed on a certain region of the bulk geometry, which extends behind the horizon (the Wheeler-DeWitt patch), is the quantity, which is dual to the complexity [6,57]. A lot of progress has been made checking these conjectures and studying complexity growth in holographic systems, see, for instance, [58–64].

In this work we rigorously compute the complexity growth in a number of random circuit models, by relating the growth in design to the growth of complexity, and are able to prove a linear growth in complexity for local random circuits in the limit of large local dimension (albeit, not for an exponentially long time). As we mention, the connection between unitary designs and quantum complexity will likely not inform complexity growth in holography as evolution by time-independent Hamiltonians will not converge to approximate designs. Thus, to study complexity growth in holography we need to explore properties beyond the Haar randomness of the evolution.

A. Strong complexity in the bulk

We briefly discuss why we believe our proposed strong definition of complexity (in terms of a distinguishing measurement), is congruent with expectations from the bulk and might be more suited for holography than the standard definition in terms of the circuit complexity.

One feature we expect complexity growth will exhibit in holography, and fast scrambling systems more generally, is the switchback effect [5]. Consider a time-evolved local operator $\mathcal{O}(t) = e^{-iHt}\mathcal{O}e^{iHt}$ (sometimes called a precursor), where $\mathcal{O}$ might be a single-site Pauli. For such an operator, we anticipate a delay in the onset of the linear complexity growth. For the traditional definition of complexity, consider the minimal circuit approximating the evolution operator e^{-iHt}. The reason for this delay is the exact cancellation of gates outside the lightcone of the spreading operator. Once the operator grows to be the size of the system (more precisely, to have support on weight n Pauli operators) after a time scale called the scrambling time, we expect the complexity of $\mathcal{O}(t)$ to begin its long time linear growth. Such an effect is also present in the bulk for both complexity-volume and action conjectures. This feature is also present in complexity growth of $\mathcal{O}(t)$ under the strong definition of complexity in Definition 2. To be concrete, consider a system of n qubits and the evolved state $e^{-iHt}\mathcal{O}e^{iHt}|\psi_0\rangle$, where H is a chaotic but local Hamiltonian and we take $|\psi_0\rangle$ to be an unentangled product state. Prior to the scrambling time, the optimal measurement to distinguish the evolving state from the maximally mixed state is a simple measurement of a qubit outside the lightcone of the evolving operator. It is not until the scrambling time, when operator has grown to have support on all sites, that the complexity of the distinguishing measurement begins to grow.

Another interesting expectation from holographic systems, where the strong and weak definitions of complexity differ, is that of one-clean qubit. This is essentially the argument given in Lemma 6, to prove that measurement complexity is a stronger definition than standard circuit complexity. Consider a simple initial state $|\psi_0\rangle$, which has been evolved for an exponential time such that $|\psi(t)\rangle$ is maximally complex. If we add a single unentangled qubit to the state $|\psi(t)\rangle \otimes |0\rangle$, then the minimal circuit complexity will be unchanged, but maximal potential complexity increases and the complexity of the state can continue to grow for a long time until it saturates at the new maximal value. For the complexity of a distinguishing measurement, adding a single clean qubit resets the complexity to an order-one value, as the optimal measurement is simply the projection onto the clean qubit. Reference [7] proposed the notion of uncomplexity as the difference of the complexity of a state or unitary from its maximal complexity and suggested an interpretation in the bulk as the total spacetime volume accessible to an infalling observer. Uncomplexity can be thought of as a resource to do useful computation. As we describe, our strong definition of complexity directly encodes this potential for useful quantum computation.

B. Entanglement growth by design

The suggestion that complexity be the dual of the long-time geometric growth in the bulk was motivated by the observation that the wormhole grows long past the times cales at which entropic quantities saturate. Given that we discuss long-time growth in complexity from a long-time growth in design, it is worth commenting on the saturation of entropies after a short growth in design order.

The entanglement entropies for k-designs were studied in Ref. [65]. Specifically, they looked at the Rényi-α entropies of a density matrix ρ: $S^{(\alpha)}(\rho) = [1/(1-\alpha)]\log[\mathrm{Tr}(\rho^\alpha)]$. For any state, the Rényis are bounded above and below by the min-entropy $S_{\min}(\rho) := \lim_{\alpha\to\infty} S^{(\alpha)}(\rho) = -\log(\|\rho\|_\infty)$ [66]. For an n-qubit system, consider the reduced density matrix $\rho_A = \mathrm{Tr}_{\bar{A}}|\psi\rangle\langle\psi|$ on a subsystem A consisting of half the qubits, so that $d_A = d_{\bar{A}}$. Reference [65] showed that for states $|\psi\rangle$ drawn from a $(k > \log d)$ design, the min-entropy of ρ_A is nearly maximal. Therefore, all entropies are nearly maximal once the design order is $k \approx n$. Considering then the time-evolved states of a fast-scrambling system, which forms unitary designs linearly in time, all entropies will saturate at a time of order n. Our arguments ensure complexity growth of approximate k-designs well beyond this entropy saturation threshold.

V. DISCUSSION

We rigorously establish a growth of the quantum complexity in the time evolution of a number of models. We prove that with overwhelming probability, an element sampled from an approximate unitary k-design has a strong complexity that scales at least linearly in k. Moreover, we can count the elements of a design of a given complexity and show that there are at least an exponential number (in k) of distinct unitaries with this complexity. Using the known relations between the evolution time and the design order k thereby establishes a lower bound on the growth of quantum complexity. Specifically, for random quantum circuits we make substantial progress on conjectures by Brown and Susskind and, using a recently established linear relation between the circuit size and design order, prove a linear growth of quantum complexity.

A number of open questions remain. For one, the results in Refs. [28,29] required taking the local dimension q to be large in a k-dependent manner. For local qubits, $T = O(n^2k^{11})$ is still the best known design depth. A proof of a linear design growth for random quantum circuits on qubits up to exponentially high moments would prove a linear growth of complexity for exponentially long times. In this work we largely focus on time-dependent evolution, but the original discussion of a long-time linear complexity growth in holographic systems was focused on time-independent Hamiltonian evolution. It remains to be seen if one can prove anything about the complexity $\mathcal{C}_\delta(e^{-iHt})$ for a specific many-body Hamiltonian H. Lastly, we largely focus on the growth regime for complexity. Nevertheless, there are a number of interesting questions at exponentially late times, when $t \geq d^2$ and complexity saturates at its maximal value.

As we emphasize, our results hold for a new and stronger notion of quantum complexity, defined in terms of optimal distinguishing measurements. We believe strong complexity to be more aptly suited for complexity in holography than circuit complexity, mirroring expectations from the bulk. More broadly, it would be interesting to explore the implications of our strong definition of complexity for quantum error correction and topological order.

ACKNOWLEDGMENTS

The authors thank Dorit Aharonov, Thom Bohdanowicz, Elizabeth Crosson, Felix Haehl, Aram Harrow, Tomas Jochym-O'Connor, Hugo Marrochio, Grant Salton, Eugene Tang, Thomas Vidick, and Beni Yoshida for inspiring discussions and valuable feedback. We also thank the anonymous reviewers for detailed comments and suggestions. All authors acknowledge funding provided by the Institute for Quantum Information and Matter, an NSF Physics Frontiers Center (NSF Grant No. PHY-1733907). J.P. is supported in part by DOE Award No. DE-SC0018407 and by the Simons Foundation It from Qubit Collaboration. R.K. is supported in part by the Office of Naval Research (Award No. N00014-17-1-2146) and the Army Research Office (Award No. W911NF121054). N.H.J. thanks the IQIM at Caltech, McGill University, and UC Berkeley for their hospitality during the completion of this work. Research at Perimeter Institute is supported by the Government of Canada through the Department of Innovation, Science and Economic Development Canada and by the Province of Ontario through the Ministry of Research, Innovation and Science.

APPENDIX A: PROOF OF THE MAIN RESULTS

1. Motivating example computations for Haar-random states

In this section, we provide valuable intuition by analyzing the complexity of Haar-random states using concentration of measure (Levy's lemma). The results presented in the main text will follow by replacing Haar-random states and unitaries with approximate k-designs and measure concentration with moment bounds. Moment bounds, however, are considerably weaker than measure concentration. This, in particular, affects constants and subleading contributions.

a. Most states have high complexity

The Hilbert space of n qudits is enormous, $d = q^n$. Nonetheless, only a tiny fraction of all possible (pure) quantum states seems to be useful for quantum computation, see, e.g., Ref. [46]. Strong state complexity (Definition 2) captures this curious aspect. In order to get a quantitative handle on the set of all pure states we endow it with the uniform measure $d\psi$ that is induced by the Haar measure on the unitary group $U(d)$. Then, random pure states $|\psi\rangle\langle\psi|$ behave like the maximally mixed state $\rho_0 = \mathbb{I}/d$ in expectation. This behavior extends to the

outcome statistics of arbitrary (fixed) measurements:

$$\mathbb{E}_{|\psi\rangle}\left[\mathrm{Tr}\left(M|\psi\rangle\langle\psi|\right)\right]=\mathrm{Tr}\left(M\mathbb{E}_{|\psi\rangle}\left[|\psi\rangle\langle\psi|\right]\right)=\mathrm{Tr}\left(M\rho_0\right). \tag{A1}$$

Concentration of measure (Levy's lemma) ensures that deviations from this average case behavior are exponentially suppressed in concrete instances:

$$\begin{aligned}&\Pr\{|\mathrm{Tr}\left[M(|\psi\rangle\langle\psi|-\rho_0)\right]|\geq\tau\}\\&\quad\leq 2\exp\left(-\frac{d\tau^2}{9\pi^3}\right)\quad\text{for any }\tau\geq 0.\end{aligned} \tag{A2}$$

We refer to Proposition 29 in Appendix D below for a proof of this well-known result. We can combine this assertion with a union bound (Boole's inequality) to conclude for any $r\in\mathbb{N}$ and $\delta\in(0,1)$

$$\begin{aligned}&\Pr\left[\mathcal{C}_\delta(|\psi\rangle)\leq r\right]\\&=\Pr\left\{\max_{M\in\mathsf{M}_r}|\mathrm{Tr}\left[M(|\psi\rangle\langle\psi|-\rho_0)\right]|\geq 1-d^{-1}-\delta\right\}\\&\leq 2|\mathsf{M}_r|\exp\left(-\frac{d(1-d^{-1}-\delta)^2}{9\pi^3}\right)\\&\leq 2.0072|\mathsf{M}_r|\exp\left(-\frac{d(1-\delta)^2}{9\pi^3}\right).\end{aligned} \tag{A3}$$

Suppose that M_r arises from combining at most r elements of a fixed universal gate set $\mathsf{G}\subset U(q^2)$. A naive counting argument reveals $|\mathsf{M}_r|\leq 2d(n+1)^r|\mathsf{G}|^r$ and we refer to Appendix B 2 below for details. We conclude that the $\Pr\left[\mathcal{C}_\delta(|\psi\rangle)\leq r\right]$ remains exponentially suppressed (in $d=q^n$) until

$$r\simeq\frac{q^n}{\log(n)}. \tag{A4}$$

To summarize, a random state is exceedingly likely to have an exponentially large strong δ-state complexity.

The Haar measure has another desirable property. It is fair in the sense that it assigns the same (infinitesimal) weight to each pure state. Such perfectly flat probability distributions allow for turning the probabilistic statement, Eq. (A3), into a quantitative one. From the definition of probability, $\Pr\left[C_\delta(|\psi\rangle)\leq r\right]$ corresponds to the ratio of low-complexity states over all states. Thus, Eq. (A3) ensures that the fraction of low-complexity states remains exponentially tiny until $r\simeq q^n/\log(n)$. In other words, *most pure states have exponentially large complexity.*

b. Most high-complexity states are far apart

In the previous subsection, we saw that concentration of measure, Eq. (A2), allows us to conclude that most quantum states have exponentially high state complexity. This argument, however, does not (yet) tell us anything about the geometric separation between high-complexity states. In principle, a large fraction of high-complexity states could result from tiny perturbations of only a few well-separated core states that have high complexity each. In other words, high-complexity states could come in few tightly packed clusters, in which case the effective number of high-complexity regions could still be comparatively small.

The probabilistic method [67] allows us to prove that extreme clustering cannot occur: *there are exponentially many high-complexity states whose pairwise distance is almost maximal.*

We show this statement by induction based on two features of Haar-random states. Firstly, we use the main result from the previous subsection. Choose $r\lesssim q^n/\log(n)$ such that Eq. (A3) ensures

$$\Pr\left[\mathcal{C}_\delta(|\psi\rangle)\leq r\right]\leq 2.0072|\mathsf{M}_r|\exp\left(-\frac{d(1-\delta)^2}{9\pi^3}\right)<\frac{1}{2}. \tag{A5}$$

The parameter r is chosen such that Haar-random states exceed this complexity with probability (at least) $1/2$. Concentration of measure also implies that a Haar-random state is very likely to be far away from any fixed state $|\phi\rangle\langle\phi|$. For any $\Delta\in(0,1)$,

$$\begin{aligned}&\Pr\left[\frac{1}{2}\||\psi\rangle\langle\psi|-|\phi\rangle\langle\phi|\|_1\leq 1-\Delta\right]\\&\quad=\Pr\left[|\langle\psi|\phi\rangle|^2\geq\Delta^2\right]\leq 3\exp\left(-\frac{\Delta^2 d}{9\pi^3}\right).\end{aligned} \tag{A6}$$

This bound readily follows from Eq. (A2) (set $M=|\phi\rangle\langle\phi|$ and perform elementary modifications).

The first step in our inductive argument is simple. Equation (A5) asserts that the probability of Haar randomly sampling a low complexity (at most r) state is smaller than $1/2$. This is equivalent to stating that the probability of Haar randomly sampling a high complexity (larger than r) is at least $1/2$. Importantly, this assertion implies that such a state exists, because the probability of sampling one is strictly positive. Choose one such state $|\phi_1\rangle$ as the first state in our list.

To construct the second state in our list, we refine this probabilistic existence argument. The probability of Haar randomly sampling a low-complexity state *or* a state that

is too close to $|\phi_1\rangle$ is bounded by

$$\begin{aligned}\Pr\left[\mathcal{C}_\delta(|\psi\rangle) \le r \cup \frac{1}{2}\||\psi\rangle\langle\psi| - |\phi_1\rangle\langle\phi_1|\|_1 \le 1-\Delta\right] \\ \le \Pr\left[\mathcal{C}_\delta(|\psi\rangle) \le r\right] + \Pr \\ \times \left[\frac{1}{2}\||\psi\rangle\langle\psi| - |\phi_1\rangle\langle\phi|\|_1 \le 1-\Delta\right] \\ < \frac{1}{2} + 3\exp\left(-\frac{\Delta^2 d}{9\pi^3}\right). \end{aligned} \tag{A7}$$

By contraposition, the probability of sampling a state that has high complexity *and* is simultaneously far away from $|\phi_1\rangle$ is at least $\frac{1}{2} - 3\exp[-(\Delta^2 d/9\pi^3)] > 0$. This implies the existence of such a state. Choose one such state $|\phi_2\rangle$ and append it to the list: $\{|\phi_1\rangle, |\phi_2\rangle\}$.

We can now inductively repeat this probabilistic existence argument and iteratively append distant high-complexity states to the list $\{|\phi_1\rangle, \ldots, |\phi_N\rangle\}$. This construction only breaks down once the list cardinality N counterbalances exponential suppression: $\frac{1}{2} - 3N\exp[-(\Delta^2 d/9\pi^3)] \le 0$, or equivalently $N \ge \frac{1}{6}\exp[(\Delta^2 d/9\pi^3)]$. Beyond this threshold, we cannot use simple union bounds and concentration of measure to ensure existence of another list element. Such a threshold, however, scales exponentially in the Hilbert-space dimension: the list $\{|\phi_1\rangle, \ldots, |\phi_N\rangle\}$ contains $N = \frac{1}{6}\exp[(\Delta^2 d/9\pi^3)]$ high-complexity states whose pairwise trace distance is at least $1-\Delta$.

We conclude this subsection with providing a bit of context. Existence proofs based on strictly positive probabilities date back to Erdős who developed them to solve several important problems in graph theory. Today, this technique is known as the *probabilistic method* and does constitute an important tool in applied math, combinatorics, and theoretical computer science [67].

2. Proof of Theorem 8

Haar-random states result from applying a Haar-random unitary $U \in U(d)$ to an arbitrary starting state, say $|\psi_0\rangle$. Now suppose that this unitary U is not chosen from the Haar measure, but from an approximate $2k$-design. By definition, this ensures that the first $2k$ moments of $|\psi\rangle\langle\psi| = U|\psi_0\rangle\langle\psi_0|U^\dagger$ accurately approximate the corresponding Haar moments. While this is too coarse to deduce exponential concentration, Eq. (A2), (this would require matching behavior for *all* moments), polynomial concentration arguments do apply. Fix a measurement $M \in \mathbb{H}_d$ and let $\bar{M} = M - [\mathrm{Tr}(M)/d]\mathbb{I}$ denote its traceless part. Markov's inequality then implies that for any $\tau > 0$

$$\begin{aligned}\Pr\{\left|\mathrm{Tr}\left[M(|\psi\rangle\langle\psi| - \rho_0)\right]\right| \ge \tau\} \\ = \Pr\{\left[\mathrm{Tr}\left(\bar{M}|\psi\rangle\langle\psi|\right)\right]^{2k} \ge \tau^{2k}\} \\ \le \tau^{-2k}\mathbb{E}\left[\mathrm{Tr}\left(\bar{M}|\psi\rangle\langle\psi|\right)^{2k}\right]. \end{aligned} \tag{A8}$$

The final expectation value corresponds to a moment of order $2k$. This is the largest moment that still approximately exhibits Haar-random behavior. Explicit bounds can be derived by exploiting this similarity and we refer to Corollary 24 below for a technical derivation:

$$\Pr\{|\mathrm{Tr}\left[M(|\psi\rangle\langle\psi| - \rho_0)\right]| \ge \tau\} \le (1+\epsilon)\left(\frac{2k}{\tau\sqrt{d}}\right)^{2k}. \tag{A9}$$

Qualitatively, this deviation bound is proportional to d^{-k} and becomes ever more stringent as the design order $2k$ increases. We can now combine this tail bound with a union bound and a counting argument for the measurement set M_r in a fashion analogous to the Haar-random case. For any $r \in \mathbb{N}$ and any $\delta \in (0,1)$ this yields

$$\begin{aligned}\Pr\left[\mathcal{C}_\delta(|\psi\rangle) \le r\right] &\le |\mathsf{M}_r|(1+\epsilon)\left(\frac{2k}{\sqrt{d}\left(1-d^{-1}-\delta\right)}\right)^{2k} \\ &\le 2(1+\epsilon)d(n+1)^r|\mathsf{G}|^r\left(\frac{16k^2}{d(1-\delta)^2}\right)^k, \end{aligned} \tag{A10}$$

where we tacitly assume $(1-\delta) \ge 2d^{-1}$ in the last step. Qualitatively, this probability remains tiny until

$$r \simeq \frac{(k-1)n - 2k\log(k)}{\log(n) + \log(|\mathsf{G}|)} \simeq \frac{k[n - 2\log(k)]}{\log(n)}, \tag{A11}$$

provided that $n \ge |\mathsf{G}|$ and $k < d/2$. We can compare this to the complexity of Haar-random states in Eq. (A4). Note that the two coincide when we consider designs of exponentially large degree. So far, this is a purely probabilistic statement. In contrast to the Haar-uniform case it is *a priori* not clear whether it is possible to transform it into a quantitative one. The reason for this is twofold: (i) the weights p_j associated with different elements from an approximate $2k$-design are typically *not* uniform. This nonuniformity extends to the distribution over the different states $|\psi_i\rangle$; (ii) collisions in the state generation: two (or more) distinct design unitaries can produce the same state.

Fortunately, the defining properties of designs ensure that these deviations cannot be too radical: the weights associated with *distinct* states $|\psi_i\rangle$ must obey $q_j \le (1+\epsilon)\binom{d+k-1}{k}^{-1}$—see Lemma 21 in Appendix C 6 below (or, equivalently, Lemma 3 in the main text). This extra condition does allow for drawing quantitative conclusions. Recall that the probability of an event E is the expected

value of its indicator function $\mathbb{1}\{E\}$. Therefore,

$$\Pr[\mathcal{C}_\delta(|\psi\rangle) > r] = \sum_j q_j \mathbb{1}\{\mathcal{C}_\delta(|\psi\rangle) > r\}$$
$$\leq (1+\epsilon)\binom{d+k-1}{k}^{-1} \sum_j \mathbb{1}\{\mathcal{C}_\delta(|\psi\rangle) > r\}. \tag{A12}$$

The sum on the rhs is simply the cardinality N of the set of states $|\psi\rangle$ with δ-state complexity greater than r and the lhs is $1 - \Pr[\mathcal{C}_\delta(|\psi\rangle) \leq r]$. Substituting the bound, Eq. (A10), into this expression establishes the claim:

$$N \geq \binom{d+k-1}{k} \times \left[\frac{1}{1+\epsilon} - 2d(n+1)^r|\mathsf{G}|^r\left(\frac{16k^2}{d(1-\delta)^2}\right)^k\right]. \tag{A13}$$

3. Proof of Theorem 9

The proof is largely analogous to the proof of Theorem 8. Fix a measurement $M \in \mathbb{H}_d \otimes \mathbb{H}_d$ and an input state $|\phi\rangle \in \mathbb{C}^d \otimes \mathbb{C}^d$. Recall that the bias of distinguishing a unitary channel $\mathcal{U} : \mathbb{H}_d \to \mathbb{H}_d$ from the depolarizing channel $\mathcal{D}$ via this measurement procedure is $\mathrm{Tr}[M(\mathcal{U}\otimes\mathcal{I} - \mathcal{D}\otimes\mathcal{I})(|\phi\rangle\langle\phi|)]$. Moreover, the depolarizing channel corresponds to the Haar average over all unitary channels: $\mathbb{E}_U(\mathcal{U}) = \mathcal{D}$, see, e.g., Lemma 26 in Appendix C 9 below. Now suppose that the corresponding unitary $U \in U(d)$ is chosen randomly from an ϵ-approximate $2k$-design. Markov's inequality yields

$$\Pr\{|\mathrm{Tr}[M\mathcal{U}\otimes\mathcal{I}(|\phi\rangle\langle\phi|)] - \mathrm{Tr}[M\mathcal{D}\otimes\mathcal{I}(|\phi\rangle\langle\phi|)]| \geq \tau\}$$
$$\leq \tau^{-2k}\,\mathbb{E}\Big(\{\mathrm{Tr}[M\mathcal{U}\otimes\mathcal{I}(|\phi\rangle\langle\phi|)] - \mathrm{Tr}[M\mathcal{D}\otimes\mathcal{I}(|\phi\rangle\langle\phi|)]\}^{2k}\Big). \tag{A14}$$

The final expectation value corresponds to the highest $2k$-design moment that still approximates Haar-random behavior. Our main technical contribution in Theorem 10 establishes tight bounds on such Haar-random moments. These generalize approximate $2k$-design ensembles $\mathcal{E}$ in a relatively straightforward fashion:

$$\mathbb{E}_\mathcal{E}\Big(\{\mathrm{Tr}[M\mathcal{U}\otimes\mathcal{I}(|\phi\rangle\langle\phi|)] - \mathrm{Tr}[M\mathcal{D}\otimes\mathcal{I}(|\phi\rangle\langle\phi|)]\}^{2k}\Big) \leq \frac{[(2k)!]^2}{d^k}\left(C_{2k} + \frac{\epsilon}{(2k)!d^{3k}}\right). \tag{A15}$$

See Corollary 23 in Appendix C 8 below for a precise statement and proof. Next, we emphasize that the crude bound $|\mathsf{M}_r| \leq (2d^2+1)n^{2r}|\mathsf{G}|^r$ applies to circuit measurements. Combining the above concentration inequality with a union bound over all measurements $M \in \mathsf{M}_r$ ensures

$$\Pr[\mathcal{C}_\delta(U) \leq r] \leq 3\left(C_{2k} + \frac{\epsilon}{(2k)!d^{3k}}\right) d^2 n^{2r}|\mathsf{G}|^r\left(\frac{64k^4}{d(1-\delta)^2}\right)^k, \tag{A16}$$

where we tacitly assume $(1-\delta) \geq 2d^{-1}$. Qualitatively, this probability remains tiny until

$$r \lesssim \frac{(k-2)[n-4k\log(k)]}{\log(n)+\log|\mathsf{G}|} \simeq \frac{k[n-4\log(k)]}{\log(n)}, \tag{A17}$$

provided that $n \geq |\mathsf{G}|$ and $k \leq d/2$. The definition of an approximate $2k$-design also imposes constraints on the weight fluctuations. Lemma 3 asserts that weights associated with distinct ensemble unitaries must obey $p_j \leq (1+\epsilon)(k!/d^{2k})$. This approximate flatness allows us to turn the probabilistic statement from above into a quantitative one:

$$\Pr[\mathcal{C}_\delta(U) > r] = \sum_j p_j \mathbb{1}\{\mathcal{C}_\delta(U) > r\} \leq (1+\epsilon)\frac{k!}{d^{2k}}\sum_j \mathbb{1}\{\mathcal{C}_\delta(U) > r\}. \tag{A18}$$

The sum on the right counts the cardinality N of distinct unitaries with δ-unitary complexity at least $r+1$, while the lhs may be lower bounded by Eq. (A16):

$$N \geq \frac{d^{2k}}{k!}\left[\frac{1}{1+\epsilon} - 3d^2n^{2r}|\mathsf{G}|^r\left(\frac{1024k^4}{d(1-\delta)^2}\right)^k\right]. \tag{A19}$$

4. Distant and distinct design elements

We show that unitary and state designs contain an exponential number $[\Omega(d^k)]$ of distinct high-complexity elements. But to really capture the ergodic nature of chaotic evolution over the unitary group, these distinct high-complexity elements should be pairwise far apart. Here we address this subtlety and show that unitary and state designs contain an exponential number of distant high-complexity unitaries or states.

a. Distant and distinct state design elements

Consider an element drawn at random from an ϵ-approximate spherical k-design $|\psi\rangle$. Equation (A10) gives that the probability the state has δ-state complexity less than r, $\mathcal{C}_\delta(|\psi\rangle) \leq r$, is bounded to be $O(d^{-k})$ when $r \lesssim kn$. We can also show that the probability an element drawn at random from an ϵ-approximate spherical k-design is close to a fixed reference state $|\phi\rangle$ is polynomially suppressed in

k. Choose $\Delta \in (0,1)$ and combine $\frac{1}{2}\||\psi\rangle\langle\psi| - |\phi\rangle\langle\phi|\|_1 = \sqrt{1-|\langle\psi,\phi\rangle|^2}$ with Markov's inequality to conclude

$$\begin{aligned}&\Pr\left[\frac{1}{2}\||\psi\rangle\langle\psi| - |\phi\rangle\langle\phi|\|_1 \leq 1-\Delta\right]\\ &= \Pr\left[|\langle\psi,\phi\rangle|^2 \geq \Delta^2\right] = \Pr\left[|\langle\psi,\phi\rangle|^{2k} \geq \Delta^{2k}\right]\\ &\leq \Delta^{-2k}\mathbb{E}_{|\psi\rangle}\left[|\langle\psi,\phi\rangle|^{2k}\right] \leq \frac{1+\epsilon}{\Delta^{2k}}\binom{d+k-1}{k}^{-1}.\end{aligned} \tag{A20}$$

The last inequality follows from a k-design moment bound similar to Eq. (18). We refer to the proof of Lemma 21 in Appendix C 6 below for a detailed derivation. Qualitatively, this bound is of order $O(d^{-k})$. We can now use a union bound to limit the probability of a random k-design state to have either low complexity *or* to be close to the reference state,

$$\begin{aligned}&\Pr\left[\mathcal{C}_\delta(|\psi\rangle) \leq r \cup \frac{1}{2}\||\psi\rangle\langle\psi| - |\phi\rangle\langle\phi|\|_1 \leq 1-\Delta\right]\\ &\leq \Pr\left[\mathcal{C}_\delta(|\psi\rangle) \leq r\right] + \Pr\\ &\times\left[\frac{1}{2}\||\psi\rangle\langle\psi| - |\phi\rangle\langle\phi|\|_1 \leq 1-\Delta\right]\\ &\leq 2(1+\epsilon)dn^r|\mathsf{G}|^r\left(\frac{16k^2}{d(1-\delta)^2}\right)^k\\ &\quad + \frac{1+\epsilon}{\Delta^{2k}}\binom{d+k-1}{k}^{-1}.\end{aligned} \tag{A21}$$

As long as $r \lesssim nk$, this bound is also of order $O(d^{-k})$ and, in turn, strictly smaller than one. We know that if the probability of the state having low complexity or being close to our fixed state is strictly less than 1, then there is a nonzero probability of a design element that is of high complexity and is far away from the fixed state. Simply stated, if $\Pr[A \cup B] < 1$ then $\Pr[\bar{A} \cap \bar{B}] > 0$.

We can iterate this procedure to construct a set of high-complexity states that are pairwise separated. As long as the probability that the design element is of low complexity or is close to all elements of the set is less than one, then there exists a design element, which is of high complexity and far away from all other design elements in the set. To construct the set $\{|\psi_1\rangle, \ldots, |\psi_N\rangle\}$, we simply need that

$$\begin{aligned}&\Pr\left[\mathcal{C}_\delta(|\psi_N\rangle \leq r \bigcup_{i=1}^{N-1}\frac{1}{2}\||\psi_N\rangle\langle\psi_N| - |\psi_i\rangle\langle\psi_i|\|_1 \leq 1-\Delta\right]\\ &< 1.\end{aligned} \tag{A22}$$

A union bound then converts this requirement into the following sufficient condition on the set cardinality N:

$$N < \Delta^{2k}\binom{d+k-1}{k}\left[\frac{1}{1+\epsilon} - 2dn^r|\mathsf{G}|^r\left(\frac{16k^2}{d(1-\delta)^2}\right)^k\right]. \tag{A23}$$

For constant $\Delta \in (0,1)$, this threshold is exponential as long as the complexity obeys $r \lesssim k$,

$$N \approx O(d^k) \quad \text{for } \mathcal{C}_\delta(|\psi\rangle) \leq r \approx k. \tag{A24}$$

We note the similarity of this bound to the bound derived for the number of distinct design elements.

a. Distant and distinct unitary design elements

Now we consider a unitary U drawn from an ϵ-approximate unitary k-design $\mathcal{E}$. Equation (A16) bounds the probability of the unitary having δ-unitary complexity less than r, $\mathcal{C}_\delta(U) \leq r$, to be $O(d^{-2k})$ when the complexity is roughly $r \lesssim nk$.

Randomly chosen k-design elements also tend to land far away from any fixed unitary. For some $V \in U(d)$ and $\Delta \in (0,1)$, Markov's inequality implies

$$\begin{aligned}\Pr\left[|\mathrm{Tr}(U^\dagger V)|^2 \geq d^2\Delta^2\right] &= \Pr\left[|\mathrm{Tr}(U^\dagger V)|^{2k} \geq d^{2k}\Delta^{2k}\right]\\ &\leq \frac{\mathbb{E}_{\mathcal{E}}\left[|\mathrm{Tr}(U^\dagger V)|^{2k}\right]}{d^{2k}\Delta^{2k}} \leq \frac{1+\epsilon}{\Delta^{2k}}\frac{k!}{d^{2k}},\end{aligned} \tag{A25}$$

where the last inequality follows from a k-design moment bound. We refer to the proof of Lemma 20 in Appendix C 6 below for a detailed derivation. Next, we apply a trick from the proof of Lemma 7 in the main text: $|\mathrm{Tr}(U^\dagger V)|^2 \geq d^2\Delta^2$ is a necessary condition for $\|\mathcal{U} - \mathcal{V}\|_\diamond < 1-\Delta$. This allows us to conclude

$$\Pr\left[\|\mathcal{U} - \mathcal{V}\|_\diamond \leq 1-\Delta\right] \leq (1+\epsilon)\frac{k!}{d^{2k}}\frac{1}{\Delta^{2k}}. \tag{A26}$$

Qualitatively, this is of order $O(d^{-2k})$.

We now have all the ingredients in place to repeat the argument from the state case. The probability of sampling a unitary that has either low complexity *or* is close to any reference unitary V is

$$\begin{aligned}&\Pr\left[\mathcal{C}_\delta(U) \leq r \cup \|\mathcal{U} - \mathcal{V}\|_\diamond \leq 1-\Delta\right]\\ &\leq 3(1+\epsilon)d^2n^{2r}|\mathsf{G}|^r\left(\frac{1024k^4}{d(1-\delta)^2}\right)^k + \frac{1+\epsilon}{\Delta^{2k}}\frac{k!}{d^{2k}},\end{aligned} \tag{A27}$$

according to a union bound. This is on the order of $O(d^{-2k}) < 1$ as long as the complexity $r \lesssim nk$. By contraposition, this ensures that there exists a design element U_1

that has both high complexity *and* is far away from V. We can use this insight to iteratively construct a set of N high-complexity design unitaries with large pairwise distances. Explicitly, to construct a set of unitaries $\{U_1, \ldots, U_N\}$, we need that

$$\Pr\left[\mathcal{C}_\delta(U_N) \leq r \bigcup_{i=1}^{N-1} \|\mathcal{U}_N - \mathcal{U}_i\|_\diamond \leq 1-\Delta\right] < 1. \quad \text{(A28)}$$

A union bound relates this condition to a sufficient upper bound on the set cardinality N:

$$N < \Delta^{2k} \frac{d^{2k}}{k!} \left[\frac{1}{1+\epsilon} - 3d^2 n^{2r} |\mathsf{G}|^r \left(\frac{1024k^4}{d(1-\delta)^2}\right)^k\right]. \quad \text{(A29)}$$

This threshold is exponential as long as the complexity obeys $r \lesssim k$:

$$N \approx O(d^{2k}) \quad \text{for } \mathcal{C}_\delta(|\psi\rangle) \leq r \approx k. \quad \text{(A30)}$$

APPENDIX B: CONCEPTUAL BACKGROUND AND CONTRIBUTIONS

1. Distinguishing states and channels

This conceptual section will review one fundamental question in probability theory, as well as two quantum generalizations. We refer to Refs. [33,68] for details. The underlying question is the following: *what is the best strategy to distinguish two (biased) coins based on a single toss?* More precisely, we consider the following game: there are two identically looking coins with different biases towards coming up heads when being tossed. These biases are known to the player. A referee then picks one of these coins uniformly at random and hands it to the player. The player is allowed to perform a single toss. Based on the result she must guess which coin she obtained and wins if this guess was correct.

a. Distinguishing classical probability distributions

Consider two (discrete) d-variate random variables. Then, we may represent the associated probability distributions by d-dimensional vectors $p, q \in \mathbb{R}^d$, which are entrywise positive ($p_i, q_i \geq 0$) and whose entries sum up to one. Likewise, a collection of events $E_1, \ldots, E_m$ can be also represented by vectors $e_1, \ldots, e_m \in \mathbb{R}^d$ that are entrywise non-negative and obey the following normalization condition: $\sum_{i=1}^m e_i = \vec{1}$. Here, $\vec{1} = (1, \ldots, 1)^T \in \mathbb{R}^d$ denotes the all-ones vector. The probability of observing the event associated with index i is

$$\Pr[i] = \langle e_i, p\rangle. \quad \text{(B1)}$$

The properties of probability and event vectors then assure $\Pr[i] \geq 0$ and $\sum_{i=1}^m \Pr[i] = 1$. Let us now return to the motivating question: what is the best strategy to distinguish two random variables, characterized by known probability vectors p and q in the single-shot limit? This is a binary question and without loss of generality we can restrict our attention to binary events. Let e_1 denote the event that leads us to guess that we observe the first random variable. The complementary event $e_2 = \vec{1} - e_1$ is then fully characterized as well. Under the additional assumption that either random variable is handed to us with equal prior probability, the probability of success becomes

$$\begin{aligned} p_{\text{cl}} &= \frac{1}{2}\Pr[1|p] + \frac{1}{2}\Pr[2|q] = \frac{1}{2}\left(\langle e_1, p\rangle + \langle e_2, q\rangle\right) \\ &= \frac{1}{2}\left(\langle e_1, p-q\rangle + \langle \vec{1}, q\rangle\right) = \frac{1}{2} + \frac{1}{2}\langle e_1, p-q\rangle. \end{aligned} \quad \text{(B2)}$$

This expression may now be optimized over all possible events e_1 in order to determine the optimal guessing strategy. The only constraints on e_1 are non-negativity and normalization. Together, they demand $0 \leq e_1 \leq \vec{1}$, where the inequality signs are to be understood componentwise. The resulting optimization problem is a *linear program* [44,69]

$$\begin{aligned} &\text{maximize} && \frac{1}{2} + \langle e_1, p-q\rangle \\ &\text{subject to} && \vec{1} \geq e_1 \geq 0, \end{aligned} \quad \text{(B3)}$$

and can be solved in a computationally tractable way. In fact, this problem is simple enough to solve analytically. The optimal e_1 is the indicator function for $p_i \geq q_i$, i.e., $e_i = \mathbb{1}\{p_i \geq q_i\}$. This is the *maximum-likelihood estimator* from statistics. Opt for the distribution that is most likely to produce the outcome that has been observed. This choice achieves an optimal success probability of

$$p_{\text{cl}}^\sharp = \frac{1}{2} + \frac{1}{4}\|p-q\|_{\ell_1}. \quad \text{(B4)}$$

Note that a success probability of $1/2$ can be trivially achieved by mere guessing. The remaining factor (multiplied by 2)

$$\beta_{\text{cl}}^\sharp = \frac{1}{2}\|p-q\|_{\ell_1} = \frac{1}{2}\sum_{i=1}^d |p_i - q_i|, \quad \text{(B5)}$$

is called the *bias* and corresponds to the *total variational distance* between p and q.

b. Distinguishing quantum states

It is useful to think of quantum states ρ as matrix generalizations of probability vectors. Similarly, positive operator-valued measurements (POVM) with m outcomes

are characterized by a collection of positive semidefinite (PSD) matrices $\{M_i\}_{i=1}^m \in \mathbb{H}_d$ that sum up to the identity matrix $\mathbb{I}$. Born's rule states that the probability of observing certain outcomes is

$$\Pr[i] = \mathrm{Tr}(M_i\rho) \quad \text{for all } 1 \le i \le m. \tag{B6}$$

This may be viewed as a noncommutative analog of the classical probability rule in Eq. (B1). One may also adapt the distinguishability game to the quantum setting: what is the probability of correctly distinguishing two quantum states ρ, σ by performing a single measurement? Once more, this is a binary question. We can without loss restrict attention to two-outcome measurements: M_1 and $M_2 = \mathbb{I} - M_1$. We associate the first outcome with opting for ρ while the second outcome flags σ. Similar to the classical case, the probability of success is

$$p_{\mathrm{QS}} = \frac{1}{2} + \frac{1}{2}(M_1, \rho - \sigma), \tag{B7}$$

which corresponds to a bias of $\beta_{\mathrm{QS}} = (M_1, \rho - \sigma)$. We may now optimize over all possible measurements M_1 to obtain the best bias possible:

$$\begin{aligned} \beta_{\mathrm{QS}}^{\sharp} = &\text{maximize} \quad (M_1, \rho - \sigma) \\ &\text{subject to} \quad \mathbb{I} \succeq M_1 \succeq 0. \end{aligned} \tag{B8}$$

The constraint denotes the positive semidefinite order ($A \succeq B$ if and only if $A - B$ is positive semidefinite). This is a semidefinite program [44,69] that is simple enough to solve analytically. The optimal measurement M_1 corresponds to the orthogonal projection onto the positive range of $\rho - \sigma$. The associated optimal bias is

$$\beta_{\mathrm{QS}}^{\sharp} = \frac{1}{2}\|\rho - \sigma\|_1, \tag{B9}$$

which is the *trace distance* of the density matrices ρ and σ. This result is known as the *Holevo-Helstrom theorem* [30,31].

Example 1: *Choose* $\rho = |\psi\rangle\langle\psi|$ *and* $\sigma = \rho_0 = (1/d)\mathbb{I}$. *Then, the (unique) optimal measurement is* $M_1 = |\psi\rangle\langle\psi|$ *and achieves a bias of*

$$\beta_{QS}^{\sharp} = \frac{1}{2}\||\psi\rangle\langle\psi| - \rho_0\|_1 = 1 - \frac{1}{d}. \tag{B10}$$

c. Distinguishing quantum channels

Quantum channels describe evolutions of quantum-mechanical systems. They are linear maps $\mathcal{A} : \mathbb{H}_d \to \mathbb{H}_{d'}$ that map density operators to density operators of potentially different dimension d'.

Suppose that we wish to distinguish two channels, say $\mathcal{A}$ and $\mathcal{B}$ based on a single channel use. For instance, input a concrete quantum state and perform a measurement on the outcome state. This indicates more freedom to maximize the probability of correct distinction by optimizing over potential input states and measurements of the channel output. The laws of quantum mechanics allow for further improving this strategy. It is possible to entangle the input state with a quantum memory: $\rho_{\mathrm{in}} \in \mathbb{H}_d \otimes \mathbb{H}_d$. We then apply the channel to the first quantum system, while the second one is left unchanged in the memory. A final two-outcome measurement $M_1 \in \mathbb{H}_{d'} \otimes \mathbb{H}_d$ on both output and memory state potentially reveals additional information. The outcome state depends on the channel in question. *A priori* there are two possibilities. Either $\rho_{\mathrm{out}} = \mathcal{A} \otimes \mathcal{I}(\rho_{\mathrm{in}})$, or $\rho_{\mathrm{out}} = \mathcal{B} \otimes \mathcal{I}(\rho_{\mathrm{in}})$. Here, $\mathcal{I}(X) = X$ denotes the identity channel acting trivially on the memory. The probability of correctly distinguishing these states—and thus the underlying channels—with a single measurement $M_1 \in \mathbb{H}_{d'} \otimes \mathbb{H}_d$ becomes

$$p_{\mathrm{QC}} = \frac{1}{2} + \mathrm{Tr}\{M_1[\mathcal{A} \otimes \mathcal{I}(\rho_{\mathrm{in}}) - \mathcal{B} \otimes \mathcal{I}(\rho_{\mathrm{in}})]\}. \tag{B11}$$

We may now optimize over all degrees of freedom to maximize the value of p_{QC}. Optimizing the measurement M_1 results in a bias that is proportional to the trace distance of the outcome states. Because of convexity, optimization over potential input states can without loss of generality be restricted to pure states:

$$\beta_{\mathrm{QC}}^{\sharp} = \frac{1}{2}\max_{|\psi\rangle\langle\psi|} \left\|\mathcal{A} \otimes \mathcal{I}(|\psi\rangle\langle\psi|) - \mathcal{B} \otimes \mathcal{I}(|\psi\rangle\langle\psi|)\right\|_1. \tag{B12}$$

This optimal bias is called the *diamond distance* between channels $\mathcal{A}$ and $\mathcal{B}$ [70].

It defines a distance measure between quantum channels that is more complicated than the trace distance between quantum states and the total variational distance between classical probability distributions, respectively. It can be difficult to compute it analytically, but does admit a computationally tractable reformulation (as a semidefinite program) [71–73].

Example 2: *Consider a unitary channel* $\mathcal{U}(\rho) = U\rho U^\dagger \in \mathbb{H}_d$ *and the completely depolarizing channel* $\mathcal{D}(\rho) = [Tr(\rho)/d]\mathbb{I} \in \mathbb{H}_d$. *Then,*

$$\frac{1}{2}\|\mathcal{U} - \mathcal{D}\|_\diamond = 1 - \frac{1}{d^2}, \tag{B13}$$

and optimal strategies are based on maximally entangling the input with the memory: Let $|\Omega\rangle = (1/\sqrt{d})\sum_{i=1}^d |i\rangle \otimes |i\rangle \in \mathbb{C}^d \otimes \mathbb{C}^d$ *be the maximally entangled (Bell) state. Set* $\rho_{in} = |\Omega\rangle\langle\Omega|$ *and measure* $M_1 = (U^\dagger \otimes \mathbb{I})|\Omega\rangle\langle\Omega|(U \otimes \mathbb{I})$.

It is easy to check that this strategy achieves the diamond distance in Eq. (B13). Proving optimality is less trivial. For instance, this claim follows from relating the diamond distance to another norm that is easier to compute. We refer to Ref. [74, Theorem 7] and Ref. [75] for details.

2. Conceptual contributions

a. Cornering "easy" unitary transformations

Fix $d = q^n$. The evolution of a closed, d-dimensional quantum-mechanical system is unitary: $\mathcal{U}(\rho) = U\rho U^\dagger$ with $U \in U(d)$. While evolutions may represent natural processes, they can also be engineered to perform certain tasks, such as quantum computing. Scalability of quantum computing hinges on the important observation that complicated evolutions (quantum gate architectures) can be decomposed into sequences of simple building blocks. A universal gate set $\mathsf{G} \subset U(q^2)$ acting on two (neighboring) qudits forms such a basic set of building blocks. For technical reasons, we assume that G contains the identity (doing nothing), as well as inverses: $g \in \mathsf{G}$ implies $g^\dagger \in \mathsf{G}$.

Universality then means that any unitary $U \in U(d)$ may be accurately approximated by a finite sequence of r unitaries chosen from G. We refer to Fig. 4 for an illustrative example. Such decompositions into sequences of elementary gates provide us with a notion of simplicity. Intuitively, a quantum cicuit V is simple if it may be generated by a G-local circuit of short size. In contrast to depth, size counts the total number of elementary gates in a circuit. For $r \in \mathbb{N}$ we define

$$\mathsf{G}_r := \{V \in U(d) : V \text{ is generated by a} \times \mathsf{G}\text{-local circuit of size } \leq r\}. \tag{B14}$$

We set $\mathsf{G}_0 = \{\mathbb{I}\}$ and the following inclusion relation follows from $\mathbb{I} \in \mathsf{G}$:

$$\mathsf{G}_0 \subseteq \mathsf{G}_1 \subseteq \cdots \subseteq \mathsf{G}_r. \tag{B15}$$

The cardinality of G_r may be bounded by a simple counting argument:

$$\left|\mathsf{G}_r\right| \leq \left(n|\mathsf{G}|\right)^r = \log_q(d)^r |\mathsf{G}|^r. \tag{B16}$$

The fact that G is a universal gate set ensures that G_r becomes dense in $U(d)$ provided that $r \to \infty$. *A priori* G_r depends on the particular choice of universal gate set G. However, the Solayev-Kitaev theorem also asserts that other universal gate sets can be accurately compiled at the cost of a constant overhead only [32].

b. Cornering "easy" measurements

The conceptual question underlying our definition of complexity is binary. Are we facing a pure state (unitary channel), or a maximally mixed state (depolarizing channel)? This allows us to restrict attention to two-outcome measurements, where we associate one outcome with each possibility.

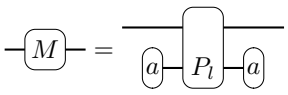
Two-outcome measurements always assume the following form: $(M, \mathbb{I} - M)$, where M obeys $\mathbb{I} \succeq M \succeq 0$. Measuring a quantum state $\rho \in \mathbb{H}_d$ results in two potential outcomes, say "yes" and "no." The probability of observing either is characterized by Born's rule (B6):

$$\begin{aligned}\Pr[\text{"yes"}] &= \operatorname{Tr}(M\rho) \quad \text{and} \quad \Pr[\text{"no"}] \\ &= \operatorname{Tr}[(\mathbb{I} - M)\rho] = 1 - \Pr[\text{"yes"}].\end{aligned} \tag{B17}$$

A *projective* two-outcome measurement is one for which M is an orthogonal projection:

$$M = VP_lV^\dagger, \quad \text{with } P_l = \sum_{i=1}^{l} |i\rangle\langle i| \text{ and } V \in U(d). \tag{B18}$$

Here $l \in [d]$ characterizes the rank of the measurement M and V is a unitary basis change to the eigenbasis of M. *Naimark's theorem*, see, e.g., Refs. [33,76], provides a powerful connection between arbitrary two-outcome measurements M and projective measurements of the form Eq. (B18). Every two-outcome measurement on $\rho \in \mathbb{H}_d$ corresponds to a projective measurement on $\rho \otimes |a\rangle\langle a| \in \mathbb{H}_d \otimes \mathbb{H}_2$, where $|a\rangle\langle a| \in \mathbb{H}_2$ is an ancilla system prepared in a pure state $|a\rangle \in \mathbb{C}^2$. Pictorially (see Appendix C 3 for an introduction of wiring diagrams),



(B19)

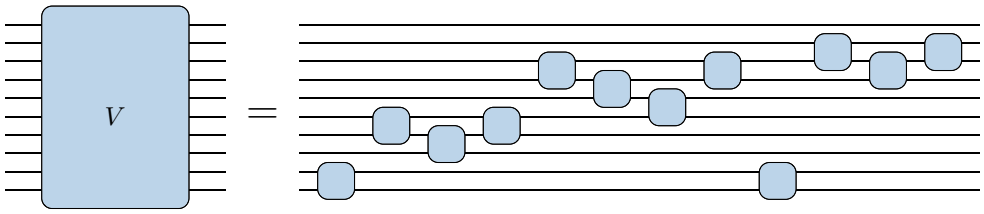


FIG. 4. *Illustration of elementary gate decompositions.* A unitary V on $n = 10$ qudits is comprised of 12 geometrically local 2-qudit gates at random positions, i.e., size$(V) = 12$.

Based on this reformulation of general two-outcome measurements, we model limited resources in the following way:

1. The ancilla state $|a\rangle \in \mathbb{C}^2$ corresponds to a (fixed) simple state, e.g., $|a\rangle = |0\rangle$.
2. The unitary $V \in U(2d)$ must be feasible to implement. More concretely we assume that it is comprised of at most r 2-qudit gates chosen from a (fixed) universal gate set $\mathsf{G} \subset U(q^2)$.
3. The projective measurement $P_l = \sum_{i=1}^{l} |i\rangle\langle i|$ is diagonal in the computational basis.

For fixed $r \in \mathbb{N}$ (circuit size for V), this framework defines the following class of measurements:

$$\mathsf{M}_r = \left\{ \mathrm{Tr}_2 \left(\mathbb{I} \otimes |a\rangle\langle a| V P_{l'} V^\dagger \right) : V \in \mathsf{G}_r, l' \in [2d] \right\} \subset \mathbb{H}_d. \tag{B20}$$

Here, $\mathrm{Tr}_2 : \mathbb{H}_d \otimes \mathbb{H}_2 \to \mathbb{H}_d$ denotes the partial trace. By construction, this set is finite and obeys

$$|\mathsf{M}_r| \leq 2d \left|\mathsf{G}_r\right| \leq 2d \left[\log_q(d) + 1\right]^r |\mathsf{G}|^r = 2d(n+1)^r |\mathsf{G}|^r. \tag{B21}$$

The last equality is contingent on $d = q^n$ (n qudits). The set M_r captures all two-outcome measurements in Hilbert-space dimension d that can be implemented by using a single ancilla qubit, as well as circuits of size at most r.

We can readily extend this family of two-outcome measurements to quantum channel discrimination. But there we need to take into account an additional quantum memory whose dimension is also d (see, e.g., Fig. 3). So, the two-outcome measurement must act on a composite system with dimension $\dim\left(\mathbb{C}^d \otimes \mathbb{C}^d\right) = d^2$. For technical reasons, we also include a single Bell measurement $(|\Omega\rangle\langle\Omega|, \mathbb{I} - \Omega\rangle\langle\Omega|) \subset \mathbb{H}_d^{\otimes 2} \simeq \mathbb{H}_{d^2}$ with $|\Omega\rangle = (1/\sqrt{d}) \sum_{i=1}^{d} |i\rangle \otimes |i\rangle$ in the definition. This implies that the total number of elementary projective measurements is $2d^2 + 1$ and we conclude

$$\begin{aligned} \mathsf{M}_r = &\left\{ \mathrm{Tr}_2 \left(\mathbb{I} \otimes |a\rangle\langle a| V P_{l'} V^\dagger \right) : V \in \mathsf{G}_r, l' \in \left[2d^2\right] \right\} \\ &\cup \left\{ V|\Omega\rangle\langle\Omega|V^\dagger : V \in \mathsf{G}_r \right\} \subset \mathbb{H}_d^{\otimes 2}. \end{aligned} \tag{B22}$$

This modification simplifies the proof of Lemma 7 and is comparatively benign. Assuming $d = q^n$ (n qudits), a simple counting argument reveals

$$|\mathsf{M}_r| \leq (2d^2 + 1)|\mathsf{G}_r| \leq (2d^2 + 1)(2n + 1)^r |\mathsf{G}|^r. \tag{B23}$$

APPENDIX C: TECHNICAL BACKGROUND AND CONTRIBUTIONS

1. Notation and basic facts from matrix analysis

Endow the vector space $\mathbb{C}^d$ with the standard inner product $\langle x|y\rangle$. A pure quantum state is a vector $\psi \in \mathbb{C}^d$ normalized to (Euclidean) unit length, i.e., $\langle \psi, \psi \rangle = 1$. We succinctly denote this by identifying normalized vectors with kets:

$$|\psi\rangle \quad \text{denotes } \psi \in \mathbb{C}^d \quad \text{with } \langle\psi|\psi\rangle = 1. \tag{C1}$$

Let $\mathbb{H}_d$ denote the space of Hermitian $d \times d$ matrices. This is a real-valued subspace of the space of all (complex-valued) $d \times d$ matrices $\mathbb{M}_d$. Fix an orthonormal basis $|1\rangle, \ldots, |d\rangle$ of $\mathbb{C}^d$. Then, the trace of a matrix X is $\mathrm{Tr}(X) = \sum_{i=1}^{d} \langle i|X|i\rangle$. The trace is cyclic, i.e., $\mathrm{Tr}(XY) = \mathrm{Tr}(YX)$ and forms the basis for defining the Schatten p-norms. In particular,

$$\begin{aligned} &\|X\|_1 = \mathrm{Tr}(|X|), |X| = \sqrt{X^2} \quad \text{(trace norm)}, \\ &\|X\|_2 = \sqrt{\mathrm{Tr}(X^2)} \quad \text{(Frobenius norm)}, \\ &\|X\|_\infty = \max_{|y\rangle} |\langle y|X|y\rangle| \quad \text{(operator norm)}. \end{aligned} \tag{C2}$$

Schatten-norms obey the following order relations:

$$\begin{aligned} \|X\|_\infty \leq \|X\|_2 \leq \|X\|_1 \quad \text{and} \quad \|X\|_1 \leq \sqrt{d}\|X\|_2 \\ \leq d\|X\|_\infty \quad \text{for all } X \in \mathbb{H}_d. \end{aligned} \tag{C3}$$

A variant of Hölder's inequality applies to traces of inner products, see, e.g., Ref. [77, Ex. IV.2.12]:

$$|\mathrm{Tr}(XY)| \leq \|X\|_1 \|Y\|_\infty \quad \text{for all } X, Y \in \mathbb{H}_d. \tag{C4}$$

The trace corresponds to a full index contraction. Partial contractions are possible for tensor products and *partial traces* are concrete examples. For $X, Y \in \mathbb{H}_d$ define

$$\mathrm{Tr}_1 (X \otimes Y) = \mathrm{Tr}(X) Y \quad \text{and} \quad \mathrm{Tr}_2 (X \otimes Y) = \mathrm{Tr}(Y) X, \tag{C5}$$

and extend this definition linearly to the tensor product $\mathbb{H}_d^{\otimes 2} \simeq \mathbb{H}_{d^2}$. This definition naturally extends to tensor products of higher order. The following tight bound connects partial traces and operator norms:

$$\begin{aligned} \max\left\{\|\mathrm{Tr}_1(X)\|_\infty, \|\mathrm{Tr}_2(X)\|_\infty\right\} \\ \leq d\|X\|_\infty \quad \text{for all } X \in \mathbb{H}_d^{\otimes 2}. \end{aligned} \tag{C6}$$

A matrix $X \in \mathbb{H}_d$ is PSD if $\langle y|X|y\rangle \geq 0$ for all $y \in \mathbb{C}^d$. We denote this feature by $X \succeq 0$. Positive semidefiniteness is preserved under partial traces:

$$X \in \mathbb{H}_d^{\otimes 2}, X \succeq 0 \quad \text{implies } \mathrm{Tr}_1(X) \succeq 0, \mathrm{Tr}_2(X) \succeq 0. \tag{C7}$$

The trace norm of PSD matrices is particularly simple: $\|X\|_1 = \mathrm{Tr}(X)$ whenever $X \succeq 0$.

2. Convex geometry and optimization

The main technical contributions of this paper are based on bounds that follow from a fundamental argument in convex optimization. Comprehensive references for convex geometry and optimization include Refs. [43,44]. A function $f : \mathbb{H}_d \to \mathbb{R}$ is *convex* if

$$f\left[\tau X + (1-\tau)Y\right] \leq \tau f(X) + (1-\tau) f(Y) \quad \text{for all } X, Y \in \mathbb{H}_d, \tau \in [0,1]. \tag{C8}$$

Linear transformations in the argument preserve this feature. Similarly, a set $\mathcal{K} \subseteq \mathbb{H}_d$ is *convex* if

$$X, Y \in \mathcal{K} \quad \text{imply } \tau X + (1-\tau)Y \in \mathcal{K} \quad \text{for all } \tau \in [0,1]. \tag{C9}$$

Let $\mathcal{K} \subseteq \mathbb{H}_d$ be a convex set. A point $X \in \mathcal{K}$ is an *extreme point* if $Y, Z \in \mathcal{K}$ and $X = \tau Y + (1-\tau)Z$ for some $\tau \in (0,1)$ necessarily imply $Y = Z = X$. Extreme points form the boundary of a convex set.

Example 3: *The set of all quantum states in $\mathbb{H}_d$ is the convex hull (i.e., the set of all convex combinations) of pure states:*

$$\{\rho \in \mathbb{H}_d : Tr(\rho) = 1, \rho \succeq 0\} = conv\left\{|\psi\rangle\langle\psi| : |\psi\rangle \in \mathbb{C}^d\right\}. \tag{C10}$$

All extreme points are pure states.

Fact 18 (Convex functions achieve their maximum at an extreme point): *Let $\mathcal{K} \subseteq \mathbb{H}_d$ be a convex set and let $f : \mathcal{K} \to \mathbb{R}$ be a convex function. Then, there exists an extreme point $X_\sharp$ of $\mathcal{K}$ such that*

$$\max_{X \in \mathcal{K}} f(X) \leq f(X_\sharp). \tag{C11}$$

This result justifies the presentation of the diamond distance in Eq. (B12). The function $X \mapsto \|\mathcal{A} \otimes \mathcal{I}(X) - \mathcal{B} \otimes \mathcal{I}(X)\|_1$ is convex (norms are convex and the channel acts like a linear transformation of the argument) and pure states are the extreme points of the set of all quantum states. Hence,

$$\max_{\rho} \|\mathcal{A} \otimes \mathcal{I}(\rho) - \mathcal{B} \otimes \mathcal{I}(\rho)\|_1 = \max_{|\psi\rangle\langle\psi|} \|\mathcal{A} \otimes \mathcal{I}(|\psi\rangle\langle\psi|) - \mathcal{B} \otimes \mathcal{I}(|\psi\rangle\langle\psi|)\|_1. \tag{C12}$$

The following technical result will prove highly valuable for establishing bounds on very general Haar moments.

Lemma 19: *Fix $A \in \mathbb{H}_d$ PSD ($A \succeq 0$). Then, the function $h(X) = Tr(XAXA)$ is non-negative and convex for all $X \in \mathbb{H}_d$.*

Proof. Apply an eigenvalue decomposition: $A = U(\sum_{i=1}^d \alpha_i |i\rangle\langle i|)U^\dagger$. The assumption that A is PSD ensures $\alpha_1, \ldots, \alpha_d \geq 0$. Next, fix $X \in \mathbb{H}_d$ arbitrary, set $\tilde{X} = U^\dagger X U$ and compute

$$\mathrm{Tr}(XAXA) = \sum_{i,j=1}^d \alpha_i \alpha_j \left|\langle i|\tilde{X}|j\rangle\right|^2 \geq 0. \tag{C13}$$

This establishes non-negativity of $h(X)$. For convexity, fix $X, Y \in \mathbb{H}_d$ and $\tau \in [0,1]$. Set $\bar{\tau} = 1 - \tau$ and note that $\tau\bar{\tau} = \tau - \tau^2 = \bar{\tau} - \bar{\tau}^2 \geq 0$. Non-negativity moreover implies $h(X - Y) \geq 0$ and we can readily deduce convexity:

$$\begin{aligned} h(\tau X + \bar{\tau} Y) &= \tau^2 \mathrm{Tr}(XAXA) + 2\tau\bar{\tau}\mathrm{Tr}(XAYA) + \bar{\tau}^2 \mathrm{Tr}(YAYA) \\ &= \tau h(X) - \tau\bar{\tau}\left[\mathrm{Tr}(XAXA) - 2\mathrm{Tr}(XAYA) \right. \\ &\quad \left. + \mathrm{Tr}(YAYA)\right] + \bar{\tau} h(Y) \\ &= \tau h(X) - \tau\bar{\tau} h(X - Y) + \bar{\tau} h(Y) \leq \tau h(X) + \bar{\tau} h(Y). \end{aligned} \tag{C14}$$

■

3. Wiring calculus

Wiring diagrams, sometimes also known as tensor network diagrams, provide a graphical way for computing contractions between tensors. Here we provide only a brief overview and refer to the recent survey [41] and lecture notes [68] for a detailed introduction. The wiring formalism associates a box with every tensor and a line emanating from the box with every index. Connected lines represent contracted indices. More precisely, we place contravariant indices of a tensor on the left of the box and covariant ones on the right. Table I contains all the essential rules necessary for the scope of this work.

Importantly lines can be bent at will without changing the value of an equation [78]. For instance, let $\rho = |\psi\rangle\langle\psi| \in \mathbb{H}_d$ be a pure quantum state and suppose that $M \in \mathbb{H}_d$ is measurement. We can then represent Born's rule pictographically as

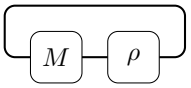
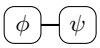
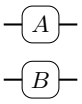
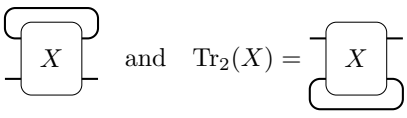
$\mathrm{Tr}(M\rho) =$  $= \langle\psi|M|\psi\rangle$. (C15)

TABLE I. Basic building blocks of wiring calculus.

ket vector	$\|\psi\rangle \in \mathbb{C}^d$	
bra vector	$\langle\phi\| \in \left(\mathbb{C}^d\right)^* \simeq \mathbb{C}^d$	
inner product (contraction)	$\langle\phi\|\psi\rangle$	
matrix	$A \in \mathbb{M}_d$	
matrix product of $A, B \in \mathbb{M}_d$	$AB \in \mathbb{M}_d$	
matrix trace (contraction)	$\mathrm{Tr}(A) \in \mathbb{C}$	
tensor product (vectors)	$\|\psi\rangle \otimes \|\phi\rangle \in (\mathbb{C}^d)^{\otimes 2}$	
tensor product (matrices)	$A \otimes B \in \mathbb{H}_d^{\otimes 2}$	

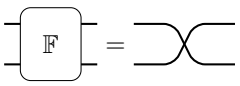
Partial traces also assume a simple form. For $X \in \mathbb{H}_d \otimes \mathbb{H}_d$

$$\mathrm{Tr}_1(X) = \quad \text{and} \quad \mathrm{Tr}_2(X) = . \tag{C16}$$



Wiring calculus is exceptionally well suited to keep track of *flip operators*. Define $\mathbb{F}|i\rangle| \otimes |j\rangle = |j\rangle \otimes |i\rangle$ via its action on computational basis elements and extend this definition linearly to $\mathbb{C}^d \otimes \mathbb{C}^d$. Then,

$$= . \tag{C17}$$

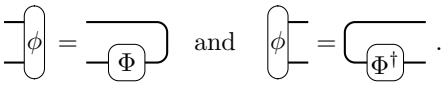


Vectorization is a linear map vec: $\mathbb{M}_d \to \mathbb{C}^d \otimes \mathbb{C}^d$ defined by its action on computational basis elements

$$|\mathrm{vec}\left(|i\rangle\langle j|\right)\rangle := |i\rangle \otimes |j\rangle, \tag{C18}$$

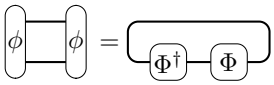
and linearly extended to all of $\mathbb{M}_d$. In wiring calculus, $|\phi\rangle = |\mathrm{vec}(\Phi)\rangle$ corresponds to bending the right (covariant) index of a matrix A to the left (into a contravariant one):

$$= \quad \text{and} \quad = . \tag{C19}$$



It is easy to see that vectorization is an isometry:

$$\langle\phi|\phi\rangle = = = \mathrm{Tr}\left(\Phi^\dagger\Phi\right) = \|\Phi\|_2^2 . \tag{C20}$$



4. Random unitaries and *k*-designs

Here we introduce a few essential concepts from quantum information theory, including a discussion of random unitaries and the notion of a design. First, recall that the Haar measure is the unique left and right invariant measure on the unitary group $U(d)$. We are often interested in moments of the Haar ensemble. Consider an operator X acting on the k-fold Hilbert space $(\mathbb{C}^d)^{\otimes k}$, the k-fold channel, or k-fold twirl, of the operator with respect to the Haar measure on the unitary group is

$$\mathcal{T}_U^{(k)}(X) = \int dU\, U^{\otimes k}(X)U^{\dagger\otimes k}. \tag{C21}$$

Similarly, we can average an operator over an ensemble of unitaries $\mathcal{E} = \{p_i, U_i\}$, a weighted subset of the full unitary group. The k-fold channel with respect to $\mathcal{E}$ is

$$\mathcal{T}_\mathcal{E}^{(k)}(X) = \sum_i p_i U_i^{\otimes k}(X)U_i^{\dagger\otimes k}, \tag{C22}$$

here written for a discrete ensemble, but such an ensemble might be discrete or continuous.

Unitary k-designs. We are often interested in how well an average over an ensemble captures an average over the full unitary group, i.e., how random the ensemble is with respect to the Haar measure on $U(d)$. A *unitary k-design* is an ensemble of unitaries $\mathcal{E} = \{p_i, U_i\}$, for which the k-fold twirl equals its Haar-random counterpart:

$$\mathcal{T}_\mathcal{E}^{(k)}(X) = \mathcal{T}_U^{(k)}(X) \quad \text{for all } X \in \mathbb{H}_d^{\otimes k}. \tag{C23}$$

This means that the ensemble $\mathcal{E}$ exactly captures the first k moments of the Haar ensemble. Unitary operator bases, such as the n-qubit Pauli group, form an exact 1-design. But very little is known about the construction of exact designs for higher k, with the notable exception of $k = 3$ and the n-qubit Clifford group [15–17]. We return to this point when discussing approximate designs.

Schur-Weyl duality. Many of the important analytic expressions for Haar averages rely on *Schur-Weyl duality* [37,38], a deep connection between irreducible representations (irreps) of the unitary group $U(d)$ and the symmetric group S_k. First, when thinking about k-fold Hilbert spaces, there is a useful set of operators that acts on this space, namely permutations of the k copies. A permutation

operator P_σ acts on the computational basis of $(\mathbb{C}^d)^{\otimes k}$ as

$$P_\sigma |i_1, \ldots, i_k\rangle = |i_{\sigma^{-1}(1)}, \ldots, i_{\sigma^{-1}(k)}\rangle. \tag{C24}$$

This action can be extended linearly to all of $(\mathbb{C}^d)^{\otimes k}$. Schur-Weyl duality is the statement that an operator acting on $(\mathbb{C}^d)^{\otimes k}$ commutes with all k-fold unitaries $U^{\otimes k}$ if and only if it is a linear combination of permutation operators

$$U^{\otimes k} X U^{\dagger \otimes k} = X \quad \longleftrightarrow \quad X = \sum_{\sigma \in S_k} c_\sigma P_\sigma. \tag{C25}$$

Many of the exact expressions for Haar moments and random unitary averages in the following subsection follow directly from this powerful result.

5. Haar integration over the unitary group

We now introduce the general formalism for integrating arbitrary moments of random unitaries over the full unitary group with respect to the Haar measure, often referred to as Weingarten calculus. Note that the k-fold twirl in Eq. (C21) describes a linear operator on the tensor product space $\mathbb{H}_d^{\otimes k}$. The associated matrix representation is called the kth moment operator, written as $O_U^{(k)} = \int dU\, U^{\otimes k} \otimes \bar{U}^{\otimes k}$, where $\bar{U}$ denotes the complex conjugate. Weingarten calculus [40,79] provides exact expressions for individual matrix elements of the moment operator:

$$\int dU\, U_{i_1 j_1} \ldots U_{i_k j_k} \bar{U}_{\ell_1 m_1} \ldots \bar{U}_{\ell_k m_k} = \sum_{\sigma,\tau \in S_k} \delta_\sigma(\vec{\imath}|\vec{\ell})\delta_\tau(\vec{\jmath}|\vec{m}) \mathcal{W}g(\sigma^{-1}\tau, d), \tag{C26}$$

where we sum over elements of the permutation group S_k and define a contraction of indices with respect to a permutation $\sigma \in S_k$ as

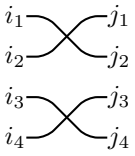
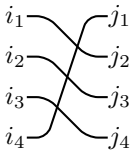
$$\delta_\sigma(\vec{\imath}|\vec{\jmath}) := \prod_{s=1}^{k} \delta_{i_s j_{\sigma(s)}} = \delta_{i_1 j_{\sigma(1)}} \ldots \delta_{i_k j_{\sigma(k)}}. \tag{C27}$$

Mixed moments of U and $\bar{U}$, i.e., averages of $U^{\otimes k} \otimes \bar{U}^{\otimes k'}$ with $k \neq k'$, vanish identically.

It is often convenient to interpret the index contraction $\delta_\sigma(\vec{\imath}|\vec{\jmath})$ as a permutation operator acting on the computational basis of the k-fold space,

$$\delta_\sigma(\vec{\imath}|\vec{\jmath}) = P_\sigma. \tag{C28}$$

For instance, two examples of contractions for $k = 4$ are

$\delta_{\{2,1,4,3\}}(\vec{\imath}|\vec{\jmath}) =$  and $\delta_{\{2,3,4,1\}}(\vec{\imath}|\vec{\jmath}) =$  . (C29)

The weight associated to a given contraction is called the Weingarten function, $\mathcal{W}g(\sigma, d)$. It is a function on elements of S_k and admits an expansion in terms of characters of the symmetric group

$$\mathcal{W}g(\sigma, d) = \frac{1}{k!} \sum_{\lambda \vdash k} \frac{f_\lambda \chi_\lambda(\sigma)}{c_\lambda(d)}, \tag{C30}$$

where we sum over the integer partitions of k that label the irreps of S_k; $\chi_\lambda(\sigma)$ is an irreducible character of λ, and f_λ is the dimension of the irrep λ. The polynomial in the denominator is defined as

$$c_\lambda(d) = \prod_{(i,j)\in\lambda} (d + j - 1), \tag{C31}$$

where we take a product over the coordinates (i, j) of the Young diagram of λ. Writing λ as an integer partition of k, with elements λ_i, the product is taken over i from 1 to $\ell(\lambda)$, the length of the partition, and j from 1 to λ_i. The expression for the Weingarten function in Eq. (C30), is valid for $k \geq d$ by restricting the sum over partitions of length $\ell(\lambda) \leq d$ [such that the polynomial $c_\lambda(d)$ in the denominator is free of zeroes].

The Weingarten functions depend only on the cycle type of the permutation, where the cycle type of $\sigma \in S_k$ is an integer partition of k. We end this brief exposition by listing the first few unitary Weingarten functions, labeled by cycle type. For $k = 1$, $\mathcal{W}g[(1), d] = (1/d)$, and for $k = 2$, we have

$$\mathcal{W}g[(1,1), d] = \frac{1}{d^2 - 1}, \quad \text{and}$$
$$\mathcal{W}g[(2), d] = -\frac{1}{d(d^2 - 1)}. \tag{C32}$$

k-fold twirl over $U(d)$. The k-fold twirl, Eq. (C21), of an operator over the unitary group can be written using Eq. (C26) as

$$\mathcal{T}_U^{(k)}(X) = \mathbb{E}_U\left[U^{\otimes k}(X)U^{\dagger\otimes k}\right] = \sum_{\sigma,\tau\in S_k} \mathcal{W}g(\sigma^{-1}\tau, d)P_\sigma \mathrm{Tr}(XP_\tau). \tag{C33}$$

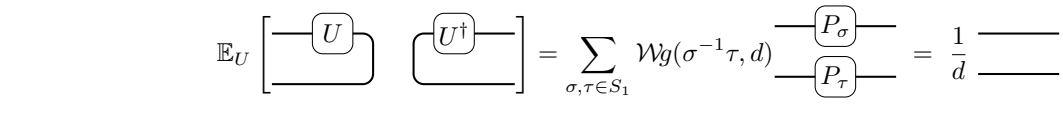
This expression equivalently follows from noting that, by the invariance of the Haar measure, the k-fold twirl $\mathcal{T}_U^{(k)}$ is invariant both under k-fold unitary conjugation and under k-fold conjugation of X.

We also note that the k-fold twirl of a permutation operator is $\mathcal{T}_U^{(k)}(P_\rho) = P_\rho$. Equation (C33), then gives that $\mathcal{W}g(\sigma^{-1}\tau, d)\mathrm{Tr}(P_\tau P_\rho) = \delta_{\sigma,\rho}$. Viewed as a matrix equation, the matrix of Weingarten functions $\mathcal{W}g_{(k)}$ is the pseudoinverse of the $k! \times k!$ matrix $G_{(k)}$ of inner products of permutation operators P_σ (the Gram matrix of P_σ's). The elements of $G_{(k)}$ are the inner products between permutation operators, $\mathrm{Tr}(P_\sigma P_\tau) = d^{\ell(\sigma^{-1}\tau)}$, where $\ell(\sigma^{-1}\tau)$ simply counts the number of closed cycles in the permutation product (equivalently, the length of the cycle type of the product):

$$\mathcal{W}g_{(k)} = G_{(k)}^{-1} \quad \text{with } \mathcal{W}g_{(k)} = \left[\mathcal{W}g(\sigma^{-1}\tau, d)\right]_{\sigma,\tau\in S_k} \quad \text{and} \quad G_{(k)} = \left[\mathrm{Tr}(P_\sigma P_\tau)\right]_{\sigma,\tau\in S_k}. \tag{C34}$$

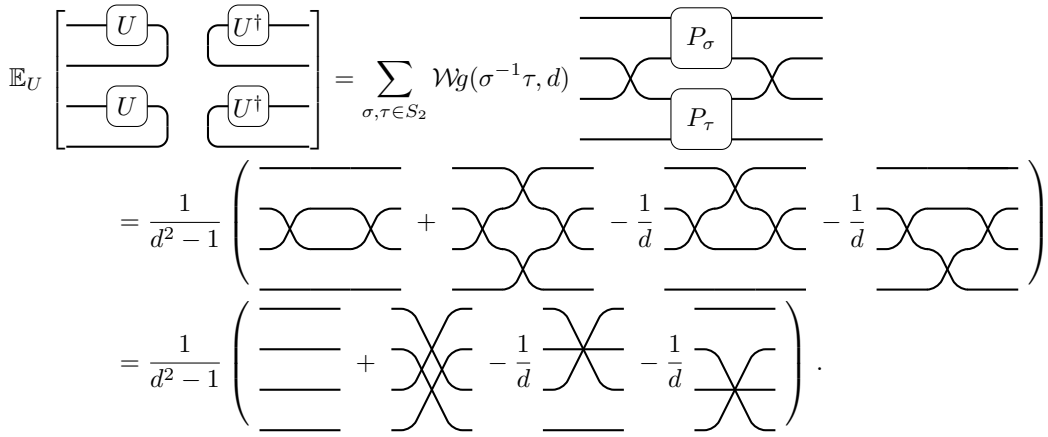
For more discussion on this, see Refs. [11,79]. The matrix inverse exists for $k \le d$. Although elegant, this derivation of the Weingarten functions quickly becomes intractable as we need to invert a $k! \times k!$ matrix. The representation theoretic definition in Eq. (C30) is straightforward to use in computing high moments.

Wiring diagrams for the first few Haar moments. To set up the calculations that will follow in the next section, we explicitly write out the wiring diagrams in the first two moments, detailing the index contractions one must take. For $k = 1$, we simply have



(C35)

For $k = 2$, we sum over elements of S_2, separately permuting the internal and external indices as



(C36)

Moments of traces. We can use the formalism introduced above to compute a few simple expressions averaged over the unitary group, which will be of use in later sections. Consider the $2k$th moment of the trace of a random unitary, $|\mathrm{Tr}(U)|^{2k}$, which we integrate over the unitary group as

$$\mathbb{E}_U\left[|\mathrm{Tr}(U)|^{2k}\right] = \sum_{\sigma,\tau\in S_k} \mathcal{W}g(\sigma^{-1}\tau, d)\mathrm{Tr}(P_\sigma P_\tau), \tag{C37}$$

with $\mathrm{Tr}(P_\sigma P_\tau) = d^{\ell(\sigma\tau)}$. View this as a matrix equation, and recall that for $k \leq d$ the Weingarten functions are the inverse of the inner products Eq. (C34). Then, we simply have the trace of the identity matrix, a sum over S_k:

$$\mathbb{E}_U\left[|\mathrm{Tr}(U)|^{2k}\right] = k!. \tag{C38}$$

This quantity is essentially the same as the frame potential [14], a quantity that quantifies the 2-norm distance between an ensemble of unitaries $\mathcal{E}$ and the Haar ensemble. The frame potential for any ensemble is lower bounded by this Haar value.

Averages of pure states. Consider a Haar random state $|\psi\rangle = U|0\rangle$, with $|0\rangle \in \mathbb{C}^d$ and $U \in U(d)$, and take the k-fold average with respect to the unitary group. Then,

$$\begin{aligned}\mathcal{T}_U^{(k)}(|\psi\rangle\langle\psi|^{\otimes k}) &= \sum_{\sigma,\tau\in S_k} \mathcal{W}g(\sigma^{-1}\tau, d) P_\sigma \mathrm{Tr}(P_\tau |\psi\rangle\langle\psi|^{\otimes k}) \\ &= \sum_{\sigma,\tau\in S_k} \mathcal{W}g(\sigma^{-1}\tau, d) P_\sigma,\end{aligned} \tag{C39}$$

as permuting and contracting the pure state moments is the same for any permutation. This also follows from Schur-Weyl duality by noting that the k-fold average is invariant under k-fold unitary conjugation and may thus be expressed as a sum of permutations. Fixing σ above, the sum over τ just gives the sum over Weingarten functions, which is

$$\sum_{\tau\in S_k} \mathcal{W}g(\tau, d) = \frac{1}{k!}\binom{k+d-1}{k}^{-1}. \tag{C40}$$

Equivalently, we can fix this coefficient by taking the trace of Eq. (C39). Thus we find that the k-fold average of a pure state is

$$\mathcal{T}_U^{(k)}(|\psi\rangle\langle\psi|^{\otimes k}) = \binom{k+d-1}{k}^{-1} \Pi_{\mathrm{sym}}, \tag{C41}$$

where $\Pi_{\mathrm{sym}} = (1/k!)\sum_{\sigma\in S_k} P_\sigma$ is the projector onto the symmetric subspace and $\binom{k+d-1}{k}$ is the corresponding dimension.

A similar calculation is to consider the moments of the expectation value of a conjugated operator $\langle\psi|U^\dagger M U|\psi\rangle$, where $|\psi\rangle \in \mathbb{C}^d$ and a Hermitian operator $M \in \mathbb{H}_d$. We find

$$\begin{aligned}&\mathbb{E}_U\left[|\langle\psi|U^\dagger M U|\psi\rangle|^k\right] \\ &\quad= \sum_{\sigma,\tau\in S_k} \mathcal{W}g(\sigma^{-1}\tau, d)\mathrm{Tr}(P_\sigma|\psi\rangle\langle\psi|)\mathrm{Tr}(P_\tau M^{\otimes k}).\end{aligned} \tag{C42}$$

Again, as permuting and contracting tensor products of a pure state just gives one, for any τ the σ sum is just a sum over Weingarten functions. Using Eq. (C40) and recalling the definition of the projector onto the symmetric subspace, we conclude

$$\mathbb{E}_U\left[\langle\psi|U^\dagger M U|\psi\rangle^k\right] = \binom{d+k-1}{k}^{-1} \mathrm{Tr}\left(\Pi_{\mathrm{sym}} M^{\otimes k}\right). \tag{C43}$$

6. Approximate k-designs and bounds on weight distributions

Weingarten calculus is a powerful tool. It characterizes twirls over the diagonal representation of the unitary group for arbitrary tensor powers $k \in \mathbb{N}$. In turn, this formula allows for computing moments of random variables that involve Haar random unitaries. These then can be used to establish *generic* features, such as concentration of measure. However, full control of *all* moments comes at a price. It is excessively difficult to sample unitaries directly from the Haar measure. Simple dimension counting highlights that circuits of exponential size are required to implement a Haar-random unitary circuit on n qudits.

The notion of k-designs introduced in Appendix C 4 addresses this issue by allowing one to interpolate between Haar-random ($k = \infty$) and highly structured ($k = 1$) ensembles. Unfortunately, very few explicit constructions of k-designs are known. This lack of efficient constructions can be overcome by relaxing the defining property of a k-design.

Definition 4 (Approximate k-design): *Fix $k \in \mathbb{N}$ and $\epsilon > 0$. A unitary ensemble $\mathcal{E} = \{p_i, U_i\}_{i=1}^N$ is an ϵ-approximate (unitary) k-design if the associated twirling channel $\mathcal{T}_\mathcal{E}^{(k)}(X) = \sum_{i=1}^n p_i U_i^{\otimes k} X (U_i^\dagger)^{\otimes k}$ obeys*

$$\left\|\mathcal{T}_\mathcal{E}^{(k)} - \mathcal{T}_U^{(k)}\right\|_\diamond \leq \frac{k!}{d^{2k}}\epsilon. \tag{C44}$$

Here, $\mathcal{T}_U^{(k)}$ denotes the twirl over the full unitary group (C33) (with respect to the Haar measure).

This definition readily extends to ensembles of infinite cardinality. Several different definitions of approximate k-designs can be found in the literature. By and large these differ in terms of the metric that is used to quantify closeness. We define an approximate design up to additive error, but choose ϵ to scale with d in a manner that mimics relative error, similar to the strong definition of a design used in Ref. [12]. This will also simplify exposition considerably.

The approximate k-design property imposes severe restrictions on associated distribution of weights and the ensemble size.

Lemma 20 (Restatement of Lemma 3): *Let* $\mathcal{E} = \{p_i, U_i\}_{i=1}^N$ *be an* ϵ*-approximate* k*-design for* $U(d)$*. Then,*

$$\max_{1\le j\le N} p_j \le (1+\epsilon)\frac{k!}{d^{2k}} \quad \text{and} \quad N \ge \frac{d^{2k}}{(1+\epsilon)k!}. \tag{C45}$$

Lower bounds on approximate k-design cardinality are known, see, e.g., Ref. [12, Lemma 26] for a similar result. We are not aware of any weight bounds in the literature.

We also consider orbits of approximate k-designs $\mathcal{E} = \{p_i, U_i\}_{i=1}^N$. Fix $|x\rangle \in \mathbb{C}^d$ arbitrary and define $|y_i\rangle = U_i|x\rangle$ for $i \in [N]$. Doing so results in a weighted set of unit vectors. These sets are called approximate complex-projective k-designs [18,80]. They approximately reproduce the first k moments of the uniform distribution on the complex unit sphere. Lower bounds on the cardinality of exact spherical k-designs are known, see, e.g., Ref. [20], but we are not aware of any statement that bounds the associated weights.

Lemma 21: *Let* $\{q_i, |y_i\rangle\}_{i=1}^{N'} \subset \mathbb{C}^d$ *be the weighted set of distinct states contained in an orbit of an* ϵ*-approximate* k*-design. Then,*

$$\max_{j\in[N']} q_j \le (1+\epsilon)\binom{d+k-1}{k}^{-1} \quad \text{and} \quad N' \ge \frac{1}{1+\epsilon}\binom{d+k-1}{k}. \tag{C46}$$

The emphasis on distinct states is justified. Two or more distinct unitaries can give rise to the same state.

Proof of Lemma 20. Fix $j \in [N] = \{1,\dots,N\}$ and use Eq. (C38) to conclude

$$\begin{aligned}\sum_{i=1}^N p_i \left|\mathrm{Tr}(U_j^\dagger U_i)\right|^{2k} &= \mathbb{E}_{\mathcal{E}}\left[|\mathrm{Tr}(U_j^\dagger U)|^{2k}\right] \le k! \\ &+ \underbrace{\mathbb{E}_{\mathcal{E}}\left[|\mathrm{Tr}(U_j^\dagger U)|^{2k}\right] - \mathbb{E}_U\left[|\mathrm{Tr}(U_j^\dagger U)|^{2k}\right]}_{\Delta}.\end{aligned} \tag{C47}$$

The approximate k-design property implies that the mismatch on the rhs remains small. Let $|\Omega\rangle = (1/\sqrt{d})\sum_{i=1}^d |i\rangle \otimes |i\rangle$ denote the maximally entangled state. Then, $\mathrm{Tr}(U) = d\langle\Omega|U\otimes\mathbb{I}|\Omega\rangle$ and we apply Definition 4 to bound

$$\begin{aligned}\Delta &= \mathbb{E}_{\mathcal{E}}\left[|\mathrm{Tr}(U_j^\dagger U)|^{2k}\right] - \mathbb{E}_U\left[|\mathrm{Tr}(U_j^\dagger U)|^{2k}\right] \\ &= d^{2k}\langle\Omega|^{\otimes k}\left(\mathbb{E}_{\mathcal{E}}\left\{\left[(U\otimes\mathbb{I})|\Omega\rangle\langle\Omega|(U\otimes\mathbb{I})^\dagger\right]^{\otimes k}\right\}\right. \\ &\quad \left. - \mathbb{E}_U\left\{\left[(U\otimes\mathbb{I})|\Omega\rangle\langle\Omega|(U\otimes\mathbb{I})^\dagger\right]^{\otimes k}\right\}\right)|\Omega\rangle^{\otimes k} \\ &\le d^{2k}\left\|\mathbb{E}_{\mathcal{E}}\left\{[\mathcal{U}\otimes\mathcal{I}(|\Omega\rangle\langle\Omega|)]^{\otimes k}\right\}\right. \\ &\quad \left. - \mathbb{E}_U\left\{[\mathcal{U}\otimes\mathcal{I}(|\Omega\rangle\langle\Omega|)]^{\otimes k}\right\}\right\|_\infty \\ &\le d^{2k}\left\|\mathcal{T}_{\mathcal{E}}^{(k)} - \mathcal{T}_U^{(k)}\right\|_\diamond \le \epsilon k!.\end{aligned} \tag{C48}$$

Combining both arguments implies $\sum_{i=1}^N p_i \left|\mathrm{Tr}(U_j^\dagger U_i)\right|^{2k} \le (1+\epsilon)k!$. This allows us to conclude

$$\begin{aligned}(1+\epsilon)k! &\ge \sum_{i=1}^N p_i \left|\mathrm{Tr}\left(U_j^\dagger U_i\right)\right|^{2k} \\ &= \sum_{i\ne j} p_i \left|\mathrm{Tr}\left(U_j^\dagger U_i\right)\right|^{2k} + p_j\left|\mathrm{Tr}\left(U_j^\dagger U_j\right)\right|^{2k} \ge p_j d^{2k},\end{aligned} \tag{C49}$$

for $j \in [N]$ arbitrary. The lower bound on the cardinality N is an immediate consequence of this weight restriction:

$$1 = \sum_{i=1}^N p_i \le \sum_{i=1}^N (1+\epsilon)\frac{k!}{d^{2k}} = N(1+\epsilon)\frac{k!}{d^{2k}}. \tag{C50}$$

■

Proof of Lemma 21. The argument is very similar to the proof of Lemma 20. Fix $j \in [N']$, set $M = |y_j\rangle\langle y_j|$ and use Eq. (C43) to conclude

$$\begin{aligned}\sum_{i=1}^{N'} q_i \left|\langle y_j, y_i\rangle\right|^{2k} &= \sum_{i=1}^N p_i \left|\langle y_j|U_i|x\rangle\right|^2 = \mathbb{E}_{\mathcal{E}}\left[\langle x|UMU^\dagger|x\rangle\right] \\ &= \binom{d+k-1}{k}^{-1}\mathrm{Tr}\left(\Pi_{\mathrm{sym}}M^{\otimes k}\right) \\ &\quad + \underbrace{\mathrm{Tr}\left(M^{\otimes k}\{\mathbb{E}_{\mathcal{E}}\left[(U|x\rangle\langle x|U^\dagger)^{\otimes k}\right] - \mathbb{E}_U\left[(U|x\rangle\langle x|U^\dagger)^{\otimes k}\right]\}\right)}_{\Delta}.\end{aligned} \tag{C51}$$

Next, observe that the Haar average obeys $\mathrm{Tr}\left(\Pi_{\mathrm{sym}}M^{\otimes k}\right) = \mathrm{Tr}\left(\Pi_{\mathrm{sym}}|y_j\rangle\langle y_j|^{\otimes k}\right) = 1$. The approximate k-design property in addition implies that the deviation from this ideal value remains small. The matrix Hoelder inequality asserts

$$\begin{aligned}\Delta &= \mathrm{Tr}\left(M^{\otimes k}\{\mathbb{E}_{\mathcal{E}}\left[(U|x\rangle\langle x|U^{\dagger})^{\otimes k}\right]-\mathbb{E}_{U}\left[(U|x\rangle\langle U^{\dagger})^{\otimes k}\right]\}\right)\\ &\leq \|M^{\otimes k}\|_{\infty}\left\|\mathcal{T}_{\mathcal{E}}^{(k)}\left[(|x\rangle\langle x|)^{\otimes k}\right]-\mathcal{T}_{U}^{(k)}\left[(|x\rangle\langle x|)^{\otimes k}\right]\right\|_{1}\\ &\leq \|M\|_{\infty}^{k}\left\|\mathcal{T}_{\mathcal{E}}^{(k)}-\mathcal{T}_{U}^{(k)}\right\|_{\diamond}\leq\epsilon\frac{k!}{d^{2k}}\leq\binom{d+k-1}{k}^{-1}\epsilon,\end{aligned} \tag{C52}$$

because $\|M\|_{\infty}=\||y_j\rangle\|_{\infty}=1$. This allows us to conclude

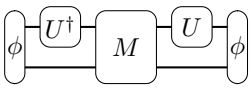
$$(1+\epsilon)\binom{d+k-1}{k}^{-1}\geq\sum_{i=1}^{N'}q_i\left|\langle y_j,y_i\rangle\right|^{2k}=q_j|\langle y_j,y_j r\rangle|^{2k}+\sum_{i\neq j}q_j|\langle y_j,y_i\rangle|^{2k}\geq q_j, \tag{C53}$$

for any $j\in\left[N'\right]$. Both weight and cardinality bound readily follow from this assertion. ■

7. A general moment bound for Haar-random unitaries

Theorem 22 (Detailed restatement of Theorem 10): *Fix* $|\phi\rangle\in(\mathbb{C}^d)^{\otimes 2}$ *and* $M\in\mathbb{H}_d^{\otimes 2}$ *such that* $\mathbb{I}\succeq M\succeq 0$. *Set*

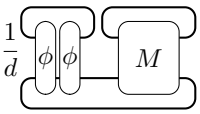
$$S_U(M,\phi):=\mathrm{Tr}\big(M\mathcal{U}\otimes\mathcal{I}(|\phi\rangle\langle\phi|)\big)=$$



(C54)

where $U\in U(d)$ *is chosen uniformly from the Haar measure. Then,*

$$\mu(M,\phi):=\mathbb{E}_U\left[S_U(M,\phi)\right]=\frac{1}{d}$$



$$=\mathrm{Tr}\left(M\mathcal{D}\otimes\mathcal{I}(|\phi\rangle\langle\phi|)\right),$$

(C55)

where $\mathcal{D}(X)=[Tr(X)/d]\mathbb{I}$ *is the depolarizing channel. Moreover, the following bounds apply to all centered moments of order* $k=1,\ldots,d^{2/3}$*:*

$$\mathbb{E}_U\Big\{\left[S_U(M,\phi)-\mu(M,\phi)\right]^k\Big\}\leq\frac{C_k(k!)^2}{d^{k/2}}. \tag{C56}$$

Here, $C_k=[1/(k+1)]\binom{2k}{k}$ *is the* kth *Catalan number.*

8. Moment bounds for approximate designs

Corollary 23: *With the same assumptions in Theorem 22, but suppose that* $U\in U(d)$ *is chosen from an* ϵ*-approximate unitary* k*-design* $\mathcal{E}$*. Then,*

$$\mathbb{E}_{\mathcal{E}}\left[\left(\underbrace{\phi\ U^\dagger\ M\ U\ \phi}_{S_U(M,\phi)}-\underbrace{\frac{1}{d}\ \phi\ \phi\ M}_{\mu(M,\phi)}\right)^k\right]\leq\frac{(k!)^2}{d^{k/2}}\left(C_k+\frac{\epsilon}{k!d^{3k/2}}\right).$$

(C57)

Proof. We can rewrite random variable and (Haar) expectation as

$$S_U(M,\phi)=\mathrm{Tr}\big[M\mathcal{U}\otimes\mathcal{I}(|\phi\rangle\langle\phi|)\big]\quad\text{and}\quad\mu(M,\phi)=\mathrm{Tr}\left[\frac{\mathbb{I}}{d}\otimes\mathrm{Tr}_1(M)\mathcal{U}\otimes\mathcal{I}(|\phi\rangle\langle\phi|)\right]. \tag{C58}$$

Combine them to obtain

$$\bar{S}_U(M,\phi) = S_U(M,\phi) - \mu(M,\phi) = \mathrm{Tr}\left[\tilde{M}\mathcal{U}\otimes\mathcal{I}(|\phi\rangle\langle\phi|)\right], \tag{C59}$$

where $\tilde{M} = M - (1/d)\mathbb{I}\otimes \mathrm{Tr}_1(M) \in \mathbb{H}_d\otimes\mathbb{H}_d$ is a traceless difference of two PSD matrices. Next, fix $k\in\mathbb{N}$ and compare the kth centered moment to its Haar-averaged counterpart:

$$\mathbb{E}_{\mathcal{E}}\left[\bar{S}_U(M,\phi)^k\right] \le \mathbb{E}_U\left[\bar{S}_U(M,\phi)^k\right] + \underbrace{\{\mathbb{E}_{\mathcal{E}}\left[\bar{S}_U(M,\phi)^k\right] - \mathbb{E}_U\left[\bar{S}_U(M,\phi)^k\right]\}}_{\Delta}. \tag{C60}$$

The first contribution is bounded by Theorem 22 and the approximate k-design property (Definition 4) ensures that the mismatch Δ remains controlled:

$$\begin{aligned}\Delta &= \mathrm{Tr}\left(\tilde{M}^{\otimes k}\left\{\mathbb{E}_{\mathcal{E}}\left[(\mathcal{U}\otimes\mathcal{I})^{\otimes k}\right] - \mathbb{E}_U\left[(\mathcal{U}\otimes\mathcal{I})^{\otimes k}\right]\right\}\left[(|\phi\rangle\langle\phi|)^{\otimes k}\right]\right)\\ &\le \|\tilde{M}^{\otimes k}\|_\infty \left\|\left(\mathbb{E}_{\mathcal{E}}\left[\mathcal{U}^{\otimes k}\otimes\mathcal{I}\right] - \mathbb{E}_U\left[\mathcal{U}^{\otimes k}\otimes\mathcal{I}\right]\right)\left[(|\phi\rangle\langle\phi|)^{\otimes k}\right]\right\|_1\\ &\le \|\tilde{M}\|_\infty^k \left\|\mathbb{E}_{\mathcal{E}}\left[\mathcal{U}^{\otimes k}\right] - \mathbb{E}_U\left[\mathcal{U}^{\otimes k}\right]\right\|_\diamond = \|\tilde{M}\|_\infty^k \left\|\mathcal{T}_{\mathcal{E}}^{(k)} - \mathcal{T}_U^{(k)}\right\|_\diamond \le \|\tilde{M}\|_\infty^k \frac{k!}{d^{2k}}\epsilon.\end{aligned} \tag{C61}$$

Finally, use the fact that $\tilde{M}$ is the difference of two PSD matrices to conclude

$$\begin{aligned}\|\tilde{M}\|_\infty &\le \max\left\{\|M\|_\infty, \|\frac{1}{d}\mathbb{I}\otimes\mathrm{Tr}_1(M)\|_\infty\right\}\\ &= \max\left\{\|M\|_\infty, \frac{1}{d}\|\mathrm{Tr}_1(M)\|_\infty\right\} \le 1,\end{aligned} \tag{C62}$$

where we also use Eq. (C6). ■

Corollary 24 (Moments of *k*-design orbits): *For* $|x\rangle\in\mathbb{C}^d$ *and a measurement* $M\in\mathbb{H}_d$ $(\mathbb{I}\succeq M\succeq 0)$ *define*

$$\bar{Q}_U(M,x) = \langle x|U^\dagger M U|x\rangle - \frac{Tr(M)}{d}, \tag{C63}$$

where U *is sampled from an* ϵ*-approximate* k*-design. Then,*

$$\begin{aligned}\mathbb{E}_{\mathcal{E}}\left[\bar{Q}_U(M,x)^k\right] &\le \binom{d+k-1}{k}^{-1}\left(d^{k/2}+\epsilon\right)\\ &\le (1+\epsilon)\left(\frac{k^2}{d}\right)^{k/2}.\end{aligned} \tag{C64}$$

Proof. Let $\bar{M} = M - [\mathrm{Tr}(M)/d]\mathbb{I}$ denote the traceless part of M and note that this reformulation cannot increase the operator norm: $\|\bar{M}\|_\infty \le \|M\|_\infty \le 1$. Moreover,

$$\begin{aligned}\mathbb{E}_{\mathcal{E}}\left[\bar{Q}_U(M,x)^k\right] \le{}& \mathbb{E}_U\left[\bar{Q}_U(M,x)^k\right]\\ &+ \underbrace{\mathbb{E}_{\mathcal{E}}\left[\bar{Q}_U(M,x)^k\right] - \mathbb{E}_U\left[\bar{Q}_U(M,x)^k\right]}_{\Delta},\end{aligned} \tag{C65}$$

and $\Delta \le \|\bar{M}\|_\infty^k \binom{d+k-1}{k}^{-1}\epsilon$ follows from arguments that are analogous to the ones presented in the proof of Lemma 21. Next, apply Eq. (C43) to the remaining Haar expectation:

$$\begin{aligned}\mathbb{E}_U\left[\bar{Q}_U(\bar{M},x)\right] &= \mathbb{E}_U\left[\langle x|U^\dagger M U|x\rangle^k\right]\\ &= \binom{d+k-1}{k}^{-1}\mathrm{Tr}\left(\Pi_{\mathrm{sym}}\bar{M}^{\otimes k}\right).\end{aligned} \tag{C66}$$

This trace can be bounded using $\mathrm{tr}(\bar{M}) = 0$, $\mathrm{tr}(\bar{M}^l) \le \mathrm{tr}(\bar{M}^2)^{l/2}$ for $l\ge 2$ and $\mathrm{tr}(\bar{M}^2) = \|\bar{M}\|_2^2 \le \|M\|_2^2$, see, e.g., Ref. [21, Lemma 17]:

$$\mathrm{Tr}\left(\Pi_{\mathrm{sym}}\bar{M}^{\otimes k}\right) \le \|\bar{M}\|_2^k \le \|M\|_2^k \le d^{k/2}\|M\|_\infty^k \le d^{k/2}. \tag{C67}$$

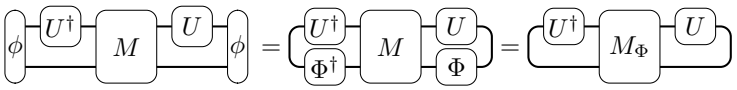
■

9. Proof of the general moment bound

This section is devoted to proving the general moment bound presented in Theorem 22 in Appendix C 7.

a. Reformulation and basic norm bounds

Fix $M \in \mathbb{H}_d \otimes \mathbb{H}_d$ PSD with $\|M\|_\infty \leq 1$ and a state $|\phi\rangle \in \mathbb{C}^d \otimes \mathbb{C}^d$. Use the vectorization correspondence $|\phi\rangle = \mathrm{vec}(\Phi)$ with $\Phi \in \mathbb{M}_{d\times d}$ to rewrite the random variable defined in Theorem 22:

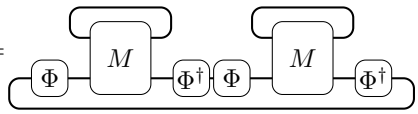
$S_U(M,\phi) =$  (C68)

Here, we implicitly define $M_\Phi := (\mathbb{I} \otimes \Phi^\dagger) M (\mathbb{I} \otimes \Phi)$. Also, recall that vectorization is an isometry, i.e., $\|\Phi\|_2 = \langle\phi|\phi\rangle = 1$. The following auxiliary result bounds the 2-norm of M_Φ and its partial contractions.

Lemma 25: *Fix a PSD matrix $M \in \mathbb{H}_d^{\otimes 2}$ with $\|M\|_\infty \leq 1$ and a matrix $\Phi \in \mathbb{M}_d$ obeying $\|\Phi\|_2 = 1$. Then, $M_\Phi = (\mathbb{I} \otimes \Phi^\dagger) M (\mathbb{I} \otimes \Phi) \in \mathbb{H}_d^{\otimes 2}$ obeys*

$$\|Tr_1(M_\Phi)\|_2 \leq d \quad \textit{and} \quad \|M_\Phi\|_2 \leq \sqrt{d}, \quad \textit{as well as } \|Tr_2(M_\Phi)\|_2 \leq \sqrt{d}. \tag{C69}$$

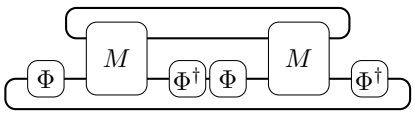
Proof. Observe

$\|\mathrm{Tr}_1(M_\Phi)\|_2^2 =$  $= \mathrm{Tr}\left(\Phi^\dagger \Phi \mathrm{Tr}_1(M) \Phi^\dagger \Phi \mathrm{Tr}_1(M)\right) =: h_1(\Phi^\dagger\Phi)\,.$ (C70)

The function $X \mapsto h_1(X)$ is convex, according to Lemma 19 in Appendix C 2 above [$M \succeq 0$ implies $\mathrm{Tr}_1(M) \succeq 0$]. Moreover, $\rho = \Phi^\dagger\Phi \in \mathbb{H}_d$ is guaranteed to be a quantum state: $\rho = \Phi^\dagger\Phi \succeq 0$ and $\mathrm{Tr}(\rho) = \|\Phi\|_2^2 = 1$. The extreme points of the convex set of all quantum states are pure states. The convex function h_1 achieves its maximum value at such an extreme point (Fact 18 in Appendix C 2) and we infer

$$h_1(\Phi^\dagger\Phi) \leq \max_{\rho \text{ state}} h_1(\rho) = \max_{|\psi\rangle} h_1(|\psi\rangle\langle\psi|) = \max_{|\psi\rangle} \langle\psi|\mathrm{Tr}_1(M)|\psi\rangle^2 = \|\mathrm{Tr}_1(M)\|_\infty^2. \tag{C71}$$

Apply Eq. (C6) to conclude the first estimate: $\|\mathrm{Tr}_1(M)\|_\infty^2 \leq d^2\|M\|_\infty \leq d^2$. The second bound can be derived in a similar fashion. Observe,

$\|M_\Phi\|_2^2 =$  $= \mathrm{Tr}\left(\Phi^\dagger\Phi \otimes \mathbb{I} M \mathbb{I} \otimes \Phi^\dagger\Phi M\right) = h_2(\Phi^\dagger\Phi)\,.$ (C72)

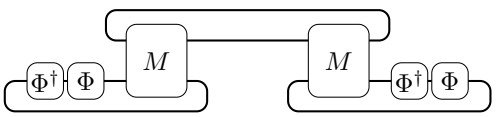
The function $h_2(X)$ is again convex, because $X \mapsto \mathbb{I} \otimes X$ is a linear transformation and $M \succeq 0$. Moreover, $\rho = \Phi^\dagger\Phi$ is again a quantum state. We infer

$$h_2(\Phi^\dagger\Phi) \leq \max_{\rho \text{ state}} h_2(\rho) = \max_{|\psi\rangle} h_2(|\psi\rangle\langle\psi|) = \max_{|\psi\rangle} \|\mathrm{Tr}_2\left(\mathbb{I} \otimes |\psi\rangle\langle\psi| M\right)\|_2^2\,, \tag{C73}$$

because convex functions achieve their maximum at the boundary of convex sets (Fact 18). Applying the relation $\|X\|_2 \leq \sqrt{d}\|X\|_\infty$ for Schatten norms in $\mathbb{H}_d$, we conclude

$$\|M_\Phi\|_2^2 \leq d \max_{|\psi\rangle} \|\mathrm{Tr}_2\left(\mathbb{I} \otimes |\psi\rangle\langle\psi| M\right)\|_\infty^2 = d\left(\max_{|\psi\rangle,|x\rangle} (\langle x| \otimes \langle\psi|) M (|x\rangle \otimes |\psi\rangle)\right)^2 \leq d\|M\|_\infty^2. \tag{C74}$$

The final bound can be established directly. Set $\rho = \Phi^\dagger\Phi$ and observe

$\|\mathrm{Tr}_2(M_\Phi)\|_2^2 =$  $= \|\mathrm{Tr}_2\left(\mathbb{I} \otimes \rho M\right)\|_2^2\,.$ (C75)

Apply $\|X\|_2 \leq \sqrt{d}\|X\|_\infty$ to simplify further

$$\|\mathrm{Tr}_2(\mathbb{I}\otimes\rho M)\|_2 \leq \sqrt{d}\|\mathrm{Tr}_2(\mathbb{I}\otimes\rho M)\|_\infty \leq \sqrt{d}\max_{|x\rangle}\left|\langle x|\mathrm{Tr}_2(\mathbb{I}\otimes\rho M)|x\rangle\right|$$
$$= \sqrt{d}\max_{|x\rangle}\left|\mathrm{Tr}\left(|x\rangle\langle x|\otimes\rho M\right)\right|. \tag{C76}$$

Finally, use matrix Hoelder (C4) to infer the advertised bound:

$$\sqrt{d}\max_{|x\rangle}|\mathrm{Tr}\left(|x\rangle\langle x|\otimes\rho M\right)| \leq \sqrt{d}\max_{|x\rangle}\||x\rangle\langle x|\otimes\rho\|_1\|M\|_\infty = \sqrt{d}\|M\|_\infty. \tag{C77}$$

■

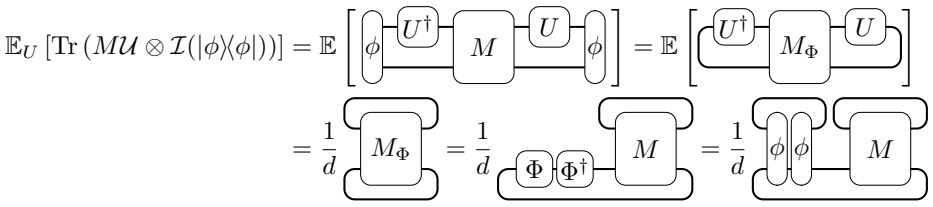
b. Expectation value and centering

The following result is well known in the literature, see, e.g., Ref. [22]. We include a self-contained derivation based on wiring diagrams for the sake of completeness.

Lemma 26 (Averaging unitary channels produces the depolarizing channel): *Fix a PSD matrix $M \in \mathbb{H}_d^{\otimes 2}$ and $|\phi\rangle \in \mathbb{C}^d\otimes\mathbb{C}^d$. Let $\mathcal{U}(X) = UXU^\dagger$ be a Haar-random unitary channel. Then,*

$$\mathbb{E}_U\{Tr\left[M\mathcal{U}\otimes\mathcal{I}(|\phi\rangle\langle\phi|)\right]\} = Tr\left[M\mathcal{D}\otimes\mathcal{I}(|\phi\rangle\langle\phi|)\right] \quad \text{with } \mathcal{D}(\rho) = \frac{Tr(\rho)}{d}\mathbb{I}. \tag{C78}$$

Proof. Averaging over a single unitary U and its adjoint decouples the register in question. Combine this with the reformulation from the previous subsection to conclude

$\mathbb{E}_U\left[\mathrm{Tr}\left(M\mathcal{U}\otimes\mathcal{I}(|\phi\rangle\langle\phi|)\right)\right] =$ 

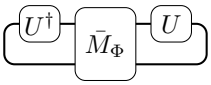
(C79)

The connection to the depolarizing channel readily follows from $\mathcal{D}\otimes\mathcal{I}(|\phi\rangle\langle\phi|) = (\mathbb{I}/d)\otimes\mathrm{Tr}_2(|\phi\rangle\langle\phi|)$. ■

Corollary 27 (Reformulation of the centered random variable): *Fix $|\phi\rangle \in \mathbb{C}^d\otimes\mathbb{C}^d$ (state) and $M \in \mathbb{H}_d^{\otimes 2}$ such that $\mathbb{I} \succeq M \succeq 0$ (measurement). For channels $\mathcal{U}(X) = UXU^\dagger$ and $\mathcal{D}(X) = [Tr(X)/d]\mathbb{I}$ define*

$$S_U(M,\phi) = Tr\left[M\mathcal{U}\otimes\mathcal{I}(|\phi\rangle\langle\phi|)\right], \quad \text{as well as} \quad \mu(M,\phi) = Tr\left[M\mathcal{D}\otimes\mathcal{I}(|\phi\rangle\langle\phi|)\right].$$

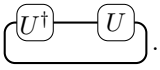
Then, we may rewrite the difference of these variables as

$\bar{S}_U(M,\phi) = S_U(M,\phi) - \mu(M,\Phi) =$ 

(C80)

where $\bar{M}_\Phi = M_\Phi - [Tr(M_\Phi)/d]\mathbb{I} \in \mathbb{H}_d^{\otimes 2}$ is the traceless part of M_Φ [i.e., $Tr(\bar{M}_\Phi) = 0$].

This reformulation immediately follows from the proof of Lemma 26, provided that we rewrite

$\mu(M,\phi) = \frac{1}{d}\mathrm{Tr}(M_\Phi) = \frac{\mathrm{Tr}(M_\Phi)}{d^2}$ 

(C81)

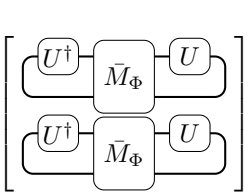
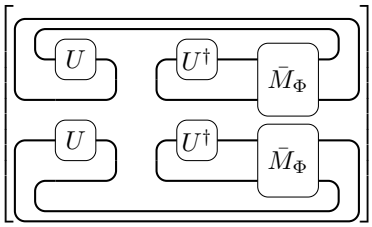
c. Bounds on centered moments

Lemma 28: *With the same assumptions and notation as in Corollary 27, suppose that $U \in U(d)$ is chosen uniformly from the Haar measure. Then, for any $k \leq d^{2/3}$*

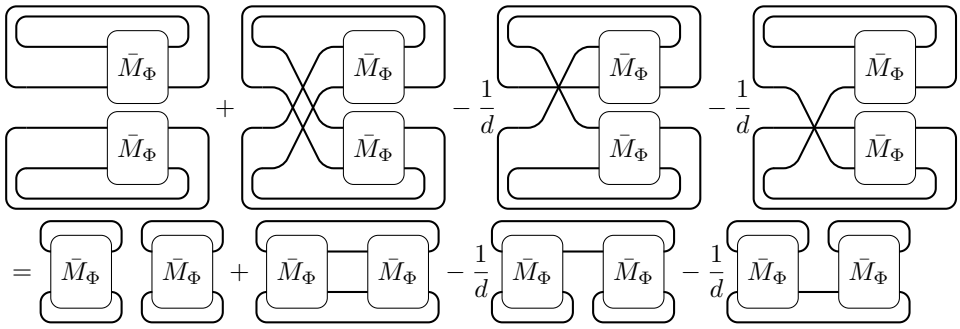
$$\mathbb{E}_U\left[\bar{S}_U(M,\phi)^k\right] \leq C_k \frac{(k!)^2}{d^{k/2}}, \tag{C82}$$

where $C_k = [1/(k-1)]\binom{2k}{k}$ is the kth Catalan number.

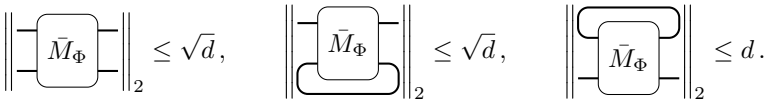
Proof. It is instructive to first analyze and understand the second moment:

$\mathbb{E}_U\left[\bar{S}_U(M,\phi)^2\right] = \mathbb{E}_U$  $= \mathbb{E}_U$  . (C83)

For $k = 2$ there are two permutations: the identity permutation $\mathbb{I} = \{1,2\}$ and swap (or flip) $S = \{2,1\}$. This results in $(k!)^2 = 4$ different contributions to the formula: $(\mathbb{I},\mathbb{I})$, (S,S), $(S,\mathbb{I})$, and $(\mathbb{I},S)$ contribute each. The associated Weingarten functions are $\mathcal{W}g[(1),d] = 1/(d^2-1)$ and $\mathcal{W}g[(2),d] = -[1/d(d^2-1)]$. Ignoring the common factor $1/(d^2-1)$, the individual contributions become

 (C84)

Each term is a full contraction that is also called a tensor network [41,42]. There are three possible constituents for each tensor network: $\bar{M}_\Phi$, $\mathrm{Tr}_2(\bar{M}_\Phi)$, and $\mathrm{Tr}_1(\bar{M}_\Phi)$. Importantly, no full self-contractions can contribute to the overall sum, because $\bar{M}_\phi$ is traceless. This ensures that networks with self-contractions—like the first term—evaluate to zero. Moreover, Lemma 25 bounds the 2-norm of each elementary constituent:

 (C85)

The final bound is considerably larger than the rest. However, the corresponding contribution in the sum (C84) is also suppressed by an additional dimension factor. This is not a coincidence: term 3 can arise only if the cycle classes of (σ,τ) differ from each other. This feature reflects itself in the Weingarten function. For the second moment, we thus obtain the following simple bound (ignoring signs):

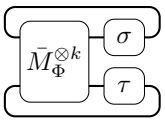
$$\mathbb{E}_U\left[\bar{S}(M,\phi)^2\right] \leq \frac{0 + d + d/d + d^2/d}{d^2-1} = \frac{2d+1}{d^2} \leq 4d^{-1}. \tag{C86}$$

It immediately follows from upper bounding individual terms using Eq. (C85).

This general strategy also applies to higher moments. Fix $k \geq 3$ arbitrary. Then, Weingarten calculus implies

$$\mathbb{E}_U\left[\bar{S}_U(M,\phi)^k\right] = \sum_{\sigma,\tau\in S_k} \mathcal{W}g_d(\sigma,\tau) N_{\sigma,\tau} \times \left[\bar{M}_\Phi, \mathrm{Tr}_2(\bar{M}_\Phi), \mathrm{Tr}_1(\bar{M}_\Phi)\right]. \quad \text{(C87)}$$

Here, each $N_{\sigma,\tau}(\cdot)$ indicates a tensor network diagram that combines (at most) three elementary building blocks according to rules that are dictated by the permutations τ and σ:

$N_{\sigma,\tau} =$  . (C88)

We can without loss restrict summation to tensor networks without self-contractions, because $\mathrm{Tr}(\bar{M}_\Phi) = 0$ ensures that such contributions vanish identically. Next, we apply a powerful general bound to individual tensor networks. Reference [42, Proposition 18] states that the value of a tensor network (without self-contractions) is bounded by the product of 2-norms of the individual constituents. For any σ, τ this implies

$$\begin{aligned} |N_{\sigma,\tau}| &= \left|N_{\sigma,\tau}\left[\bar{M}_\Phi, \mathrm{Tr}_2(\bar{M}_\Phi), \mathrm{Tr}_1(\bar{M}_\Phi)\right]\right| \\ &\leq \|\bar{M}_\Phi\|_2^{\nu_1} \|\mathrm{Tr}_2(\bar{M}_\Phi)\|_2^{\nu_2} \|\mathrm{Tr}_1(\bar{M}_\Phi)\|_2^{\nu_3}, \end{aligned} \quad \text{(C89)}$$

where $\nu_1, \nu_2, \nu_3 \in [k]$ denote the number of times each basic building block occurred in the network. Clearly, $\nu_1 + \nu_2 + \nu_3 = k$ and we can combine this with Eq. (C85) to conclude

$$|N_{\sigma,\tau}| \leq d^{\nu_1/2} d^{\nu_2/2} d^{\nu_3} = d^{k/2+\nu_3/2}. \quad \text{(C90)}$$

The final contribution $d^{\nu_3/2}$ is always counterbalanced by the Weingarten function, i.e., the dangerous terms are always suppressed by powers of $1/d$. As we discuss, the Weingarten functions $\mathcal{W}g(\sigma,d)$ depend only on the cycle type of the permutation σ. The asymptotic behavior is $\mathcal{W}g(\sigma,d) \sim 1/d^{2k-\ell(\sigma)}$, where ℓ is the length of the cycle type, i.e., the number of cycles in the permutation. The leading-order terms are those for which the cycle type is $(1,1,\ldots,1)$, the partition of $2k$ into 1's. For $\mathcal{W}g(\sigma^{-1}\tau,d)$ this corresponds to terms with $\sigma = \tau$. Returning to the problem at hand, we contract the upper indices of the k copies of $\bar{M}_\Phi$ with respect to σ and the lower indices with τ, as shown in Eq. (C88). The leading-order terms are those in which we act similarly on upper and lower indices. In order to generate terms in the tensor-network contraction of M's containing a dangerous contribution, $\mathrm{Tr}_1(\bar{M}_\Phi)$, the lengths of the cycle types of the two permutations must differ by at least one in order to generate a contraction, a length one cycle, in the σ indices:

$$\begin{aligned} \mathbb{E}_U\left[\bar{S}_U(M,\phi)^k\right] &\leq \sum_{\tau,\sigma\in S_k} |\mathcal{W}g(\sigma^{-1}\tau,d)| N_{\sigma,\tau} \\ &\leq \sum_{\sigma\in S_k} \mathcal{W}g[(1,\ldots,1),d] d^{k/2} \\ &\quad + \sum_{\tau\neq\sigma\in S_k} \mathcal{W}g(\sigma^{-1}\tau,d) d^{k/2+\nu_3/2}. \end{aligned} \quad \text{(C91)}$$

Although, the $\mathrm{Tr}_1(\bar{M}_\Phi)$ terms will only contribute at subleading order, they appear with a larger contribution in powers of d. Thus, to rigorously upper bound the expression, we need bounds on the Weingarten functions as well as on the number of terms ν_3 which appear in a given tensor network $N_{\sigma,\tau}$.

Precise upper bounds on the Weingarten functions are known [40,81]. For our purposes, it will be convenient to use the (slightly weaker) bound in Ref. [82], which states that for $k \leq d^{2/3}$

$$|\mathcal{W}g(\sigma,d)| \leq \frac{3}{2}\frac{C_{k-1}}{d^{2k-\ell(\sigma)}}, \quad \text{(C92)}$$

where C_k is the kth Catalan number.

Now we establish that $\nu_3(\sigma,\tau)$, the number of dangerous terms $\mathrm{Tr}_1(\bar{M}_\Phi)$ terms in a given $N_{\sigma,\tau}$, is bounded by the distance between the permutations σ and τ as $\nu_3(\sigma,\tau) \leq 2d(\sigma,\tau)$. First we note a few facts about the symmetric group. $d(\sigma,\tau)$ is defined as the minimal number of transpositions needed to take σ to τ, and defines a distance between the permutations. Specifically, $d(\sigma,\tau)$ is a metric on the Cayley graph of the symmetric group with the generating set of transpositions. The length of the cycle type of a permutation $\sigma \in S_k$ is related to the number of transpositions needed to build σ from the identity permutation as $\ell(\sigma) = k - d(\sigma,\mathbb{I})$. Furthermore, a transposition changes the number of cycles in a permutation by exactly one.

Recalling Eq. (C88), the terms $\mathrm{Tr}_1(\bar{M}_\Phi)$ appear only in a given tensor-network diagram when the permutation σ has a fixed point where τ does not, i.e., there is a self-contraction in the σ indices of $\bar{M}_\Phi^{\otimes k}$ and not the τ indices. This implies that σ has a length one cycle at a point where τ does not. As $d(\sigma,\tau)$ is the minimal number of transpositions required to take σ to τ, and a transposition can only change the number of cycles by exactly 1, then for every two dangerous terms the distance between the permutations σ and τ must increase by at least one. This shows that $\nu_3(\sigma,\tau)$ is bounded as

$$\nu_3(\sigma,\tau) \leq 2d(\sigma,\tau) = 2[k - \ell(\sigma^{-1}\tau)]. \quad \text{(C93)}$$

Returning to the general moment bound, we can apply the bound on Weingarten functions in Eq. (C92) and the bound

on ν_3 to show that

$$\begin{aligned}\mathbb{E}_U\left[\bar{S}_U(M,\phi)^k\right] &\le \sum_{\tau,\sigma\in S_k} |\mathcal{W}g(\sigma^{-1}\tau,d)|N_{\sigma,\tau} \\ &\le \sum_{\tau,\sigma\in S_k} \frac{3}{2} C_{k-1} d^{\ell(\sigma^{-1}\tau)-2k+k/2+\nu_3/2} \\ &\le \sum_{\tau,\sigma\in S_k} \frac{3}{2} C_{k-1} d^{-k/2} \le C_k (k!)^2 d^{-k/2}, \end{aligned} \tag{C94}$$

which establishes the claim. ■

10. ε-coverings of local random circuits

We want to extend our results in Sec. III A on complexity growth to local random circuits, where the gates are chosen Haar randomly from $U(q^2)$. Obviously, the ensemble of size T circuits is continuous and statements about the number of states of a certain complexity become less meaningful. Nevertheless, we can consider an ε-covering of the ensemble of local random quantum circuits (RQCs) in order to make concrete statements about complexity growth.

We say that a set of unitaries V is an ε-covering of a set of unitaries U if for all $U \in \mathsf{U}$ there is some $V \in \mathsf{V}$ such that $\|U(\cdot)U^\dagger - V(\cdot)V^\dagger\|_\diamond \le \varepsilon$.

Consider the set of local random circuits of size T, where again we act on n local qudits with local dimension q and with local gates chosen Haar randomly from $U(q^2)$. Following Lemma 27 from Ref. [12], we can bound the size of an ε-covering of the set $\mathcal{E}_{\mathrm{RQC}}$ size T local RQCs. Approximating each local gate to accuracy ε/T, we construct a covering in diamond norm of each gate with size $\le \left(10T/\varepsilon\right)^{q^4}$. For the n^T choices of gates in the circuit, we conclude that there exists an ε-covering $\widetilde{\mathcal{E}}_{\mathrm{RQC}}$ of size T RQCs with cardinality

$$|\widetilde{\mathcal{E}}_{\mathrm{RQC}}| \le n^T \left(\frac{10T}{\varepsilon}\right)^{Tq^4}. \tag{C95}$$

Furthermore, if an ensemble $\mathcal{E}$ forms an ϵ-approximate unitary k-design, then the ε-covering of $\mathcal{E}$ will form an ϵ'-approximate unitary design with $\epsilon' = \epsilon + 2d^{2k}\varepsilon$ (from Proposition 8 in Ref. [12]). Using the lower bound on the cardinality of an approximate design in Lemma 20 and the upper bound on the cardinality of an ε-covering of size T local random circuits in Eq. (C95), means that for $\tilde{\mathcal{E}}_{\mathrm{RQC}}$ to form an approximate design, we must have

$$\frac{1}{1+\epsilon'}\frac{d^{2k}}{k!} \le |\widetilde{\mathcal{E}}_{\mathrm{RQC}}| \le n^T \left(\frac{10T}{\varepsilon}\right)^{Tq^4}. \tag{C96}$$

This gives a lower bound on the size for local random circuits to form k-designs

$$T \ge \frac{2kn\log q}{q^4 \log k}. \tag{C97}$$

Therefore, an optimal random circuit implementation of a unitary design will have at least an essentially linear scaling in both n and k.

APPENDIX D: CONCENTRATION OF MEASURE FOR HAAR-UNIFORM VECTORS

Proposition 29: *Fix $M \in \mathbb{H}_d$ with $\|M\|_\infty \le 1$ and suppose that $|\psi\rangle \in \mathbb{C}^d$ is chosen uniformly from the complex unit sphere. Then,*

$$\begin{aligned} &Pr\left[|\langle\psi|M|\psi\rangle - \mathbb{E}\left(\langle\psi|M|\psi\rangle\right)| \ge \tau\right] \\ &\qquad \le 2\exp\left(-\frac{d\tau^2}{9\pi^3}\right) \quad \textit{for any } \tau \ge 0. \end{aligned} \tag{D1}$$

The proof is standard and we include it in this Appendix for completion. It is based on Levy's lemma, i.e., concentration of measure on the real-unit sphere $\mathbb{S}^{2d-1} \subset \mathbb{R}^{2d}$. A function $f : \mathbb{S}^{2d-1} \to \mathbb{R}$ is L-Lipschitz (with respect to the Euclidean norm $\|\cdot\|_{\ell_2}$ on $\mathbb{R}^{2d}$) if

$$|f(x) - f(y)| \le L\|x-y\|_{\ell_2} \quad \text{for all } x,y \in \mathbb{S}^{2d-1}. \tag{D2}$$

Theorem 30 (Levy's lemma): *Let $f : \mathbb{S}^{2d-1} \to \mathbb{R}$ be a L-Lipschitz function on the unit sphere. Then, the following relation is true if x is chosen uniformly from $\mathbb{S}^{2d-1}$:*

$$Pr\{|f(x) - \mathbb{E}[f(x)]| \ge \tau\} \le 2\exp\left(-\frac{4d\tau^2}{9\pi^3 L^2}\right). \tag{D3}$$

Proof of Proposition 29. The complex unit sphere in d dimensions admits an isometric embedding—with respect to the Euclidean norm—onto the real-valued unit sphere $\mathbb{S}^{2d-1} \subseteq \mathbb{R}^{2d}$:

$$|\psi\rangle \mapsto |x\rangle = \mathrm{Re}(|\psi\rangle) \oplus \mathrm{Im}(|\psi\rangle) \in \mathbb{S}^{2d-1}. \tag{D4}$$

This embedding maps the uniform distribution on the complex unit sphere in $\mathbb{C}^d$ to the uniform distribution on the real-valued unit sphere in $\mathbb{R}^{2d}$. Under this embedding, the function of interest $\langle\psi|M|\psi\rangle$ becomes

$$\begin{aligned}\langle\psi|M|\psi\rangle &= \langle \mathrm{Re}(\psi)|M|\mathrm{Re}(\psi)\rangle \\ &\quad + \langle \mathrm{Im}(\psi)|M|\mathrm{Im}(\psi)\rangle = \langle x|M\oplus M|x\rangle, \end{aligned} \tag{D5}$$

because M is Hermitian. Its expectation is also preserved and Lemma 31 immediately below states that this function

is Lipschitz with constant $2\|M\|_\infty \leq 2$. The claim then readily follows from Levy's lemma (Theorem 30). ■

Lemma 31: *Fix $M \in \mathbb{H}_d$. Then, the following relation is true for any pair of unit-norm vectors $x, y \in \mathbb{S}^{2d-1} \subset \mathbb{R}^{2d}$*

$$|\langle x|M \oplus M|x\rangle - \langle y|M \oplus M|y\rangle| \leq 2\|M\|_\infty \|x - y\|_{\ell_2}. \tag{D6}$$

Proof. Fix $x, y \in \mathbb{S}^{2d-1}$ and apply Hoelder's inequality:

$$\begin{aligned} &|\langle x|M \oplus M|x\rangle - \langle y|M \oplus M|y\rangle|^2 \\ &\quad = \mathrm{Tr}\left[M \oplus M(|x\rangle\langle x| - |y\rangle\langle y|)\right]^2 \\ &\quad \leq \|M \oplus M\|_\infty^2 \||x\rangle\langle x| - |y\rangle\langle y|\|_1^2. \end{aligned} \tag{D7}$$

The block structure of $M \oplus M$ ensures $\|M \oplus M\|_\infty = \|M\|_\infty$, while the remaining term is the trace norm of a difference of pure states. This can be computed analytically and we obtain

$$\begin{aligned} &\||x\rangle\langle x| - |y\rangle\langle y|\|_1^2 \\ &\quad = 4\left(1 - \langle x, y\rangle^2\right) = 4\left(1 + \langle x, y\rangle\right)\left(1 - \langle x, y\rangle\right) \\ &\quad \leq 4\left(2 - 2|\langle x, y\rangle|\right), \end{aligned} \tag{D8}$$

because $\langle x, y\rangle \leq \|x\|_{\ell_2}\|y\|_{\ell_2} \leq 1$ Finally,

$$\begin{aligned} 2 - 2\langle x, y\rangle &= \langle x, x\rangle - \langle x, y\rangle - \langle y, x\rangle + \langle y, y\rangle \\ &= \langle x - y, x - y\rangle = \|x - y\|_{\ell_2}^2, \end{aligned} \tag{D9}$$

and the claim follows. ■

APPENDIX E: DESIGNS AND THE TRADITIONAL DEFINITION OF COMPLEXITY

In the bulk of the paper we focus on a stronger notion of complexity than the standard definition, an operational definition involving the complexity of the distinguishing measurement to differentiate the state from the maximally mixed state. A more traditional definition is often considered in the literature, which involves building a quantum circuit that approximates the state when evolved from an initial state. This intuitive notion of complexity is related to the minimal size of such a circuit.

In this Appendix, we work through the counting arguments in Appendix A for the complexity of elements of a k-design using the more traditional (albeit weaker) definition of complexity. We refer to this as the *weak complexity* of a state or unitary to distinguish it from the operational definitions presented in Sec. II A.

Consider a system of n qudits with local dimension q, such that the total dimension is $d = q^n$. Let $\mathsf{G} \subset U(q^2)$ denote a universal gate set of elementary 2-local gates, and let G_r be the set of circuits of size r built from our gate set G.

Definition 5 (Weak δ-state complexity): *For $\delta \in [0, 1]$, we say that a state $|\psi\rangle$ has δ-state complexity of at most r if there exists a unitary circuit $V \in \mathsf{G}_r$ such that*

$$\frac{1}{2}\left\||\psi\rangle\langle\psi| - V|0\rangle\langle 0|V^\dagger\right\|_1 \leq \delta, \quad \textit{which we denote as} \quad \times \mathcal{C}'_\delta(|\psi\rangle)) \leq r.$$

We want to be able to make precise statements about the complexity of sets of states. More specifically, if we consider a complex projective design, the requirement that they form a k-design is sufficiently restrictive to deduce a quantitative statement about the complexity of the constituent states.

Theorem 32 (Weak complexity of state designs): *Consider an ϵ-approximate complex projective k-design $\mathcal{E} = \{p_i, |\psi_i\rangle\}_{i=1}^N$. Then there are at least*

$$\frac{d^k}{k!}\frac{1}{1+\epsilon} - \frac{n^r|\mathsf{G}|^r}{(1-\delta^2)^k}, \tag{E1}$$

states with weak δ-state complexity $\mathcal{C}'_\delta(|\psi_i\rangle)) > r$.

The number of high complexity states is exponentially large in k for complexity

$$r \lesssim \frac{k(n - \log k)}{\log n}. \tag{E2}$$

Turning now to the complexity of unitaries, the traditional definition of complexity is the minimal size of a circuit, built from our gate set, which approximates that unitary.

Definition 6 (Weak δ-unitary complexity): *For $\delta \in [0, 1]$, we say that a unitary U has δ-unitary complexity of at most r if there exists a circuit $V \in \mathsf{G}_r$ such that*

$$\frac{1}{2}\left\|\mathcal{U} - \mathcal{V}\right\|_\diamond \leq \delta, \quad \textit{which we denote as } \mathcal{C}'_\delta(U) \leq r,$$

where $\mathcal{U}(\rho) = U\rho U^\dagger$ and $\mathcal{V}(\rho) = V\rho V^\dagger$.

Again, we ask if the structure of a unitary k-design allows us to conclude anything about the complexity of unitaries. Once more, we find that we can turn the statement that k-design elements have a certain expected complexity into a quantitative one.

Theorem 33 (Weak complexity of unitary designs): *Consider an ϵ-approximate unitary k-design $\mathcal{E} = \{p_i, U_i\}_{i=1}^N$. Then there are at least*

$$\frac{d^{2k}}{k!}\frac{1}{1+\epsilon} - \frac{n^r|\mathsf{G}|^r}{(1-\delta^2)^k}, \tag{E3}$$

unitaries in $\mathcal{E}$ with weak δ-unitary complexity $\mathcal{C}'_\delta(U_i) > r$.

The number of high-complexity unitaries is again exponentially large in k for complexity less than

$$r \lesssim \frac{k(2n - \log k)}{\log n}. \tag{E4}$$

We now provide details and proofs of the above statements about the complexity of spherical and unitary designs.

1. Weak state complexity for spherical designs

Proof of Theorem 32. First, as stated in Lemma 6, we note that the definition of weak δ-state complexity in Definition 5 is equivalently written as

$$|\langle\psi|V|0\rangle|^2 \geq 1 - \delta^2. \tag{E5}$$

We can show this by first noting that $X := |\psi\rangle\langle\psi| - V|0\rangle\langle 0|V^\dagger$ has rank at most two. Directly computing the eigenvalues of X from

$$\begin{aligned} \mathrm{Tr}(X) &= \lambda_1 + \lambda_2 = 0 \quad \text{and} \\ \mathrm{Tr}(X^2) &= \lambda_1^2 + \lambda_2^2 = 2 - 2|\langle\psi|V|0\rangle|^2, \end{aligned} \tag{E6}$$

we find $\lambda_{1,2} = \pm\sqrt{1 - |\langle\psi|V|0\rangle|^2}$. Then as $\|X\|_1 = |\lambda_1| + |\lambda_2|$ we have that

$$\frac{1}{2}\left\||\psi\rangle\langle\psi| - V|0\rangle\langle 0|V^\dagger\right\|_1 = \sqrt{1 - |\langle\psi|V|0\rangle|^2}, \tag{E7}$$

from which the claim follows.

We want to ask, given some state $|\psi\rangle$ chosen uniformly from an ϵ-approximate spherical k-design, what is the probability that the state has δ complexity at most r: $\mathcal{C}'_\delta(|\psi\rangle) \leq r$? We know that the state will have δ complexity r if there exists a $V \in \mathsf{G}_r$ such that Eq. (E5) holds. A union bound then gives that

$$\begin{aligned} \Pr\left[\mathcal{C}'_\delta(|\psi\rangle) \leq r\right] &= \Pr\left[\bigcup_{V\in\mathsf{G}_r}\left\{|\langle\psi|V|0\rangle|^2 \geq 1-\delta^2\right\}\right] \\ &\leq \sum_{V\in\mathsf{G}_r}\Pr\left[|\langle\psi|V|0\rangle|^2 \geq 1-\delta^2\right]. \end{aligned} \tag{E8}$$

We can bound the probability that a state drawn from a spherical k-design satisfies Eq. (E5) as a straightforward consequence of Markov's inequality:

$$\begin{aligned} &\Pr\left[|\langle\psi|V|0\rangle|^2 \geq 1-\delta^2\right] \\ &\quad= \Pr\left[|\langle\psi|V|0\rangle|^{2k} \geq \left(1-\delta^2\right)^k\right] \\ &\quad\leq \frac{\mathbb{E}_{|\psi\rangle}\left[|\langle\psi|V|0\rangle|^{2k}\right]}{(1-\delta^2)^k} \leq \frac{(1+\epsilon)\binom{d+k-1}{k}^{-1}}{(1-\delta^2)^k}. \end{aligned} \tag{E9}$$

In the last step here, we use Eq. (C43) and proceeding similarly as in the proof of Lemma 21 in Appendix C 6, noting that for a fixed state $|\phi\rangle$ and $|\psi\rangle$ averaged over an ϵ-approximate spherical k-design, we have

$$\mathbb{E}_{|\psi\rangle}\left[|\langle\psi|\phi\rangle|^{2k}\right] \leq (1+\epsilon)\binom{d+k-1}{k}^{-1}. \tag{E10}$$

This claim readily follows from an argument similar to the proof of Lemma 21. Returning to Eq. (E8), we find that the probability that a state in a spherical design has complexity of at most r is

$$\Pr\left[\mathcal{C}'_\delta(|\psi\rangle) \leq r\right] \leq (1+\epsilon)\binom{d+k-1}{k}^{-1}\frac{n^r|\mathsf{G}|^r}{(1-\delta^2)^k}, \tag{E11}$$

using the bound on the expectation and a bound on the cardinality of G_r.

We now turn to proving the primary claim. Negating the above assertion implies that

$$\Pr\left[\mathcal{C}'_\delta(|\psi\rangle) > r\right] \geq 1 - (1+\epsilon)\binom{d+k-1}{k}^{-1}\frac{n^r|\mathsf{G}|^r}{(1-\delta^2)^k}. \tag{E12}$$

Furthermore, we may also write this probability as the expectation of the associated event, which yields

$$\begin{aligned} \Pr\left[\mathcal{C}'_\delta(|\psi\rangle) > r\right] &= \mathbb{E}_{|\psi\rangle}\left[\mathbb{1}\left\{\mathcal{C}'_\delta(|\psi\rangle) > r\right\}\right] \\ &= \sum_i p_i\,\mathbb{1}\left\{\mathcal{C}'_\delta(|\psi_i\rangle) > r\right\} \\ &\leq (1+\epsilon)\binom{d+k-1}{k}^{-1} N, \end{aligned} \tag{E13}$$

where $\mathbb{1}$ is the indicator function, and in the last step we use the bound on the weights of an ϵ-approximate spherical k-design in Lemma 21. N denotes the number of states in the spherical design $|\psi_i\rangle$ with weak δ complexity greater than r. Combining the previous two equations, we find that

$$N \geq \frac{d^k}{k!}\frac{1}{1+\epsilon} - \frac{n^r|\mathsf{G}|^r}{(1-\delta^2)^k}, \tag{E14}$$

which completes the proof. ■

2. Weak unitary complexity for unitary designs

Proof of Theorem 33. We start by noting an equivalent definition of weak δ-unitary complexity as shown in the proof of Lemma 7. A necessary, but in general not sufficient, condition for weak unitary complexity in Definition 6 is

$$\left|\mathrm{Tr}(V^\dagger U)\right|^2 \geq d^2(1-\delta^2). \tag{E15}$$

Now we again ask, given some unitary U chosen uniformly from an ϵ-approximate unitary k-design, what is the probability that it has δ-unitary complexity at most r: $\mathcal{C}'_\delta(U) \leq r$? As this holds if there exists a $V \in \mathsf{G}_r$ such that the channels are close in diamond distance, a union bound then gives that

$$\begin{aligned}\Pr\left[\mathcal{C}'_\delta(U) \leq r\right] &= \Pr\left[\bigcup_{V\in G_r}\left\{\frac{1}{2}\left\|\mathcal{U}-\mathcal{V}\right\|_\diamond \leq \delta\right\}\right] \\ &\leq \sum_{V\in G_r}\Pr\left[\left|\mathrm{Tr}(V^\dagger U)\right|^2 \geq d^2(1-\delta^2)\right],\end{aligned} \tag{E16}$$

using the reformulation above. We can bound the probability that a unitary drawn from a k-design satisfies this condition again by using Markov's inequality:

$$\begin{aligned}&\Pr\left[\left|\mathrm{Tr}(V^\dagger U)\right|^2 \geq d^2(1-\delta^2)\right] \\ &= \Pr\left[\left|\mathrm{Tr}(V^\dagger U)\right|^{2k} \geq d^{2k}(1-\delta^2)^k\right] \\ &\leq \frac{\mathbb{E}_{\mathcal{E}}\left[\left|\mathrm{Tr}(V^\dagger U)\right|^{2k}\right]}{d^{2k}(1-\delta^2)^k} \leq \frac{(1+\epsilon)k!}{d^{2k}(1-\delta^2)^k},\end{aligned} \tag{E17}$$

where in the last step, we use the moments of traces for unitary designs and as in Lemma 20 in Appendix C 6 above find that for a fixed unitary V and a unitary U averaged over an ϵ-approximate unitary k-design, we have

$$\mathbb{E}_{\mathcal{E}}\left[\left|\mathrm{Tr}(V^\dagger U)\right|^{2k}\right] \leq (1+\epsilon)k!. \tag{E18}$$

Returning to the expression above in Eq. (E16), we find that the probability $\mathcal{C}'_\delta(U) \leq r$ is

$$\Pr\left[\mathcal{C}'_\delta(U) \leq r\right] \leq (1+\epsilon)\frac{k!}{d^{2k}}\frac{n^r|\mathsf{G}|^r}{(1-\delta^2)^k}, \tag{E19}$$

using the bound on the expectation and a bound on the cardinality of G_r. Negating the expression gives a lower bound on the probability that a unitary in a k-design has complexity greater than r. Furthermore, we may also write this probability as the expectation

$$\Pr\left[\mathcal{C}'_\delta(U) > r\right] = \sum_i p_i \mathbb{1}\left\{\mathcal{C}'_\delta(U_i) > r\right\} \leq (1+\epsilon)\frac{k!}{d^{2k}}N, \tag{E20}$$

where we use the bound on the unitary design weights in Lemma 20. N denotes the number of untiaries in a k-design with weak δ complexity greater than r. Combining the previous two equations, we find that

$$N \geq \frac{d^{2k}}{k!}\frac{1}{1+\epsilon} - \frac{n^r|\mathsf{G}|^r}{(1-\delta^2)^k}, \tag{E21}$$

which completes the proof. ■

[1] D. Poulin, A. Qarry, R. Somma, and F. Verstraete, Quantum Simulation of Time-Dependent Hamiltonians and the Convenient Illusion of Hilbert Space, Phys. Rev. Lett. **106**, 170501 (2011).

[2] E. Bernstein and U. Vazirani, Quantum complexity theory, SIAM J. Comput. **26**, 1411 (1997).

[3] X. Chen, Z. C. Gu, and X. G. Wen, Local unitary transformation, long-range quantum entanglement, wave function renormalization, and topological order, Phys. Rev. **B82**, 155138 (2010).

[4] L. Susskind, Computational complexity and black hole horizons, Fortsch. Phys. **64**, 44 (2016), [Fortsch. Phys. **64**, 24 (2016)].

[5] D. Stanford and L. Susskind, Complexity and shock wave geometries, Phys. Rev. **D90**, 126007 (2014).

[6] A. R. Brown, D. A. Roberts, L. Susskind, B. Swingle, and Y. Zhao, Complexity, action, and black holes, Phys. Rev. **D93**, 086006 (2016).

[7] A. R. Brown and L. Susskind, Second law of quantum complexity, Phys. Rev. **D97**, 086015 (2018).

[8] L. Susskind, Black holes and complexity classes, ArXiv:1802.02175.

[9] S. Aaronson, The complexity of quantum states and transformations: From quantum money to black holes, ArXiv:1607.05256.

[10] T. C. Bohdanowicz and F. G. S. L. Brandão, Universal Hamiltonians for exponentially long simulation, ArXiv:1710.02625.

[11] D. A. Roberts and B. Yoshida, Chaos and complexity by design, JHEP **04**, 121 (2017).

[12] F. G. S. L. Brandão, A. W. Harrow, and M. Horodecki, Local random quantum circuits are approximate polynomial-designs, Commun. Math. Phys. **346**, 397 (2016).

[13] C. Dankert, R. Cleve, J. Emerson, and E. Livine, Exact and approximate unitary 2-designs and their application to fidelity estimation, Phys. Rev. **A80**, 012304 (2009).

[14] D. Gross, K. Audenaert, and J. Eisert, Evenly distributed unitaries: On the structure of unitary designs, J. Math. Phys. **48**, 052104 (2007).

[15] Z. Webb, The clifford group forms a unitary 3-design, Quantum Info. Comput. **16**, 1379 (2016).

[16] H. Zhu, Multiqubit clifford groups are unitary 3-designs, Phys. Rev. A **96**, 062336 (2017).

[17] R. Kueng and D. Gross, Qubit stabilizer states are complex projective 3-designs, ArXiv:1510.02767.

[18] A. Ambainis and J. Emerson, in *Twenty-Second Annual IEEE Conference on Computational Complexity (CCC'07)* (2007), p. 129.

[19] O. Szehr, F. Dupuis, M. Tomamichel, and R. Renner, Decoupling with unitary approximate two-designs, New J. Phys. **15**, 053022 (2013).

[20] A. J. Scott, Tight informationally complete quantum measurements, J. Phys. A: Math. Gen. **39**, 13507 (2006).

[21] R. Kueng, H. Rauhut, and U. Terstiege, Low rank matrix recovery from rank one measurements, Appl. Comput. Harmon. Anal. **42**, 88 (2017).

[22] J. Emerson, R. Alicki, and K. Życzkowski, Scalable noise estimation with random unitary operators, J. Opt. B: Quantum Semiclass. Opt **7**, S347 (2005).

[23] P. Hayden and J. Preskill, Black holes as mirrors: Quantum information in random subsystems, JHEP **09**, 120 (2007).

[24] Note that here we discuss the size of the circuit, if we parallelize the application of gates, the depth of the circuit required to form an approximate design scales linearly in n.

[25] N. Lashkari, D. Stanford, M. Hastings, T. Osborne, and P. Hayden, Towards the fast scrambling conjecture, JHEP **04**, 022 (2013).

[26] E. Onorati, O. Buerschaper, M. Kliesch, W. Brown, A. H. Werner, and J. Eisert, Mixing properties of stochastic quantum Hamiltonians, Commun. Math. Phys. **355**, 905 (2017).

[27] Y. Nakata, C. Hirche, M. Koashi, and A. Winter, Efficient quantum pseudorandomness with nearly time-independent Hamiltonian dynamics, Phys. Rev. **X7**, 021006 (2017).

[28] N. Hunter-Jones, Unitary designs from statistical mechanics in random quantum circuits, ArXiv:1905.12053.

[29] J. Haferkamp and N. Hunter-Jones, Improved spectral gaps for random quantum circuits: Large local dimensions and all-to-all interactions, ArXiv:2012.05259.

[30] A. S. Holevo, Optimal quantum measurements, Teoret. Mat. Fiz. **17**, 319 (1973).

[31] C. W. Helstrom, *Quantum Detection and Estimation Theory*, Mathematics in Science and Engineering (Academic Press, New York, NY, 1976).

[32] C. M. Dawson and M. A. Nielsen, The Solovay-Kitaev algorithm, Quantum Info. Comput. **6**, 81 (2006).

[33] J. Watrous, *The Theory of Quantum Information* (Cambridge University Press, Cambridge, 2018).

[34] For $q = 2$ a depth-two circuit comprised of n Hadamard gates and n CNOTs suffices.

[35] A. Harrow and S. Mehraban, Approximate unitary t-designs by short random quantum circuits using nearest-neighbor and long-range gates, ArXiv:1809.06957.

[36] A. W. Harrow and R. A. Low, Efficient quantum tensor product expanders and k-designs, Lect. Notes Comput. Sci. **5687**, 548 (2009).

[37] W. Fulton and J. Harris, *Representation Theory: A First Course*, Graduate Texts in Mathematics (Springer, New York, 1991).

[38] M. Christandl, PhD thesis, University of Cambridge, 2006.

[39] D. Weingarten, Asymptotic behavior of group integrals in the limit of infinite rank, J. Math. Phys. **19**, 999 (1978).

[40] B. Collins and P. Śniady, Integration with respect to the haar measure on unitary, orthogonal and symplectic group, Commun. Math. Phys. **264**, 773 (2006).

[41] J. C. Bridgeman and C. T. Chubb, Hand-waving and interpretive dance: An introductory course on tensor networks, J. Phys. **A50**, 223001 (2017).

[42] M. Kliesch, R. Kueng, J. Eisert, and D. Gross, Guaranteed recovery of quantum processes from few measurements, Quantum **3**, 171 (2019).

[43] R. T. Rockafellar, *Convex Analysis*, Princeton Mathematical Series, Vol. 28 (Princeton University Press, Princeton, NJ, 1970).

[44] A. Barvinok, *A Course in Convexity*, Graduate Studies in Mathematics, Vol. 54 (American Mathematical Society, Providence, RI, 2002).

[45] J. Cotler, N. Hunter-Jones, J. Liu, and B. Yoshida, Chaos, complexity, and random matrices, JHEP **11**, 048 (2017).

[46] D. Gross, S. T. Flammia, and J. Eisert, Most Quantum States are too Entangled to be Useful as Computational Resources, Phys. Rev. Lett. **102**, 190501 (2009).

[47] A. Bouland, B. Fefferman, and U. Vazirani, Computational pseudorandomness, the wormhole growth paradox, and constraints on the AdS/CFT duality, ArXiv:1910.14646.

[48] Z. Ji, Y.-K. Liu, and F. Song, in *Advances in Cryptology—CRYPTO 2018* (Springer, 2018), p. 126.

[49] In addition to containing inverses, Ref. [12] also required that the gate set G be comprised of algebraic entries, but recent results suggest that both these restrictions may be relaxed [83,84].

[50] Recently, Ref. [35] showed that higher-dimensional local random quantum circuits form approximate designs in $O[n^{1/D}\mathrm{poly}(k)]$ depth, with some (high-degree) polynomial dependence on k. Theorem 9 then gives a polynomial growth of complexity for these higher-dimensional circuits.

[51] We note that Ref. [28] computed the circuit depth, whereas the discussion here involves the circuit size, giving an extra factor of n.

[52] A. Nahum, S. Vijay, and J. Haah, Operator spreading in random unitary circuits, Phys. Rev. **X8**, 021014 (2018).

[53] T. Zhou and A. Nahum, Emergent statistical mechanics of entanglement in random unitary circuits, Phys. Rev. **B99**, 174205 (2019).

[54] J. Bourgain and A. Gamburd, A spectral gap theorem in SU(d), J. Eur. Math. Soc. **14**, 1455 (2012).

[55] J. Cotler and N. Hunter-Jones, Spectral decoupling in many-body quantum chaos, JHEP **12**, 205 (2020).

[56] L. Susskind, Entanglement is not enough, Fortsch. Phys. **64**, 49 (2016).

[57] A. R. Brown, D. A. Roberts, L. Susskind, B. Swingle, and Y. Zhao, Holographic Complexity Equals Bulk Action? Phys. Rev. Lett. **116**, 191301 (2016).

[58] S. Chapman, H. Marrochio, and R. C. Myers, Complexity of formation in holography, JHEP **01**, 062 (2017).

[59] D. Carmi, R. C. Myers, and P. Rath, Comments on holographic complexity, JHEP **03**, 118 (2017).

[60] M. Alishahiha, Holographic complexity, Phys. Rev. **D92**, 126009 (2015).

[61] D. Carmi, S. Chapman, H. Marrochio, R. C. Myers, and S. Sugishita, On the time dependence of holographic complexity, JHEP **11**, 188 (2017).
[62] P. Caputa, N. Kundu, M. Miyaji, T. Takayanagi, and K. Watanabe, Liouville action as path-integral complexity: From continuous tensor networks to AdS/CFT, JHEP **11**, 097 (2017).
[63] C. A. Agón, M. Headrick, and B. Swingle, Subsystem complexity and holography, JHEP **02**, 145 (2019).
[64] K. Goto, H. Marrochio, R. C. Myers, L. Queimada, and B. Yoshida, Holographic complexity equals which action? JHEP **02**, 160 (2019).
[65] Z.-W. Liu, S. Lloyd, E. Y. Zhu, and H. Zhu, Entanglement, quantum randomness, and complexity beyond scrambling, JHEP **07**, 041 (2018).
[66] To see this, recall the relation between Schatten α-norms in d dimensions: $\|\rho\|_\infty \leq \|\rho\|_\alpha \leq d^{1/\alpha}\|\rho\|_\infty$. This ensures that for any state ρ, we have $S_{\min}(\rho) \leq S^{(\alpha)}(\rho) \leq S_{\min}(\rho) + (\log d/\alpha)$. Note that as we take α to be greater than n, these Rényi entropies concentrate ever sharper around the min-entropy.
[67] N. Alon and J. H. Spencer, *The Probabilistic Method*, Wiley Series in Discrete Mathematics and Optimization (John Wiley & Sons, Hoboken, NJ, 2016), 4th ed.
[68] R. Kueng, Quantum and classical information processes with tensors (lecture notes), Spring, 2019. Caltech course notes: https://iqim.caltech.edu/classes.
[69] S. Boyd and L. Vandenberghe, *Convex Optimization* (Cambridge University Press, Cambridge, 2004).
[70] A. Y. Kitaev, Quantum computations: Algorithms and error correction, Russ. Math. Surv. **52**, 1191 (1997).
[71] J. Watrous, Semidefinite programs for completely bounded norms, Theory Comput. **5**, 217 (2009).
[72] A. Ben-Aroya and A. Ta-Shma, On the complexity of approximating the diamond norm, Quantum Info. Comput. **10**, 77 (2010).
[73] J. Watrous, Simpler semidefinite programs for completely bounded norms, Chic. J. Theoret. Comput. Sci. **8**, 1 (2013).
[74] M. Kliesch, R. Kueng, J. Eisert, and D. Gross, Improving compressed sensing with the diamond norm, IEEE Trans. Inf. Theory **62**, 7445 (2016).
[75] U. Michel, M. Kliesch, R. Kueng, and D. Gross, Comments on improving compressed sensing with the diamond norm—saturation of the norm inequalities between diamond and nuclear norm, IEEE Trans. Inf. Theory **64**, 7443 (2018).
[76] V. Paulsen, *Completely Bounded Maps and Operator Algebras*, Cambridge Studies in Advanced Mathematics, Vol. 78 (Cambridge University Press, Cambridge, 2002).
[77] R. Bhatia, *Matrix Analysis*, Graduate Texts in Mathematics, Vol. 169 (Springer-Verlag, New York, 1997).
[78] Technically, this is only true for bending lines an even number of times. A single bend corresponds to transposition, which is basis dependent and not equivalent to conjugation. This subtlety, however, will rarely feature in our arguments.
[79] B. Collins, Moments and cumulants of polynomial random variables on unitary groups, the Itzykson-Zuber integral, and free probability, Int. Math. Res. Not. **2003**, 953 (2003).
[80] J. M. Renes, R. Blume-Kohout, A. J. Scott, and C. M. Caves, Symmetric informationally complete quantum measurements, J. Math. Phys. **45**, 2171 (2004).
[81] B. Collins and S. Matsumoto, Weingarten calculus via orthogonality relations: New applications, Lat. Am. J. Probab. Math. Stat. **14**, 631 (2017).
[82] A. Montanaro, Weak multiplicativity for random quantum channels, Commun. Math. Phys. **319**, 535 (2013).
[83] R. Mezher, J. Ghalbouni, J. Dgheim, and D. Markham, Unitary t-designs from *relaxed* seeds, ArXiv:1911.03704.
[84] M. Oszmaniec, A. Sawicki, and M. Horodecki, Epsilon-nets, unitary designs and random quantum circuits, ArXiv:2007.10885.

Chapter 5

Improving near-term quantum algorithms by derandomization

or: Efficient estimation of Pauli observables by derandomization

Abstract

We consider the problem of jointly estimating expectation values of many Pauli observables, a crucial subroutine in variational quantum algorithms. Starting with randomized measurements, we propose an efficient derandomization procedure that iteratively replaces random single-qubit measurements by fixed Pauli measurements; the resulting deterministic measurement procedure is guaranteed to perform at least as well as the randomized one. In particular, for estimating any L low-weight Pauli observables, a deterministic measurement on only of order $\log(L)$ copies of a quantum state suffices. In some cases, for example, when some of the Pauli observables have high weight, the derandomized procedure is substantially better than the randomized one. Specifically, numerical experiments highlight the advantages of our derandomized protocol over various previous methods for estimating the ground-state energies of small molecules.

Authors

Hsin-Yuan (Robert) Huang, Richard Kueng, John Preskill.

Journal

Physical Review Letters **127**:030503 (2021).

Confirmation of declaration of author contributions (Hsin-Yuan Huang)

Publication:

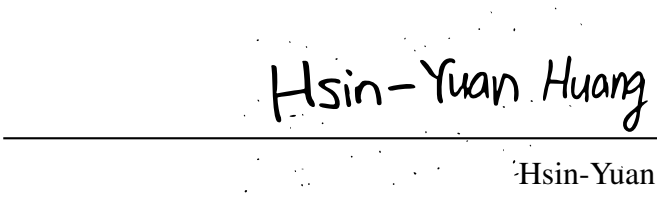
H.Y. Huang, R. Kueng, J. Preskill, Efficient estimation of Pauli observables by derandomization, *Physical Review Letters* **127**:030503 (2021)

Declaration of author contributions:

Hsin-Yuan Huang and Richard Kueng developed the theoretical aspects of this work. Hsin-Yuan Huang conducted the numerical experiments and wrote the open-source code. John Preskill provided guidance and conceived the applications. All authors wrote the manuscript.

Confirmation by co-author:

I confirm this declaration of author contributions, as well as my co-authorship.



Hsin-Yuan Huang

Confirmation of declaration of author contributions (John Preskill)

Publication:

H.Y. Huang, R. Kueng, J. Preskill, Efficient estimation of Pauli observables by derandomization, *Physical Review Letters* **127**:030503 (2021)

Declaration of author contributions:

Hsin-Yuan Huang and Richard Kueng developed the theoretical aspects of this work. Hsin-Yuan Huang conducted the numerical experiments and wrote the open-source code. John Preskill provided guidance and conceived the applications. All authors wrote the manuscript.

Confirmation by co-author:

I confirm this declaration of author contributions, as well as my co-authorship.

John Preskill

Efficient Estimation of Pauli Observables by Derandomization

Hsin-Yuan Huang,[1,2,*] Richard Kueng,[3] and John Preskill[1,2,4,5]
[1]Institute for Quantum Information and Matter, Caltech, Pasadena, California 91125, USA
[2]Department of Computing and Mathematical Sciences, Caltech, Pasadena, California 91125, USA
[3]Institute for Integrated Circuits, Johannes Kepler University Linz, A-4040, Austria
[4]Walter Burke Institute for Theoretical Physics, Caltech, Pasadena, California 91125, USA
[5]AWS Center for Quantum Computing, Pasadena, California 91125, USA

(Received 19 March 2021; accepted 14 June 2021; published 16 July 2021)

We consider the problem of jointly estimating expectation values of many Pauli observables, a crucial subroutine in variational quantum algorithms. Starting with randomized measurements, we propose an efficient derandomization procedure that iteratively replaces random single-qubit measurements by fixed Pauli measurements; the resulting deterministic measurement procedure is guaranteed to perform at least as well as the randomized one. In particular, for estimating any L low-weight Pauli observables, a deterministic measurement on only of order $\log(L)$ copies of a quantum state suffices. In some cases, for example, when some of the Pauli observables have high weight, the derandomized procedure is substantially better than the randomized one. Specifically, numerical experiments highlight the advantages of our derandomized protocol over various previous methods for estimating the ground-state energies of small molecules.

DOI: 10.1103/PhysRevLett.127.030503

Introduction.—Noisy intermediate-scale quantum (NISQ) devices are becoming available [1]. Though less powerful than fully error-corrected quantum computers, NISQ devices used as coprocessors might have advantages over classical computers for solving some problems of practical interest. For example, variational algorithms using NISQ hardware have potential applications to chemistry, materials science, and optimization [2–10].

In a typical NISQ variational algorithm, we need to estimate expectation values for a specified set of operators $\{O_1, O_2, \ldots, O_L\}$ in a quantum state ρ that can be prepared repeatedly using a programmable quantum system. To obtain precise estimates, each operator must be measured many times, and finding a reasonably efficient procedure for extracting the desired information is not easy in general. In this Letter, we consider the special case where each O_j is a Pauli operator; this case is of particular interest for near-term applications.

Suppose we have quantum hardware that produces multiple copies of the n-qubit state ρ. Furthermore, for every copy, we can measure all the qubits independently, choosing at our discretion to measure each qubit in the X, Y, or Z basis. We are given a list of L n-qubit Pauli operators (each one a tensor product of n Pauli matrices), and our task is to estimate the expectation values of all L operators in the state ρ, with an additive error no larger than ε for each operator. We would like to perform this task using as few copies of ρ as possible.

If all L Pauli operators have relatively low weight (act nontrivially on only a few qubits), there is a simple randomized protocol that achieves our goal quite efficiently: For each of M copies of ρ, and for each of the n qubits, we chose uniformly at random to measure X, Y, or Z. Then we can achieve the desired prediction accuracy with high success probability if $M = O(3^w \log L/\varepsilon^2)$, assuming that all L operators on our list have weight no larger than w [11,12]. If the list contains high-weight operators, however, this randomized method is not likely to succeed unless M is very large.

In this Letter, we describe a deterministic protocol for estimating Pauli-operator expectation values that always performs at least as well as the randomized protocol and performs much better in some cases. This deterministic protocol is constructed by *derandomizing* the randomized protocol. The key observation is that we can compute a lower bound on the probability that randomized measurements on M copies successfully achieve the desired error ε for every one of our L target Pauli operators. Furthermore, we can compute this lower bound even when the measurement protocol is partially deterministic and partially randomized; that is, when some of the measured single-qubit Pauli operators are fixed, and others are still sampled uniformly from $\{X, Y, Z\}$.

Hence, starting with the fully randomized protocol, we can proceed step by step to replace each randomized single-qubit measurement by a deterministic one, taking care in each step to ensure that the new partially randomized protocol, with one additional fixed measurement, has success probability at least as high as the preceding protocol. When all measurements have been fixed, we have a fully

deterministic protocol. In numerical experiments, we find that this deterministic protocol substantially outperforms randomized protocols [12–16]. The improvement is especially significant when the list of target observables includes operators with relatively high weight. Further performance gains are possible by executing (at least) linear-depth circuits before measurements [17–20]. Such procedures do, however, require deep quantum circuits. In contrast, our protocol only requires single-qubit Pauli measurements, which are more amenable to execution on near-term devices.

The manuscript is organized as follows. We first provide some statistical background, explain the randomized measurement protocol, then analyze the derandomization procedure. We then provide numerical results showing how the derandomized protocol improves on previous methods. We conclude with remarks and outlooks. Further examples and details of proofs are in the Supplemental Material [21].

Statistical background.—Let ρ be a fixed, but unknown, quantum state on n qubits. We want to accurately predict L expectation values

$$\omega_\ell(\rho) = \mathrm{tr}(O_{\mathbf{o}_\ell}\rho) \quad \text{for } 1 \le \ell \le L, \tag{1}$$

where each $O_{\mathbf{o}_\ell} = \sigma_{\mathbf{o}_\ell[1]} \otimes \cdots \otimes \sigma_{\mathbf{o}_\ell[n]}$ is a tensor product of single-qubit Pauli matrices, i.e., $\mathbf{o}_\ell = [\mathbf{o}_\ell[1], \ldots, \mathbf{o}_\ell[n]]$ with $\mathbf{o}_\ell[k] \in \{I, X, Y, Z\}$. To extract meaningful information, we perform M (single-shot) Pauli measurements on independent copies of ρ. There are 3^n possible measurement choices. Each of them is characterized by a full-weight Pauli string $\mathbf{p}_m \in \{X, Y, Z\}^n$ and produces a random string of n outcome signs $\mathbf{q}_m \in \{\pm 1\}^n$.

Not every Pauli measurement $\mathbf{p}_m$ $(1 \le m \le M)$ provides actionable advice about every target observable $\mathbf{o}_\ell$ $(1 \le \ell \le L)$. The two must be compatible in the sense that the latter corresponds to a marginal of the former; i.e., it is possible to obtain $\mathbf{o}_\ell$ from $\mathbf{p}_m$ by replacing some local nonidentity Pauli matrices with I. If this is the case, we write $\mathbf{o}_\ell \triangleright \mathbf{p}_m$ and say that measurement $\mathbf{p}_m$ "hits" target observable $\mathbf{o}_\ell$. For instance, $[X, I], [I, X], [X, X] \triangleright [X, X]$, but $[Z, I], [I, Z], [Z, Z] \not\triangleright [X, X]$. We can approximate each $\omega_\ell(\rho)$ by empirically averaging (appropriately marginalized) measurement outcomes that belong to Pauli measurements that hit $\mathbf{o}_\ell$,

$$\hat{\omega}_\ell = \frac{1}{h(\mathbf{o}_\ell; [\mathbf{p}_1, \ldots, \mathbf{p}_M])} \sum_{m : \mathbf{o}_\ell \triangleright \mathbf{p}_m} \prod_{j : \mathbf{o}_\ell[j] \neq I} \mathbf{q}_m[j], \tag{2}$$

where $h(\mathbf{o}_\ell; [\mathbf{p}_1, \ldots, \mathbf{p}_M]) = \sum_{m=1}^{M} \mathbf{1}\{\mathbf{o}_\ell \triangleright \mathbf{p}_m\} \in \{0, 1, \ldots, M\}$ counts how many Pauli measurements hit target observable $\mathbf{o}_\ell$.

It is easy to check that each $\hat{\omega}_\ell$ exactly reproduces $\omega_\ell(\rho)$ in expectation [provided that $h(\mathbf{o}_\ell; \mathbf{P}) \ge 1$]. Moreover, the probability of a large deviation improves exponentially with the number of hits.

Lemma 1. (Confidence bound). Fix $\varepsilon \in (0, 1)$ (accuracy) and $1 - \delta \in (0, 1)$ (confidence). Suppose that Pauli observables $\mathbf{O} = [\mathbf{o}_1, \ldots, \mathbf{o}_L]$ and Pauli measurements $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_M]$ are such that

$$\mathrm{Conf}_\varepsilon(\mathbf{O}; \mathbf{P}) := \sum_{\ell=1}^{L} \exp\left(-\frac{\varepsilon^2}{2} h(\mathbf{o}_\ell; \mathbf{P})\right) \le \frac{\delta}{2}. \tag{3}$$

Then, the associated empirical averages (2) obey

$$|\hat{\omega}_\ell - \omega_\ell(\rho)| \le \varepsilon \quad \text{for all } 1 \le \ell \le L \tag{4}$$

with probability (at least) $1 - \delta$.

See Supplemental Material Sec. B.1 for a detailed derivation [21]. We call the function defined in Eq. (3) the "confidence bound." It is a statistically sound summary parameter that checks whether a set of Pauli measurements ($\mathbf{P}$) allows for confidently predicting a collection of Pauli observables ($\mathbf{O}$) up to accuracy ε each.

Randomized Pauli measurements.—Intuitively speaking, a small confidence bound (3) implies a good Pauli estimation protocol. But how should we choose our M Pauli measurements ($\mathbf{P}$) in order to achieve $\mathrm{Conf}_\varepsilon(\mathbf{O}; \mathbf{P}) \le \delta/2$? The randomized measurement toolbox [12,13,16,22,23] provides a perhaps surprising answer to this question. Let $w(\mathbf{o}_\ell)$ denote the *weight* of Pauli observable $\mathbf{o}_\ell$, i.e., the number of qubits on which the observable acts nontrivially: $w(\mathbf{o}_\ell) = \sum_{k=1}^{n} \mathbf{1}\{\mathbf{o}_\ell[k] \neq I\}$. These weights capture the probability of hitting $\mathbf{o}_\ell$ with a completely random measurement string: $\mathrm{Prob}_{\mathbf{p}}[\mathbf{o}_\ell \triangleright \mathbf{p}] = 1/3^{w(\mathbf{o}_\ell)}$. In turn, a total of M randomly selected Pauli measurements will, on average, achieve $\mathbb{E}_{\mathbf{P}}[h(\mathbf{o}_\ell; \mathbf{P})] = M/3^{w(\mathbf{o}_\ell)}$ hits, regardless of the actual Pauli observable $\mathbf{o}_\ell$ in question. This insight allows us to compute expectation values of the confidence bound (3)

$$\mathbb{E}_{\mathbf{P}}[\mathrm{Conf}_\varepsilon(\mathbf{O}; \mathbf{P})] = \sum_{\ell=1}^{L} (1 - \nu/3^{w(\mathbf{o}_\ell)})^M, \tag{5}$$

where $\nu = 1 - \exp(-\varepsilon^2/2) \in (0, 1)$. Each of the L terms is exponentially suppressed in $\varepsilon^2 M/3^{w(\mathbf{o}_\ell)}$. Concrete realizations of a randomized measurement protocol are extremely unlikely to deviate substantially from this expected behavior (see, e.g., [11]). Combined with Lemma 1, this observation implies a powerful error bound.

Theorem 1. (Theorem 3 in Ref. [11].)—Empirical averages (2) obtained from M randomized Pauli measurements allow for ε-accurately predicting L Pauli expectation values $\mathrm{tr}(O_{\mathbf{o}_1}\rho), \ldots, \mathrm{tr}(O_{\mathbf{o}_L}\rho)$ up to additive error ε given that $M \propto \log(L) \max_\ell 3^{w(\mathbf{o}_\ell)}/\varepsilon^2$.

In particular, order $\log(L)$ randomized Pauli measurements suffice for estimating any collection of L low-weight Pauli observables. It is instructive to compare this result

to other powerful statements about randomized measurements, most notably the "classical shadow" paradigm [12,16]. For Pauli observables and Pauli measurements, the two approaches are closely related. The estimators (2) are actually simplified variants of the classical shadow protocol (in particular, they do not require median of means prediction) and the requirements on M are also comparable. This is no coincidence; information-theoretic lower bounds from [12] assert that there are scenarios where the scaling $M \propto \log(L) \max_\ell 3^{w(\mathbf{o}_\ell)}/\varepsilon^2$ is asymptotically optimal and cannot be avoided.

Nevertheless, this does not mean that randomized measurements are *always* a good idea. High-weight observables do pose an immediate challenge, because it is extremely unlikely to hit them by chance alone.

Derandomized Pauli measurements.—The main result of this Letter is a procedure for identifying "good" Pauli measurements that allow for accurately predicting many (fixed) Pauli expectation values. This procedure is designed to interpolate between two extremes: (i) completely randomized measurements (good for predicting many local observables) and (ii) completely deterministic measurements that directly measure observables sequentially (good for predicting few global observables).

Note that we can efficiently compute concrete confidence bounds (3), as well as expected confidence bounds averaged over all possible Pauli measurements (5). Combined, these two formulas also allow us to efficiently compute expected confidence bounds for a list of measurements that is partially deterministic and partially randomized. Suppose that $\mathbf{P}^\sharp$ subsumes deterministic assignments for the first $(m-1)$ Pauli measurements, as well as concrete choices for the first $(k-1)$ Pauli labels of the mth measurement, see Fig. 1 (center). There are three possible choices for the next Pauli assignment: $\mathbf{P}^\sharp[k,m] = W$ with $W = X, Y, Z$. For each choice, we can explicitly compute the resulting conditional expectation value,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{P}}[\mathrm{Conf}_\varepsilon(\mathbf{O};\mathbf{P})|\mathbf{P}^\sharp, \mathbf{P}[k,m] = W] \\
&= \sum_{\ell=1}^{L} \exp\left(-\frac{\varepsilon^2}{2}\sum_{m'=1}^{m-1}\prod_{k'=1}^{n}\mathbf{1}\{\mathbf{o}_\ell[k'] \triangleright \mathbf{P}^\sharp[k',m']\}\right) \\
&\quad\times\left(1-\nu\frac{\mathbf{1}\{\mathbf{o}_\ell[k]\triangleright W\}}{3^{w_{\overline{k}}(\mathbf{o}_\ell)}}\prod_{k'=1}^{k-1}\mathbf{1}\{\mathbf{o}_\ell[k']\triangleright\mathbf{P}^\sharp[k',m]\}\right) \\
&\quad\times(1-\nu 3^{-w(\mathbf{o}_\ell)})^{M-m},
\end{aligned} \tag{6}$$

where $\nu = 1-\exp(-\varepsilon^2/2)$, $w_{\overline{k}}(\mathbf{o}_\ell) = w([\mathbf{o}_\ell[k+1],\ldots,\mathbf{o}_\ell[n]])$ and $\mathbf{o}_\ell[k'] \triangleright \mathbf{P}^\sharp[k',m]$ if $\mathbf{o}_\ell[k'] = I$ or $\mathbf{o}_\ell[k'] = \mathbf{P}^\sharp[k',m]$. This formula allows us to build deterministic measurements one Pauli label at a time.

We start by envisioning a collection of M completely random n-qubit Pauli measurements. That is, each Pauli label is random and Eq. (5) captures the expected confidence bound averaged over *all* 3^{nM} assignments. There are three possible choices for the first label in the first Pauli measurement: $\mathbf{P}[1,1] = X$, $\mathbf{P}[1,1] = Y$, and $\mathbf{P}[1,1] = Z$. At least one concrete choice does not further increase the confidence bound averaged over all remaining Pauli signs,

$$\begin{aligned}
&\min_{W\in\{X,Y,Z\}} \mathbb{E}_{\mathbf{P}}[\mathrm{Conf}_\varepsilon(\mathbf{O};\mathbf{P})|\mathbf{P}[1,1] = W] \\
&\le \frac{1}{3}\sum_{W\in\{X,Y,Z\}} \mathbb{E}_{\mathbf{P}}[\mathrm{Conf}_\varepsilon(\mathbf{O};\mathbf{P})|\mathbf{P}[1,1] = W] \\
&= \mathbb{E}_{\mathbf{P}}[\mathrm{Conf}_\varepsilon(\mathbf{O};\mathbf{P})].
\end{aligned} \tag{7}$$

Crucially, Eq. (6) allows us to efficiently identify a minimizing assignment

$$\mathbf{P}^\sharp[1,1] = \operatorname*{argmin}_{W\in\{X,Y,Z\}} \mathbb{E}_{\mathbf{P}}[\mathrm{Conf}_\varepsilon(\mathbf{O};\mathbf{P})|\mathbf{P}[1,1] = W]. \tag{8}$$

Doing so replaces an initially random single-qubit measurement setting by a concrete Pauli label that minimizes the conditional expectation value over all remaining (random)

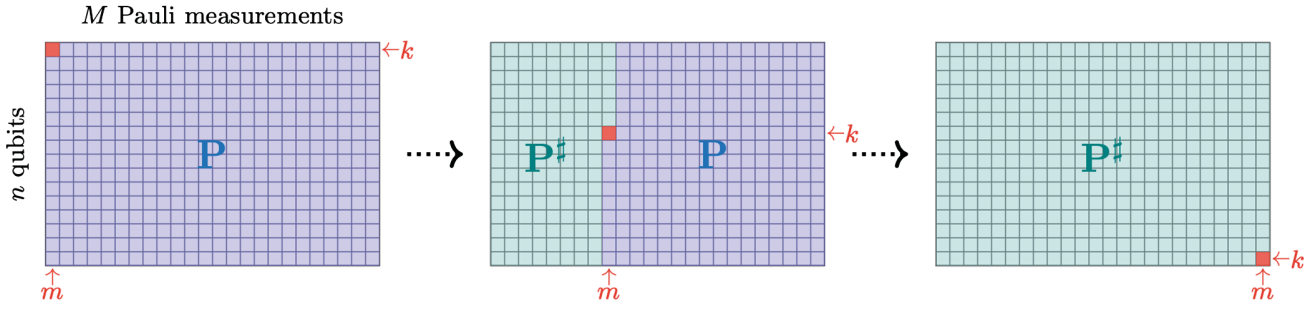


FIG. 1. Illustration of the derandomization algorithm (Algorithm 1): We envision M randomized n-qubit measurements as a two-dimensional array composed of $n \times M$ Pauli labels. Blue squares are place holders for random Pauli labels, while green squares denote deterministic assignments (either X, Y, or Z). Starting with a completely unspecified array (left), the algorithm iteratively checks how a concrete Pauli assignment (red square) affects the confidence bound [Eq. (3)] averaged over all remaining assignments. A simple update rule [Eq. (8)] replaces the initially random label with a deterministic assignment that keeps the remaining confidence bound expectation as small as possible (center). Once the entire grid is traversed, no randomness is left (right) and the algorithm outputs M deterministic n-qubit Pauli measurements.

Algorithm 1. The derandomization algorithm proposed in this work for finding an efficient scheme for measuring a collection of n-qubit Pauli observables.

Derandomization.

Input: measurement budget M, accuracy ε, and L n-qubit Pauli observables $\mathbf{O} = [\mathbf{o}_1, \ldots, \mathbf{o}_L]$.
Output: M Pauli measurements $\mathbf{P}^\sharp \in \{X, Y, Z\}^{n\times M}$.

function DERANDOMIZATION $(\mathbf{O}, M, \varepsilon)$
  initialize $\mathbf{P}^\sharp = [[]]$ (empty $n \times M$ array)
  for $m = 1$ to M **do** ▷ loop over measurements
    for $k = 1$ to n **do** ▷ loop over qubits
      for $W = X, Y, Z$ **do compute**
        $f(W) = \mathbb{E}_{\mathbf{P}}[\mathrm{Conf}_\varepsilon(\mathbf{O};\mathbf{P})|$
          $\mathbf{P}^\sharp, \mathbf{P}[k,m] = W]$
        [see Eq. (6) for a precise formula]
      $\mathbf{P}^\sharp[k,m] \leftarrow \mathrm{argmin}_{W\in\{X,Y,Z\}} f(W)$
  output $\mathbf{P}^\sharp \in \{X, Y, Z\}^{n\times M}$

assignments. This procedure is known as derandomization [24–26] and can be iterated. Figure 1 provides visual guidance, while pseudo-code can be found in Algorithm 1. There are a total of $n \times M$ iterations. Step (k, m) is contingent on comparing three conditional expectation values $\mathbb{E}_{\mathbf{P}}[\mathrm{Conf}_\varepsilon(\mathbf{O};\mathbf{P})|\mathbf{P}^\sharp, \mathbf{P}[k,m] = W]$ and assigning the Pauli label that achieves the smallest score. These update rules are constructed to ensure that (appropriate modifications of) Eq. (7) remain valid throughout the procedure. Combining all of them implies the following rigorous statement about the resulting Pauli measurements $\mathbf{P}^\sharp$.

Theorem 2. (Derandomization promise).—Algorithm 1 is guaranteed to output Pauli measurements $\mathbf{P}^\sharp$ with below average confidence bound: $\mathrm{Conf}_\varepsilon(\mathbf{O};\mathbf{P}^\sharp) \le \mathbb{E}_{\mathbf{P}}[\mathrm{Conf}_\varepsilon(\mathbf{O};\mathbf{P})]$.

We see that derandomization produces deterministic Pauli measurements that perform at least as favorably as (averages of) randomized measurement protocols. But the actual difference between randomized and derandomized Pauli measurements can be much more pronounced. In the examples we considered, derandomization reduces the measurement budget M by at least an order of magnitude, compared to randomized measurements. Furthermore, because Algorithm 1 implements a greedy update procedure, we have no assurance that our derandomized measurement procedure is globally optimal or even close to optimal. Using dynamic programming, the derandomization algorithm runs in time $\mathcal{O}(nML)$; see Supplemental Material Sec. C 3 for a detailed implementation [21].

Numerical experiments.—The ability to accurately estimate many Pauli observables is an essential subroutine for variational quantum eigensolvers (VQEs) [4,8–10,27]. Randomized Pauli measurements [11,12]—also known as classical shadows in this context—offer a conceptually simple solution that is efficient both in terms of quantum hardware and measurement budget.

Derandomization can and should be viewed as a refinement of the original classical shadows idea. Supported by rigorous theory (Theorem 2), this refinement is only contingent on an efficient classical preprocessing step, namely, running Algorithm 1. It does not incur any extra cost in terms of quantum hardware and classical postprocessing, but can lead to substantial performance gains. Numerical experiments visualized in Ref. [12], Fig. 5, have revealed unconditional improvements of about one order of magnitude for a particular VQE experiment [28] (simulating quantum field theories).

In this section, we present additional numerical studies that support this favorable picture. These address a slight variation of Algorithm 1 that does not require fixing the total measurement budget M in advance. We focus on the "electronic structure problem": determine the ground-state energy for molecules with unknown electronic structure. This is one of the most promising VQE applications in quantum chemistry and material science. Different encoding schemes—most notably Jordan-Wigner (JW) [29], Bravyi-Kitaev (BK) [30] and parity (P) [30,31]—allow for mapping molecular Hamiltonians to qubit Hamiltonians that correspond to sums of Pauli observables. Several benchmark molecules have been identified whose encoded Hamiltonians are just simple enough for an explicit classical minimization, so that we can compare Pauli estimation techniques with the exact answer.

Figure 2 illustrates one such comparison. We fix a benchmark molecule BeH_2, a BK encoding and plot the ground-state energy approximation error against the

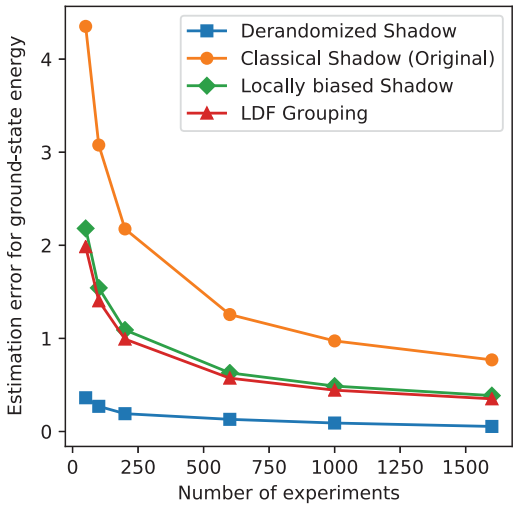


FIG. 2. BeH_2 ground-state energy estimation error (in Hartree) under Bravyi-Kitaev encoding [30] for different measurement schemes: The error for derandomized shadow is the root-mean-squared error (RMSE) over ten independent runs. The error for the other methods shows the RMSE over infinitely many runs and can be evaluated efficiently using the variance of one experiment [14].

TABLE I. Average estimation error using 1000 measurements for different molecules, encodings, and measurement schemes: The first column shows the molecule and the corresponding ground-state electronic (Elec.) energy (in Hartree). We consider the following abbreviations: Derandomized Pauli measurements (Derand.), locally biased classical shadow (Local S.), largest degree first (LDF) heuristic, and original classical shadow (Shadow) [12]

Molecule (E_{GS})	Encodings	Derand.	Local S.	LDF	Shadow
H_2 (−1.86)	*JW*	0.06	0.13	0.15	0.41
	P	0.03	0.14	0.19	0.48
	BK	0.06	0.14	0.19	0.75
Li H (−8.91)	*JW*	0.03	0.12	0.23	0.52
	P	0.03	0.16	0.29	0.87
	BK	0.04	0.26	0.27	0.40
BeH_2 (−19.04)	*JW*	0.06	0.26	0.37	1.29
	P	0.09	0.36	0.49	1.77
	BK	0.06	0.49	0.44	0.97
H_2O (−83.60)	*JW*	0.12	0.51	1.02	1.68
	P	0.22	0.65	1.63	2.52
	BK	0.20	1.17	1.45	3.25
NH_3 (−66.88)	*JW*	0.18	0.59	0.94	3.79
	P	0.21	0.83	1.61	2.13
	BK	0.12	0.73	1.45	1.89

number of Pauli measurements. The plot highlights that derandomization outperforms the original classical shadows procedure (randomized Pauli measurements) [12], locally biased classical shadows [12], and another popular technique known as LDF grouping [14,32]. The discrepancy between randomized and derandomized Pauli measurements is particularly pronounced.

This favorable picture extends to a variety of other benchmark molecules and other encoding schemes, see Table I. For a fixed measurement budget, derandomization consistently leads to a smaller estimation error than other state-of-the-art techniques. One could also repeat the measurement scheme found by the derandomization algorithm multiple times to improve the estimation error; see Supplemental Material Sec. C.4 [21]. Finally, we note that in the presence of measurement noise, the various approaches we have considered are likely to suffer about equally, as they were all based on single-qubit Pauli measurements. One could mitigate such noise by incorporating recently proposed noise inversion techniques [33,34].

Conclusion and outlook.—We consider the problem of predicting many Pauli expectation values from few Pauli measurements. Derandomization [24–26] provides an efficient procedure that replaces originally randomized single-qubit Pauli measurements by specific Pauli assignments. The resulting Pauli measurements are deterministic, but inherit *all* advantages of a fully randomized measurement protocol. Furthermore, the derandomization procedure can accurately capture the fine-grained structure of the observables in question. Predicting molecular ground-state energies based on derandomized Pauli measurements scales favorably and improves upon many existing techniques [11,14,16,32]. Source code for an implementation of the proposed procedure is available at [35].

Randomized measurements have also been used to estimate entanglement entropy [12,36–38], topological invariants [39,40], benchmark physical devices [12,22,41,42], and predict outcomes of physical experiments [43]. Derandomization provides a principled approach for adapting randomized measurement procedures to fine-grained structure and is closely related to an algorithmic technique—multiplicative weight update [44]—commonly used in machine learning and game theory. So far, we have only considered estimations of Pauli observables, but measurement design via derandomization should apply more broadly; we look forward to applying derandomization to other tasks such as estimating non-Pauli observables and entanglement entropies. Additional improvements in performance might be achieved by modifying the cost function $f(W)$ used in Algorithm 1, for example, by greedily assigning more than one single-qubit Pauli measurement in each iteration.

The authors thank Andreas Elben, Stefan Hillmich, Steven T. Flammia, Jarrod McClean, and Lorenzo Pastori for valuable input and inspiring discussions. H. H. is supported by the J. Yang and Family Foundation. J. P. acknowledges funding from the U.S. Department of Energy Office of Science, Office of Advanced Scientific Computing Research, (DE-NA0003525, DE-SC0020290), and the National Science Foundation (PHY-1733907). The Institute for Quantum Information and Matter is a NSF Physics Frontiers Center.

*hsinyuan@caltech.edu

[1] J. Preskill, Quantum computing in the NISQ era and beyond, Quantum **2**, 79 (2018).

[2] K. Bharti, A. Cervera-Lierta, T. H. Kyaw, T. Haug, S. Alperin-Lea, A. Anand, M. Degroote, H. Heimonen, J. S. Kottmann, T. Menke, W.-K. Mok, S. Sim, L.-C. Kwek, and A. Aspuru-Guzik, Noisy intermediate-scale quantum (NISQ) algorithms, arXiv:2101.08448.

[3] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio *et al.*, Variational quantum algorithms, arXiv:2012.09265.

[4] C. Hempel, C. Maier, J. Romero, J. McClean, T. Monz, H. Shen, P. Jurcevic, B. P. Lanyon, P. Love, R. Babbush, A. Aspuru-Guzik, R. Blatt, and C. F. Roos, Quantum Chemistry Calculations on a Trapped-Ion Quantum Simulator, Phys. Rev. X **8**, 031022 (2018).

[5] H.-Y. Huang, K. Bharti, and P. Rebentrost, Near-term quantum algorithms for linear systems of equations, arXiv:1909.07344.

[6] H.-Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, and J. R. McClean, Power of data in quantum machine learning, Nat. Commun. **12**, 2631 (2021).

[7] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets, Nature (London) **549**, 242 (2017).

[8] P. J. J. O'Malley, R. Babbush, I. D. Kivlichan, J. Romero, J. R. McClean, R. Barends, J. Kelly, P. Roushan, A. Tranter, N. Ding *et al.*, Scalable Quantum Simulation of Molecular Energies, Phys. Rev. X **6**, 031007 (2016).

[9] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien, A variational eigenvalue solver on a photonic quantum processor, Nat. Commun. **5**, 4213 (2014).

[10] G. A. Quantum *et al.*, Hartree-Fock on a superconducting qubit quantum computer, Science **369**, 1084 (2020).

[11] T. J. Evans, R. Harper, and S. T. Flammia, Scalable Bayesian Hamiltonian learning, arXiv:1912.07636.

[12] H.-Y. Huang, R. Kueng, and J. Preskill, Predicting many properties of a quantum system from very few measurements, Nat. Phys. **16**, 1050 (2020).

[13] A. Elben, B. Vermersch, C. F. Roos, and P. Zoller, Statistical correlations between locally randomized measurements: A toolbox for probing entanglement in many-body quantum states, Phys. Rev. A **99**, 052323 (2019).

[14] C. Hadfield, S. Bravyi, R. Raymond, and A. Mezzacapo, Measurements of quantum Hamiltonians with locally-biased classical shadows, arXiv:2006.15788.

[15] M. Ohliger, V. Nesme, and J. Eisert, Efficient and feasible state tomography of quantum many-body systems, New J. Phys. **15**, 015024 (2013).

[16] M. Paini and A. Kalev, An approximate description of quantum states, arXiv:1910.10543.

[17] O. Crawford, B. van Straaten, D. Wang, T. Parks, E. Campbell, and S. Brierley, Efficient quantum measurement of Pauli operators in the presence of finite sampling error, Quantum **5**, 385 (2021).

[18] W. J. Huggins, J. R. McClean, N. C. Rubin, Z. Jiang, N. Wiebe, K. B. Whaley, and R. Babbush, Efficient and noise resilient measurements for quantum chemistry on near-term quantum computers, npj Quantum Inf. **7**, 23 (2021).

[19] A. F. Izmaylov, T.-C. Yen, R. A. Lang, and V. Verteletskyi, Unitary partitioning approach to the measurement problem in the variational quantum eigensolver method, J. Chem. Theory Comput. **16**, 190 (2020).

[20] T.-C. Yen and A. F. Izmaylov, Cartan sub-algebra approach to efficient measurements of quantum observables, arXiv:2007.01234.

[21] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.127.030503 for additional proofs of the theorems and details of the numerical experiments.

[22] A. Elben, R. Kueng, H.-Y. R. Huang, R. van Bijnen, C. Kokail, M. Dalmonte, P. Calabrese, B. Kraus, J. Preskill, P. Zoller, and B. Vermersch, Mixed-State Entanglement from Local Randomized Measurements, Phys. Rev. Lett. **125**, 200501 (2020).

[23] M. Ohliger, V. Nesme, and J. Eisert, Efficient and feasible state tomography of quantum many-body systems, New J. Phys. **15**, 015024 (2013).

[24] N. Alon and J. H. Spencer, *The Probabilistic Method*, 3rd ed. Wiley-Interscience Series in Discrete Mathematics and Optimization (Wiley, New York, 2008).

[25] R. Motwani and P. Raghavan, *Randomized Algorithms* (Cambridge University Press, Cambridge, England, 1995).

[26] V. V. Vazirani, *Approximation Algorithms* (Springer, New York, 2001).

[27] A. Kandala, A. Mezzacapo, K. Temme, M. Takita, M. Brink, J. M. Chow, and J. M. Gambetta, Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets, Nature (London) **549**, 242 (2017).

[28] C. Kokail, C. Maier, R. van Bijnen, T. Brydges, M. K. Joshi, P. Jurcevic, C. A. Muschik, P. Silvi, R. Blatt, C. F. Roos *et al.*, Self-verifying variational quantum simulation of lattice models, Nature (London) **569**, 355 (2019).

[29] P. Jordan and E. Wigner, Über das paulische äquivalenzverbot, Z. Phys. **47**, 631 (1928).

[30] S. B. Bravyi and A. Y. Kitaev, Fermionic quantum computation, Ann. Phys. (Amsterdam) **298**, 210 (2002).

[31] J. T. Seeley, M. J. Richard, and P. J. Love, The Bravyi-Kitaev transformation for quantum computation of electronic structure, J. Chem. Phys. **137**, 224109 (2012).

[32] V. Verteletskyi, T.-C. Yen, and A. F. Izmaylov, Measurement optimization in the variational quantum eigensolver using a minimum clique cover, J. Chem. Phys. **152**, 124114 (2020).

[33] S. Chen, W. Yu, P. Zeng, and S. T. Flammia, Robust shadow estimation, arXiv:2011.09636.

[34] D. E. Koh and S. Grewal, Classical shadows with noise, arXiv:2011.11580.

[35] H.-Y. Huang, R. Kueng, and J. Preskll, Source code for the derandomization procedure, https://github.com/momohuang/predicting-quantum-properties (accessed 2021).

[36] T. Brydges, A. Elben, P. Jurcevic, B. Vermersch, C. Maier, B. P. Lanyon, P. Zoller, R. Blatt, and C. F. Roos, Probing rényi entanglement entropy via randomized measurements, Science **364**, 260 (2019).

[37] A. Rath, R. van Bijnen, A. Elben, P. Zoller, and B. Vermersch, Importance sampling of randomized measurements for probing entanglement, arXiv:2102.13524.

[38] V. Vitale, A. Elben, R. Kueng, A. Neven, J. Carrasco, B. Kraus, P. Zoller, P. Calabrese, B. Vermersch, and M. Dalmonte, Symmetry-resolved dynamical purification in synthetic quantum matter, arXiv:2101.07814.

[39] Z.-P. Cian, H. Dehghani, A. Elben, B. Vermersch, G. Zhu, M. Barkeshli, P. Zoller, and M. Hafezi, Many-Body Chern Number from Statistical Correlations of Randomized Measurements, Phys. Rev. Lett. **126**, 050501 (2021).

[40] A. Elben, J. Yu, G. Zhu, M. Hafezi, F. Pollmann, P. Zoller, and B. Vermersch, Many-body topological invariants from randomized measurements in synthetic quantum matter, Sci. Adv. **6**, eaaz3666 (2020).

[41] J. Choi, A. L. Shaw, I. S. Madjarov, X. Xie, J. P. Covey, J. S. Cotler, D. K. Mark, H.-Y. Huang, A. Kale, H. Pichler, F. G. S. L. Brandão, S. Choi, and M. Endres, Emergent

randomness and benchmarking from many-body quantum chaos, arXiv:2103.03535.

[42] 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, Phys. Rev. A **77**, 012307 (2008).

[43] H.-Y. Huang, R. Kueng, and J. Preskill, Information-theoretic bounds on quantum advantage in machine learning, Phys. Rev. Lett. **126**, 190505 (2021).

[44] S. Arora, E. Hazan, and S. Kale, The multiplicative weights update method: A meta-algorithm and applications, Theory Comput. **8**, 121 (2012).

Appendix A: Illustrative derandomization examples

The exact workings of Algorithm 1 depend on the structure of the set of Pauli observables. In this appendix section, we provide several examples to illustrate the mechanism of the derandomization procedure.

1. Many local Pauli observables.

Many near-term applications of quantum devices rely on repeatedly estimating a large number of low-weight Pauli observables. For example, low-energy eigenstates of a many-body Hamiltonian may be prepared and studied using a variational method, in which the Hamiltonian, a sum of local terms, is measured many times. Using randomized measurements, we can predict many low-weight observables simultaneously at comparatively little cost. It is known that a logarithmic number of randomized Pauli measurements allows for accurately predicting a polynomial number of low-weight observables [22].

This desirable feature provably extends to derandomized measurements. From Theorem 2 and Eq. (5), we infer that the measurement budget $M = 4\log(2L/\delta)\max_\ell 3^{\mathrm{w}(\mathbf{o}_\ell)}/\varepsilon^2$ suffices to ensure that Algorithm 1 outputs Pauli measurements $\mathbf{P}^\sharp$ that obey $\textsc{Conf}_\varepsilon(\mathbf{O};\mathbf{P}) \leq \delta/2$. With Lemma 1, we may convert this into an error bound: empirical averages (2) formed from appropriate measurement outcomes are guaranteed to obey $|\hat{\omega}_\ell - \mathrm{tr}(O_{\mathbf{o}_\ell}\rho)| \leq \varepsilon$ for all $1 \leq \ell \leq L$ with high probability (at least $1-\delta$). This error bound is roughly on par with the best rigorous result about predicting local Pauli observables from randomized Pauli measurements [16]. But this argument implicitly assumes that $\textsc{Conf}_\varepsilon(\mathbf{O};\mathbf{P}^\sharp)$ (which we can compute) is comparable to $\mathbb{E}_{\mathbf{P}}\left[\textsc{Conf}_\varepsilon(\mathbf{O};\mathbf{P})\right]$ (which is characterized by Eq. (5)). This assumption is extremely pessimistic, because often $\textsc{Conf}_\varepsilon(\mathbf{O};\mathbf{P}^\sharp) \ll \mathbb{E}_{\mathbf{P}}\left[\textsc{Conf}_\varepsilon(\mathbf{O};\mathbf{P})\right]$. If this is the case, derandomized Pauli measurements perform substantially better.

2. Few global Pauli observables.

We have seen that derandomized measurements never perform worse than randomized measurements. But they can perform much better. This discrepancy is best illustrated with a simple example: design Pauli measurements to predict both a complete Y-string ($\mathbf{o}_1 = [Y,\ldots,Y]$) and a complete Z-string ($\mathbf{o}_2 = [Z,\ldots,Z]$). Here, randomized measurements are a terrible idea, because it is exponentially unlikely to hit either string by chance alone.

Contrast this with derandomization. For the very first assignment ($k=1$,$m=1$), Algorithm 1 starts by computing three conditional expectations. Comparing them reveals $f(Y) = f(Z) < f(X)$ and the algorithm determines that assigning X is likely a bad idea. The two remaining choices should be equivalent and the algorithm assigns, say, $\mathbf{P}^\sharp[1,1] = Y$. This initial choice does affect the expected confidence bound associated with the second Pauli label ($k=2$,$m=1$): $f(Y) < f(X) = f(Z)$. Taking into account the already assigned first Pauli label, both X and Z become equally unfavorable and the algorithm sticks to assigning $\mathbf{P}^\sharp[2,1] = Y$. This situation now repeats itself until the first Pauli measurement is completely assigned: $\mathbf{p}_1^\sharp = [Y,\ldots,Y] = \mathbf{o}_1$. The algorithm has successfully kept track of an entire global Pauli string.

It is now time to assign the first Pauli label of the second Pauli measurement ($k=1$, $m=2$). While X is still a bad idea, taking into account that we have already measured $\mathbf{o}_1$ once also breaks the symmetry between Y and Z assignments: $f(Z) < f(Y) < f(X)$. So the algorithm chooses $\mathbf{P}^\sharp[1,2] = Z$ and subsequently sticks to assigning Z for all qubits: $\mathbf{p}_2^\sharp = [Z,\ldots,Z] = \mathbf{o}_2$. Having measured both $\mathbf{o}_1$ and $\mathbf{o}_2$ an equal number of times restores the initial symmetry and the algorithm basically resets. This process resets until all M Pauli measurements are assigned and Algorithm 1 outputs $\mathbf{P}^\sharp = [\mathbf{o}_1,\mathbf{o}_2,\ldots,\mathbf{o}_1,\mathbf{o}_2]$. In words: measure both global observables equally often. Although statistically optimal, this measurement protocol is neither surprising nor particularly interesting. What is encouraging, though, is that Algorithm 1 has (re-)discovered it all by itself.

3. Very many global Pauli observables (non-example):

The derandomization algorithm is not without flaws. The greedy update rule in line 8 of Algorithm 1 can be misguided to produce non-optimal results. This happens, for instance, for a very large collection of global Pauli observables that appears to have favorable structure but actually doesn't. For instance, set $\mathbf{o}_1 = [X,\ldots,X]$ and $\mathbf{o}_\ell = [Z;\tilde{\mathbf{o}}_\ell]$, where $\tilde{\mathbf{o}}_\ell \in \{X,Y,Z\}^{n-1}$ ranges through all 3^{n-1} possible Pauli strings of size $(n-1)$. There are $L = 3^{n-1}+1$ target observables, all of which are global and therefore incompatible. However, 3^{n-1} of them start with a Pauli-Z label. This imbalance leads the algorithm to believe that assigning $\mathbf{P}^\sharp[1,m] = Z$ for all $1 \leq m \leq M$ is always a good idea (provided that M is not much larger than 3^{n-1}). By doing so, it completely ignores the first target observable which starts with an X-label. But at the same time, it cannot capitalize on this particular decision, because observables $\mathbf{o}_2$ to $\mathbf{o}_L$ are actually incompatible. This results in an imbalanced

output $\mathbf{P}^\sharp$ that treats observables $\mathbf{o}_2$ to $\mathbf{o}_L$ roughly equally, but completely forgets about $\mathbf{o}_1$. Needless to say, the resulting confidence bound will not be minimal either. We emphasize that this highly stylized non-example is not motivated by actual applications. Instead it is intended to illustrate how greedy update procedures can get stuck in local minima.

Appendix B: Additional details and proofs

1. Proof of Lemma 1

Let us briefly recapitulate the general setting. A n-qubit Pauli measurement $\mathbf{p} \in \{X, Y, Z\}^n$ produces a random string of n signs $\hat{\mathbf{q}} \in \{\pm 1\}^n$. Information about the underlying n-qubit state ρ is encoded in the distribution of outcome strings

$$\Pr\left[\hat{\mathbf{q}} = \mathbf{q} | \mathbf{p}, \rho\right] = \mathrm{tr}\left(\bigotimes_{j=1}^{m} \tfrac{1}{2}\left(\sigma_I + \mathbf{q}[j]\sigma_{\mathbf{p}[j]}\right)\rho\right) \quad \text{for all } \mathbf{q} \in \{\pm 1\}^n. \tag{B1}$$

Now, suppose that $\mathbf{o} \in \{I, X, Y, Z\}^n$ is another Pauli string that is hit by $\mathbf{p}$ ($\mathbf{o} \rhd \mathbf{p}$). Then, we can appropriately marginalize n-qubit outcome strings $\mathbf{q} \in \{\pm 1\}^n$ to reproduce $\omega(\rho) = \mathrm{tr}\left(O_{\mathbf{o}}\rho\right)$ in expectation:

$$\begin{aligned}
\mathbb{E} \prod_{j:\mathbf{o}[j]\neq I} \mathbf{q}[j] &= \sum_{\mathbf{q}\in\{\pm 1\}^n} \Pr\left[\mathbf{q}|\mathbf{p},\rho\right] \prod_{j:\,\mathbf{o}_j\neq I} \mathbf{q}[j] \\
&= \sum_{\mathbf{q}\in\{\pm 1\}^n} \mathrm{tr}\left(\bigotimes_{j:\mathbf{o}[j]\neq I} \tfrac{1}{2}\left(\mathbf{q}[j] + \sigma_{\mathbf{p}[j]}\right) \bigotimes_{j:\mathbf{o}[j]=I} \tfrac{1}{2}\left(\sigma_I + \mathbf{q}[j]\sigma_{\mathbf{p}[j]}\right)\rho\right) \\
&= \tfrac{1}{2^n} \sum_{\mathbf{q}\in\{\pm 1\}^n} \mathrm{tr}\left(\bigotimes_{j:\,\mathbf{o}[j]\neq I} \sigma_{\mathbf{o}[j]} \bigotimes_{j:\,\mathbf{o}[j]=I} \sigma_I \rho\right) = \mathrm{tr}\left(\bigotimes_{j=1}^{n} \sigma_{\mathbf{o}[j]}\rho\right) = \mathrm{tr}\left(O_{\mathbf{o}}\rho\right),
\end{aligned} \tag{B2}$$

whenever $\mathbf{o} \rhd \mathbf{p}$ (which ensures $\mathbf{o}[j] = \mathbf{p}[j]$ whenever $\mathbf{o}[j] \neq I$). Now, suppose that we perform a total of M Pauli measurements $\mathbf{p}_1, \ldots, \mathbf{p}_M$. The above relation suggests to approximate Pauli observables $\omega_\ell(\rho) = \mathrm{tr}(O_{\mathbf{o}_\ell}\rho)$ by empirical averages:

$$\hat{\omega}_\ell = \begin{cases} \frac{1}{h(\mathbf{o}_\ell;\mathbf{P})} \sum_{m:\mathbf{o}_\ell \rhd \mathbf{p}_m} \prod_{j:\mathbf{o}_\ell[j]\neq I} \mathbf{q}_m[j] & \text{if } h(\mathbf{o}_\ell;\mathbf{P}) \geq 1 \\ 0 & \text{if } h(\mathbf{o}_\ell;\mathbf{P}) = 0. \end{cases} \tag{B3}$$

Here, $h(\mathbf{o}_\ell;\mathbf{P}) = \sum_{m=1}^{M} \mathbf{1}\left\{\mathbf{o}_\ell \rhd \mathbf{p}_m\right\}$ denotes the *hitting count*, i.e. the number of times a Pauli measurement $\mathbf{p}_m$ provides meaningful information about observable $\mathbf{o}_\ell$. If $h(\mathbf{o}_\ell;\mathbf{P}) = 0$, not a single Pauli measurement is compatible with the target observable in question and we set $\hat{\omega}_\ell = 0$, because we do not have any actionable advice. The above procedure allows us to jointly estimate L Pauli observables based on M Pauli measurement outcomes. The quality of reconstruction is exponentially suppressed in the number of times we hit each target Pauli observable.

Lemma 2. *Fix a collection of M Pauli measurements $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_M]$, a collection of L Pauli observables $\omega_\ell(\rho) = \mathrm{tr}\left(O_{\mathbf{o}_\ell}\rho\right)$. Then, for all $\varepsilon > 0$*

$$\Pr\left[\max_{1\leq\ell\leq L} |\hat{\omega}_\ell - \omega_\ell(\rho)| \geq \varepsilon\right] \leq 2\sum_{\ell=1}^{L} \exp\left(-\tfrac{\varepsilon^2}{2} h(\mathbf{o}_\ell;\mathbf{P})\right). \tag{B4}$$

Lemma 1 in the main text is an immediate consequence of this concentration inequality.

Proof. The union bound – also known as Boole's inequality – states that the probability associated with a union of events is upper bounded by the sum of individual event probabilities. For the task at hand, it implies

$$\Pr\left[\max_{1\leq\ell\leq L} |\hat{\omega}_\ell - \omega_\ell(\rho)| \geq \varepsilon\right] = \Pr\left[\bigcup_{\ell=1}^{L} \left\{|\hat{\omega}_\ell - \omega_\ell| \geq \varepsilon\right\}\right] \leq \sum_{\ell=1}^{L} \Pr\left[|\hat{\omega}_\ell - \omega_\ell(\rho)| \geq \varepsilon\right]. \tag{B5}$$

This allows us to treat individual deviation probabilities separately. Fix $1 \leq \ell \leq L$ and note that $\hat{\omega}_\ell$ is an empirical average of $M_\ell = h(\mathbf{o}_\ell;\mathbf{P})$ random signs $s_i^{(\ell)} = \prod_{j:\,\mathbf{o}_\ell[j]\neq I} \mathbf{q}_i[j] \in \{\pm 1\}$ that are independent each (they arise from different measurement outcomes). Empirical averages of independent signed random variables

tend to concentrate sharply around their true expectation value $\mathbb{E}s_i^{(\ell)} = \mathrm{tr}(O_{\mathbf{o}_\ell}\rho)$. Hoeffding's inequality makes this intuition precise and asserts for any $\varepsilon > 0$

$$\Pr\left[|\hat{\omega}_\ell - \omega_\ell(\rho)| \geq \varepsilon\right] = \Pr\left[\left|\tfrac{1}{M_\ell}\sum_{i=1}^{M_\ell}\left(s_i^{(\ell)} - \mathbb{E}s_i^{(\ell)}\right)\right| \geq \varepsilon\right] \leq 2\exp\left(-\tfrac{\varepsilon^2}{2}M_\ell\right). \tag{B6}$$

The claim follows, because such an exponential bound is valid for each term in Eq. (B5). This also includes terms with zero hits ($M_\ell = 0$), because $\Pr\left[|\hat{\omega}_\ell - \omega_\ell| \geq \varepsilon\right] \leq 1 = \exp\left(-0/2\right)$ – and the claim follows. □

2. Derivation of Eq. (6)

Note that each hitting count $h(\mathbf{o}_\ell; \mathbf{P}) = \sum_{m=1}^{M} \mathbf{1}\left\{\mathbf{o}_\ell \rhd \mathbf{p}_m\right\}$ is a sum of M indicator functions that can take binary values each. This structure allows us to rewrite the confidence bound (3) as

$$\begin{aligned}\textsc{Conf}_\varepsilon(\mathbf{O};\mathbf{P}) &= \sum_{\ell=1}^{L}\exp\left(-\tfrac{\varepsilon^2}{2}h(\mathbf{o}_\ell;\mathbf{P})\right) = \sum_{\ell=1}^{L}\prod_{m'=1}^{M}\exp\left(-\tfrac{\varepsilon^2}{2}\mathbf{1}\left\{\mathbf{o}_\ell \rhd \mathbf{p}_{m'}\right\}\right) \\ &= \sum_{\ell=1}^{L}\prod_{m'=1}^{M}\left(1 - \nu\mathbf{1}\left\{\mathbf{o}_\ell \rhd \mathbf{p}_{m'}\right\}\right),\end{aligned} \tag{B7}$$

where $\nu = 1 - \exp\left(-\varepsilon^2/2\right) \in (0,1)$. Next, note that each remaining indicator function can be further decomposed into a product of more elementary indicator functions:

$$\mathbf{1}\left\{\mathbf{o}_\ell \rhd \mathbf{p}_{m'}\right\} = \prod_{k'=1}^{n}\mathbf{1}\left\{\mathbf{o}_\ell[k'] \rhd \mathbf{p}_{m'}[k']\right\} = \prod_{k'=1}^{n}\left(\mathbf{1}\left\{\mathbf{o}_\ell[k'] = I\right\} + \mathbf{1}\left\{\mathbf{o}_\ell[k'] = \mathbf{p}_{m'}[k']\right\}\right). \tag{B8}$$

Finally, note that a randomly assigned single-qubit label $\mathbf{p}_m[j] \in \{X, Y, Z\}$ hits non-identity Pauli label $\mathbf{o}_\ell[j] \neq I$ with probability $1/3$. More precisely,

$$\mathbb{E}_{\mathbf{p}_m[j]}\left[\mathbf{1}\left\{\mathbf{o}_\ell[j] \rhd \mathbf{p}_m[j]\right\}\right] = \Pr_{\mathbf{p}_m[j]}\left[\mathbf{o}_\ell[j] \rhd \mathbf{p}_m[j]\right] = (1/3)^{\mathbf{1}\{\mathbf{o}_\ell[j]\neq I\}} = \begin{cases}1/3 & \text{if } \mathbf{o}_\ell[j] \neq I, \\ 1 & \text{if } \mathbf{o}_\ell[j] = I.\end{cases} \tag{B9}$$

Together with independence, this observation allows us to compute expectation values of confidence bounds that are partially assigned already. Let $\mathbf{P}^\sharp$ denote the already assigned part that encompasses the first $m-1$ Pauli measurements, as well as the first k single-qubit labels of the m-th Pauli measurement: $\mathbf{P}^\sharp = \left[\mathbf{p}_1^\sharp, \ldots, \mathbf{p}_{m-1}^\sharp\right] \cup \left[\mathbf{p}_m^\sharp[1], \ldots, \mathbf{p}_m[k]^\sharp\right]$. We also assume that all remaining Pauli labels are assigned independently and uniformly at random ($\Pr\left[\mathbf{p}_{m'}[k'] = X\right] = \Pr\left[\mathbf{p}_{m'}[k'] = Y\right] = \Pr\left[\mathbf{p}_{m'}[k'] = Z\right] = 1/3$). Independence ensures that the conditional expectation factorizes nicely into individual components:

$$\begin{aligned}\mathbb{E}_{\mathbf{P}}\left[\textsc{Conf}_\varepsilon(\mathbf{O};\mathbf{P})|\mathbf{P}^\sharp\right] =& \sum_{\ell=1}^{L}\prod_{m'=1}^{m-1}\left(1 - \nu\mathbf{1}\{\mathbf{o}_\ell \rhd \mathbf{p}_{m'}^\sharp\}\right) \\ &\times\left(1 - \nu\prod_{k'=1}^{k}\{\mathbf{o}_\ell[k'] \rhd \mathbf{p}_m[k']\}\prod_{k'=k+1}^{n}\mathbb{E}_{\mathbf{p}_m[k']}\{\mathbf{o}_\ell[k'] \rhd \mathbf{p}_m[k']\}\right) \\ &\times\prod_{m'=m+1}^{M}\left(1 - \nu\prod_{k'=1}^{n}\mathbb{E}_{\mathbf{p}_{m'}[k']}\mathbf{1}\{\mathbf{o}_\ell[k'] \rhd \mathbf{p}_{m'}[k']\}\right) \\ =& \sum_{\ell=1}^{L}\prod_{m'=1}^{m-1}\left(1 - \nu\mathbf{1}\{\mathbf{o}_\ell \rhd \mathbf{p}_{m'}^\sharp\}\right)\left(1 - \nu\prod_{k'=1}^{k}\{\mathbf{o}_\ell[k'] \rhd \mathbf{p}_m[k']\}\prod_{k'=k+1}^{n}(1/3)^{\mathbf{1}\{\mathbf{o}_\ell[k']\neq I\}}\right) \\ &\times\prod_{m'=m+1}^{M}\left(1 - \nu\prod_{k'=1}^{n}(1/3)^{\mathbf{1}\{\mathbf{o}_\ell[k']\neq I\}}\right).\end{aligned} \tag{B10}$$

Now, note that the exponent $\sum_{k'=k+1}^{n}\mathbf{1}\{\mathbf{o}_\ell[k'] \neq I\} = \mathrm{w}_{\neg k}(\mathbf{o}_\ell)$ captures the weight of the reduced Pauli string $[\mathbf{o}_\ell[k+1], \ldots, \mathbf{o}_\ell[n]]$ (in particular, $\mathrm{w}_{\neg 0}(\mathbf{o}_\ell) = \mathrm{w}(\mathbf{o}_\ell)$) Reading Eq. (B7) backwards to recognize $\prod_{m'=1}^{m-1}\left(1 - \nu\mathbf{1}\{\mathbf{o}_\ell \rhd \mathbf{p}_{m'}^\sharp\}\right) = \exp\left(-\tfrac{\varepsilon^2}{2}h(\mathbf{o}_\ell; [\mathbf{p}_1^\sharp, \ldots, \mathbf{p}_{m-1}^\sharp])\right)$ further simplifies the expression:

$$\mathbb{E}_{\mathbf{P}}\left[\textsc{Conf}_\varepsilon(\mathbf{O};\mathbf{P}^\sharp)|\mathbf{P}^\sharp\right] = \sum_{\ell=1}^{L}\exp\left(-\tfrac{\varepsilon^2}{2}h(\mathbf{o}_\ell; [\mathbf{p}_1^\sharp, \ldots, \mathbf{p}_{m-1}^\sharp])\right)\left(1 - \nu\prod_{k'=1}^{k}\{\mathbf{o}_\ell[k'] \rhd \mathbf{p}_m[k']\}3^{-\mathrm{w}_{\neg k}(\mathbf{o}_\ell)}\right) \tag{B11}$$

$$\times \left(1 - \nu 3^{-\mathrm{w}(\mathbf{o}_\ell)}\right)^{M-m}.$$

Appendix C: Further details regarding numerical experiments

1. Quantum chemistry applications

We consider a molecular electronic Hamiltonian that has been encoded into an n-qubit system. The Hamiltonian can be written as a sum of Pauli observables.

$$H = \sum_{P \in \{I,X,Y,Z\}^n} \alpha_P P. \tag{C1}$$

The number of qubits for different molecules is given by

$$\mathrm{H}_2 : n = 8,\ \mathrm{LiH} : n = 12,\ \mathrm{BeH}_2 : n = 14,\ \mathrm{H}_2\mathrm{O} : n = 14,\ \mathrm{NH}_3 : n = 16. \tag{C2}$$

Each molecule is represented by a fermionic Hamiltonian in a minimal STO-3G basis, ranging from 4 to 16 spin orbitals. The 8-qubit H_2 example is represented using a 6-31G basis. The fermionic Hamiltonian is mapped to a qubit Hamiltonian using three different common encodings: *Jordan-Wigner* (JW) [27], *Bravyi-Kitaev* (BK) [5] and *Parity* (P) [5, 44]. The Pauli decomposition considered here has already been featured in many existing works; see [4, 18, 28] for more details.

In our numerical experiments, the measurement procedure is applied to the exact ground state of the encoded n-qubit Hamiltonian H:

$$\rho = |g\rangle\langle g|, \quad \text{where} \quad |g\rangle = \arg\min_{|\psi\rangle} \langle\psi| H |\psi\rangle . \tag{C3}$$

The ground state $|g\rangle$ is obtained by exact diagonalization using the Lanczos method, see e.g. [33] for a recent survey. We focus on root-mean squared error (RMSE) to quantify the measurement error. For M independent repetitions of the measurement procedure giving rise to M estimates $\hat{E}_1, \ldots, \hat{E}_M$, the RMSE is given by:

$$\mathrm{RMSE} = \sqrt{\frac{1}{M}\sum_{i=1}^{M} (\hat{E}_i - E_{\mathrm{GS}})^2}, \tag{C4}$$

where E_{GS} is the exact ground state electronic energy $\mathrm{tr}(H\rho) = \langle\psi| H |\psi\rangle$. We consider the ground state electronic energy of the molecule without the static Coulomb repulsion energy between the nuclei. Hence the total ground state energy of the molecule is the sum of the ground state electronic energy and the static Coulomb repulsion energy (Born-Oppenheimer approximation). We do not focus on the static Coulomb repulsion energy because it is not encoded in the molecular electronic Hamiltonian H and is considered to be a fixed value.

2. Methods we compare to

We elaborate the alternative measurement procedures with which we compared our derandomized procedure.

1. *LDF grouping*: The largest-degree-first (LDF) grouping strategy and other heuristics have been considered and investigated in [47]. The conclusion is that the LDF grouping strategy results in good performance (differing from the best heuristics by at most 10%) and is generally recommended. The measurement error (RMSE) of LDF grouping strategy can be computed exactly given an exact representation of the ground state $|g\rangle$; see [18] for details.

2. *Classical shadow*: The measurement procedure measures each qubit in a random X, Y, Z Pauli basis. This procedure is known to allow estimation of any L few-body observables from only order $\log(L)$ measurements [11, 16, 22]. However, the performance would degrade significantly when we consider many-body observables. Hence, this approach will likely perform less well for molecular Hamiltonians due to the presence of many high-weight Pauli observables.

3. *Locally-biased classical shadow*: This is an improvement over classical shadows, proposed by [18], designed to overcome disadvantages in estimating the expectation of many-body observables. The idea is to bias the distribution over different Pauli bases (X, Y or Z) for each qubit to minimize the variance when we measure the quantum Hamiltonian given in Equation (C1). Ref. [18] demonstrated that this approach would yield similar or better performance compared to LDF grouping and outperforms classical shadows.

3. Details of the derandomization algorithm

In what follows, we provide a detailed description of the cost function used to derandomize the single-qubit Pauli observables for our numerical experiments. In Algorithm 1, we used the cost function

$$f(W) = \mathbb{E}_{\mathbf{P}}\big[\mathrm{Conf}_{\varepsilon}(\mathbf{O};\mathbf{P})|\mathbf{P}^{\sharp}, \mathbf{P}[k,m] = W\big]. \tag{C5}$$

The conditional expectation is given by Eq. (6) and is restated here for convenience

$$\begin{aligned}\mathbb{E}_{\mathbf{P}}\left[\textsc{Conf}_{\varepsilon}(\mathbf{O};\mathbf{P})|\mathbf{P}^{\sharp}\right] = &\sum_{\ell=1}^{L}\exp\left(-\frac{\varepsilon^2}{2}\sum_{m'=1}^{m-1}\prod_{k'=1}^{n}\mathbf{1}\left\{\mathbf{o}_{\ell}[k'] \rhd \mathbf{P}^{\sharp}[k',m']\right\}\right)\\ &\times\left(1-\nu\prod_{k'=1}^{k}\mathbf{1}\left\{\mathbf{o}_{\ell}[k'] \rhd \mathbf{P}^{\sharp}[k',m]\right\}3^{-\mathrm{w}_{\neg k}(\mathbf{o}_{\ell})}\right)\\ &\times\left(1-\nu 3^{-\mathrm{w}(\mathbf{o}_{\ell})}\right)^{M-m},\end{aligned}$$

where $\nu = 1-\exp(-\varepsilon^2/2)$ and $\mathrm{w}_{\neg k}(\mathbf{o}_{\ell}) = \mathrm{w}([\mathbf{o}_{\ell}[k+1],\ldots,\mathbf{o}_{\ell}[n]])$. This formula requires us to fix the total number of measurements M beforehand. However, one may want to keep measuring until certain criteria are satisfied, e.g., that all of the L Pauli observables has been measured sufficiently many times. In such a scenario, it is unclear what M should be. One approach is to try out various different values of M and choose the one that works best. In the numerical experiments, we consider the following alternative strategy, where we simply remove $\left(1-\nu 3^{-\mathrm{w}(\mathbf{o}_{\ell})}\right)^{M-m}$ since it only depends on the weight of the Pauli observable $\mathbf{o}_{\ell}$. The results are similar and one does not have to choose M beforehand. The precise formula we used in Algorithm 1 is now given by a modified cost function instead of the conditional expectation value,

$$f(W) = C(\mathbf{P}^{\sharp}, \mathbf{P}[k,m] = W). \tag{C6}$$

The modified cost function is a sum of single-observable cost functions $\exp\left(-V(\mathbf{o}_{\ell},\mathbf{P}^{\sharp})\right)$,

$$C(\mathbf{P}^{\sharp}) = \sum_{\ell=1}^{L}\exp\left(-V(\mathbf{o}_{\ell},\mathbf{P}^{\sharp})\right), \tag{C7}$$

$$\begin{aligned}V(\mathbf{o}_{\ell},\mathbf{P}^{\sharp}) = &\frac{\eta}{2}\sum_{m'=1}^{m-1}\prod_{k'=1}^{n}\mathbf{1}\left\{\mathbf{o}_{\ell}[k'] \rhd \mathbf{P}^{\sharp}[k',m']\right\}\\ &-\log\left(1-\frac{\nu}{3^{\mathrm{w}([\mathbf{o}_{\ell}[k+1],\ldots,\mathbf{o}_{\ell}[n]])}}\prod_{k'=1}^{k}\mathbf{1}\left\{\mathbf{o}_{\ell}[k'] \rhd \mathbf{P}^{\sharp}[k',m]\right\}\right),\end{aligned} \tag{C8}$$

where $\eta, \nu > 0$ are hyperparameters that need to be chosen properly. In the numerical experiments, we consider $\eta = 0.9$ and $\nu = 1-\exp(-\eta/2)$. The larger $V(\mathbf{o}_{\ell},\mathbf{P}^{\sharp})$ is, the lower the single-observable cost function $\exp\left(-V(\mathbf{o}_{\ell},\mathbf{P}^{\sharp})\right)$ will be. The following discussion provides an intuitive understanding for the role of the two terms in $V(\mathbf{o}_{\ell},\mathbf{P}^{\sharp})$.

1. The first term in $V(\mathbf{o}_{\ell},\mathbf{P}^{\sharp})$ is proportional to

$$\sum_{m'=1}^{m-1}\prod_{k'=1}^{n}\mathbf{1}\left\{\mathbf{o}_{\ell}[k'] \rhd \mathbf{P}^{\sharp}[k',m']\right\}, \tag{C9}$$

   which determines how many times the Pauli observable $\mathbf{o}_{\ell}$ has been measured in the first $m-1$ Pauli measurements. If the Pauli observable $\mathbf{o}_{\ell}$ has been measured many times, then $V(\mathbf{o}_{\ell},\mathbf{P}^{\sharp})$ is large, and therefore $\exp\left(-V(\mathbf{o}_{\ell},\mathbf{P}^{\sharp})\right)$ is close to zero.

2. The second term in $V(\mathbf{o}_{\ell},\mathbf{P}^{\sharp})$ is approximately equal to the following by Taylor expansion,

$$\frac{\nu}{3^{\mathrm{w}([\mathbf{o}_{\ell}[k+1],\ldots,\mathbf{o}_{\ell}[n]])}}\prod_{k'=1}^{k}\mathbf{1}\left\{\mathbf{o}_{\ell}[k'] \rhd \mathbf{P}^{\sharp}[k',m]\right\}. \tag{C10}$$

   It would be nonzero only when $\mathbf{o}_{\ell}[k'] \rhd \mathbf{P}^{\sharp}[k',m]$ for all $k' = 1,\ldots,k$. Furthermore if the weight of $[\mathbf{o}_{\ell}[k+1],\ldots,\mathbf{o}_{\ell}[n]]$ is smaller, then the single-observable cost function $\exp\left(-V(\mathbf{o}_{\ell},\mathbf{P}^{\sharp})\right)$ incurred by $\mathbf{o}_{\ell}$ would be smaller.

When the entire set of M measurements has been decided, $V(\mathbf{o}_\ell, \mathbf{P}^\sharp)$ will consist only of the first term and is proportional to the number of times the observable $\mathbf{o}_\ell$ has been measured.

For quantum chemistry applications, the coefficients of different Pauli observable are different, e.g., in Eq. (C1), the Hamiltonian H consists of Pauli observable P with varying coefficients α_P. In such a case, one would want to measure each Pauli observable $\mathbf{o}_\ell$ with a number of times proportional to $|\alpha_{\mathbf{o}_\ell}|$ [34]. In order to include the proportionality to $|\alpha_{\mathbf{o}_\ell}|$, we consider the following modified cost function that depends on the coefficients α,

$$C_\alpha(\mathbf{P}^\sharp) = \sum_{l=1}^{L} \exp\left(-V(\mathbf{o}_\ell, \mathbf{P}^\sharp)/w_{\mathbf{o}_\ell}\right), \text{ where } w_{\mathbf{o}_\ell} = \frac{|\alpha_{\mathbf{o}_\ell}|}{\max_p |\alpha_{\mathbf{o}_p}|}. \tag{C11}$$

The definition of $V(\mathbf{o}_\ell, \mathbf{P}^\sharp)$ is given in Eq. (C8). Recall that $V(\mathbf{o}_\ell, \mathbf{P}^\sharp)$ will be proportional to the number of times the observable $\mathbf{o}_\ell$ has been measured, hence the weight factor $w_{\mathbf{o}_\ell}$ will promote the proportionality of $V(\mathbf{o}_\ell, \mathbf{P}^\sharp)$ to $w_{\mathbf{o}_\ell} \propto |\alpha_{\mathbf{o}_\ell}|$. While the cost function is derived from derandomizing the powerful randomized procedure [22], it is not clear if this is the optimal cost function. We believe other cost functions that are tailored to the particular application could yield even better performance; we leave such an exploration as goal for future work.

For illustration purposes, we present a Python implementation for the derandomization algorithm. We maintain two arrays (lists in Python language): `num_of_measurements_so_far` and `num_of_matches_needed_in_this_round`. The former array stores the number of successful measurements for each Pauli observable o_ℓ, while the latter array stores the required number of matches needed to measure each pauli observable o_ℓ. Keeping these two arrays enables us to compute the cost function $f(W)$ in time $\mathcal{O}(L)$. Because the cost function is computed a total of $\mathcal{O}(M \times n)$ times, the time complexity of the derandomization algorithm amounts to $\mathcal{O}(nML)$. Furthermore, in the C++ implementation of the derandomization algorithm – available at `https://github.com/momohuang/predicting-quantum-properties/blob/master/data_acquisition_shadow.cpp` – we implement an improved algorithm using a basic data structure for updating the two arrays, which results in a time complexity of $\mathcal{O}(L \times \sum_{\ell=1}^{L} \mathrm{w}(o_\ell))$, where $\mathrm{w}(o_\ell)$ is the number of non-identity component in the ℓth Pauli observable o_ℓ. We note that the time complexity for performing the derandomization algorithm is equal to the time complexity for estimating the expectation value of the Pauli observables after measurements (up to a multiplicative constant).

```
def derandomized_classical_shadow(all_observables,\
        num_of_measurements_per_observable, system_size, weight=None):
    #
    # Implementation of the derandomized classical shadow
    #
    #     all_observables: a list of Pauli observables, each Pauli observable is a list of tuple
    #                      of the form ("X", position) or ("Y", position) or ("Z", position)
    #     num_of_measurements_per_observable: int for the number of measurement for each observable
    #     system_size: int for how many qubits in the quantum system
    #     weight: None or a list of coefficients for each observable
    #             None — neglect this parameter
    #             a list — modify the number of measurements for each observable
    #.                    by the corresponding weight
    #
    if weight is None:
        weight = [1.0] * len(all_observables)
    assert(len(weight) == len(all_observables))

    def cost_function(num_of_measurements_so_far, num_of_matches_needed_in_this_round):
        eta = 0.9 # a hyperparameter that can be tuned
        nu = 1 - math.exp(-eta / 2)

        cost = 0
        for i, zipitem in enumerate(zip(num_of_measurements_so_far,\
                                        num_of_matches_needed_in_this_round)):
            measurement_so_far, matches_needed = zipitem
            if num_of_measurements_so_far[i] >= math.floor(weight[i] *\
                                                num_of_measurements_per_observable):
                continue

            if system_size < matches_needed:
                V = eta / 2 * measurement_so_far
```

```
        else:
            V = eta / 2 * measurement_so_far - math.log(1 - nu / (3 ** matches_needed))
        cost += math.exp(-V / weight[i])
    return cost

def match_up(qubit_i, dice_roll_pauli, single_observable):
    for pauli, pos in single_observable:
        if pos != qubit_i:
            continue
        else:
            if pauli != dice_roll_pauli:
                return -1
            else:
                return 1
    return 0

num_of_measurements_so_far = [0] * len(all_observables)
measurement_procedure = []

for repetition in range(num_of_measurements_per_observable * len(all_observables)):
    # A single round of parallel measurement over "system_size" number of qubits
    num_of_matches_needed_in_this_round = [len(P) for P in all_observables]
    single_round_measurement = []

    for qubit_i in range(system_size):
        cost_of_outcomes = dict([("X", 0), ("Y", 0), ("Z", 0)])

        for dice_roll_pauli in ["X", "Y", "Z"]:
            # Assume the dice rollout to be "dice_roll_pauli"
            for i, single_observable in enumerate(all_observables):
                result = match_up(qubit_i, dice_roll_pauli, single_observable)
                if result == -1:# impossible to measure
                    num_of_matches_needed_in_this_round[i] += 100 * (system_size+10)
                if result == 1: # match up one Pauli X/Y/Z
                    num_of_matches_needed_in_this_round[i] -= 1

            cost_of_outcomes[dice_roll_pauli] = cost_function(num_of_measurements_so_far,\
                        num_of_matches_needed_in_this_round)

            # Revert the dice roll
            for i, single_observable in enumerate(all_observables):
                result = match_up(qubit_i, dice_roll_pauli, single_observable)
                if result == -1:# impossible to measure
                    num_of_matches_needed_in_this_round[i] -= 100 * (system_size+10)
                if result == 1: # match up one Pauli X/Y/Z
                    num_of_matches_needed_in_this_round[i] += 1

        for dice_roll_pauli in ["X", "Y", "Z"]:
            if min(cost_of_outcomes.values()) < cost_of_outcomes[dice_roll_pauli]:
                continue
            # The best dice roll outcome will come to this line
            single_round_measurement.append(dice_roll_pauli)
            for i, single_observable in enumerate(all_observables):
                result = match_up(qubit_i, dice_roll_pauli, single_observable)
                if result == -1:# impossible to measure
                    num_of_matches_needed_in_this_round[i] += 100 * (system_size+10)
                if result == 1: # match up one Pauli X/Y/Z
                    num_of_matches_needed_in_this_round[i] -= 1
            break

    measurement_procedure.append(single_round_measurement)

    for i, single_observable in enumerate(all_observables):
```

```
            if num_of_matches_needed_in_this_round[i] == 0: # finished measuring all qubits
                num_of_measurements_so_far[i] += 1

        success = 0
        for i, single_observable in enumerate(all_observables):
            if num_of_measurements_so_far[i] >= math.floor(weight[i] *\
                                                    num_of_measurements_per_observable):
                success += 1

        if success == len(all_observables):
            break

    return measurement_procedure
```

4. Error scaling by repeating the deterministic measurement procedure

Since it may be relatively inconvenient to change the measurement setting, an experimentalist may prefer to measure the same Pauli observable multiple times, each time on an independently prepared copy of the quantum state. Naturally, repeating each Pauli observable measurement multiple times, while keeping the number of distinct Pauli operators measured fixed, improves the prediction performance.

To illistrate this improvement, we consider the estimation error (in Hartree) for the BeH_2 ground state energy under Bravyi-Kitaev encoding [5]. The numerical results are shown in Figure 4, indicating how the error decreases as we increase the number of measurement repetitions for a particular deterministic measurement procedure designed by the derandomization procedure. The variance in the estimation error also decreases substantially as the number of repetitions increases.

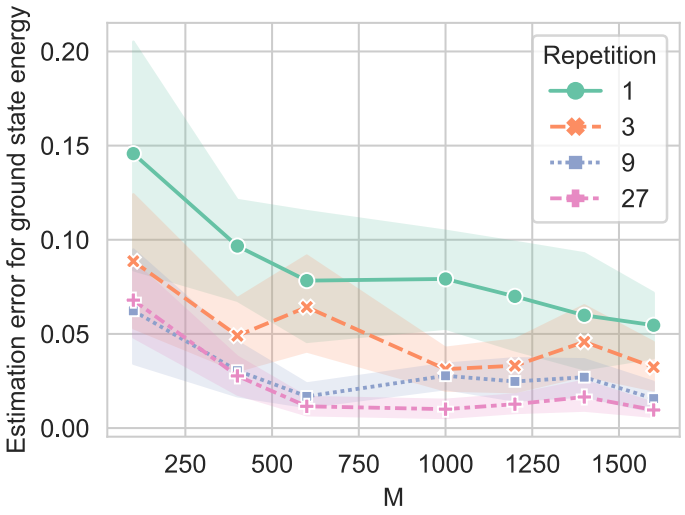


Figure 4: BeH_2 *ground state energy estimation error (in Hartree) under Bravyi-Kitaev encoding [5] for various numbers of repetitions:* The horizontal axis is the number of measured Pauli observables in the measurement procedure found by the derandomization algorithm. The vertical axis is the estimation error, the absolute value of the difference between the estimated ground state energy and the true energy. We consider repeating the deterministic measurement procedure 1, 3, 9, and 27 times. The shaded region surrounding each curve indicates the standard deviation due to quantum measurement fluctuations for the fixed deterministic Pauli measurement scheme.

Chapter 6

Quantum algorithms for convex optimization

or: Faster quantum and classical SDP approximations for quadratic binary optimization

Abstract

We give a quantum speedup for solving the canonical semidefinite programming relaxation for binary quadratic optimization. This class of relaxations for combinatorial optimization has so far eluded quantum speedups. Our methods combine ideas from quantum Gibbs sampling and matrix exponent updates. A de-quantization of the algorithm also leads to a faster classical solver. For generic instances, our quantum solver gives a nearly quadratic speedup over state-of-the-art algorithms. We also provide an efficient randomized rounding procedure that converts approximately optimal SDP solutions into constant factor approximations of the original quadratic optimization problem.

Authors

Fernando G.S.L. Brandão, Richard Kueng, Daniel Stilck França.

Journal

to appear in *Quantum* (2021)

Confirmation of declaration of author contributions (Fernando G.S.L. Brandão)

Publication:

F.G.S.L. Brandão, R. Kueng, D. Stilck França, Faster quantum and classical SDP approximations for quadratic binary optimization, under review at *Quantum* (2021)

Declaration of author contributions:

Daniel Stilck França and Richard Kueng developed the theoretical aspects of this work. Fernando G.S.L. Brandão conceived the original project and provided guidance. All authors wrote the manuscript.

Confirmation by co-author:

I confirm this declaration of author contributions, as well as my co-authorship.

Fernando G.S.L. Brandão

Confirmation of declaration of author contributions (Daniel Stilck França)

Publication:

F.G.S.L. Brandão, R. Kueng, D. Stilck França, Faster quantum and classical SDP approximations for quadratic binary optimization, under review at *Quantum* (2021)

Declaration of author contributions:

Daniel Stilck França and Richard Kueng developed the theoretical aspects of this work. Fernando G.S.L. Brandão conceived the original project and provided guidance. All authors wrote the manuscript.

Confirmation by co-author:

I confirm this declaration of author contributions, as well as my co-authorship.

Daniel Stilck França

Faster quantum and classical SDP approximations for quadratic binary optimization

Fernando G.S L. Brandão,[1,2,3], Richard Kueng[1,2,4], and and Daniel Stilck França[5,6]

[1]Institute for Quantum Information and Matter, California Institute of Technology, Pasadena, CA, USA
[2]Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, USA
[3]AWS Center for Quantum Computing, Pasadena, CA, USA
[4]Institute for Integrated Circuits, Johannes Kepler University Linz, Austria
[5]QMATH, Department of Mathematical Sciences, University of Copenhagen, Denmark
[6]Department of Mathematics, Technische Universität München, Germany

We give a quantum speedup for solving the canonical semidefinite programming relaxation for binary quadratic optimization. This class of relaxations for combinatorial optimization has so far eluded quantum speedups. Our methods combine ideas from quantum Gibbs sampling and matrix exponent updates. A de-quantization of the algorithm also leads to a faster classical solver. For generic instances, our quantum solver gives a nearly quadratic speedup over state-of-the-art algorithms. Such instances include approximating the ground state of spin glasses and MaxCut on Erdös-Rényi graphs. We also provide an efficient randomized rounding procedure that converts approximately optimal SDP solutions into approximations of the original quadratic optimization problem.

1 Introduction

Quadratic optimization problems with binary constraints are an important class of optimization problems. Given a (real-valued) symmetric $n \times n$ matrix A the task is to compute

$$\text{maximize} \quad \langle x|A|x\rangle \quad \text{subject to} \quad x \in \{\pm 1\}^n \qquad \text{(MaxQP)}. \tag{1}$$

This problem arises naturally in many applications across various scientific disciplines, e.g. image compression [OP83], latent semantic indexing [Kol98], community detection [MS16], correlation clustering [CW04, MMMO17] and structured principal component analysis, see e.g. [KT19a, KT19b] and references therein. Mathematically, MaxQPs (1) are closely related to computing the $\infty \to 1$ norm of matrices. This norm, in turn, closely relates to the *cut norm* (replace $x \in \{\pm 1\}^n$ by $x \in \{0,1\}^n$), as both norms can only differ by a constant factor. These norms are an important concept in theoretical computer science [FK99, AFdlVKK03, AN06], since problems such as identifying the largest cut in a graph (MaxCut) can be naturally formulated as instances of these norms. This connection highlights that optimal solutions of (1) are NP-hard to compute in the worst case. Despite their intrinsic hardness, quadratic

optimization problems do admit a canonical *semidefinite programming* (SDP) relaxation[1] [GW95]:

$$\text{maximize} \quad \operatorname{tr}(AX) \quad \text{subject to} \quad \operatorname{diag}(X) = \mathbf{1},\ X \geq 0 \qquad \text{(MaxQP SDP)} \tag{2}$$

Here, $X \geq 0$ indicates that the $n \times n$ matrix X is positive semidefinite (psd), i.e. $\langle y|X|y\rangle \geq 0$ for all $y \in \mathbb{R}^n$. SDPs comprise a rich class of convex optimization problems that can be solved efficiently under mild assumptions, e.g. by using interior point methods [BV04].

Perhaps surprisingly, the optimal value of the MaxQP relaxation often provides a constant factor approximation to the optimal value of the original quadratic problem. However, the associated optimal matrix $X^\sharp$ is typically *not* in one-to-one correspondence with an optimal feasible point $x^\sharp \in \{\pm 1\}^n$ of the original problem (1). Several randomized rounding procedures have been devised to overcome this drawback since the pioneering work of [GW95]. These transform $X^\sharp$ into a random binary vector $\tilde{x} \in \{\pm 1\}^n$ that achieves $\langle \tilde{x}|A|\tilde{x}\rangle \geq \gamma \max_{x \in \{\pm 1\}^n} \langle x|A|x\rangle$ in expectation for some constant γ. Explicit values of γ are known for instance for the case of A being the adjacency matrix of a graph [GW95] or positive semidefinite [AN06].

Although tractable in a theoretical sense, the runtime associated with general-purpose SDP solvers quickly becomes prohibitively expensive in both memory and time. This practical bottleneck has spurred considerable attention in the theoretical computer science community over the past decades [AHK05, BM05, BVB16, TYUC17]. (Meta) algorithms, like *matrix multiplicative weights* (MMW) [AHK05] solve the MaxQP SDP (2) up to multiplicative error $\epsilon \|A\|_{\ell_1}$ in runtime $\mathcal{O}((n/\epsilon)^{2.5} s)$, where s denotes the column sparsity of A. Further improvements are possible if the problem description A has additional structure, such as A being the adjacency matrix of a graph [AK16].

Very recently, a line of works pointed out that quantum computers can solve certain SDPs even faster [BS17, vAGGdW17, vAG19, BKL+17, KP20]. However, the current runtime guarantees depend on problem-specific parameters. These parameters scale particularly poorly for most combinatorial optimization problems, including the MaxQP SDP, and negate any potential advantage.

In this work, we tackle this challenge and overcome the shortcomings of existing quantum SDP solvers by considering the following further relaxation of problem (2):

$$\begin{aligned} \text{find} \quad & X \quad \text{(renormalized, relaxed, feasibility MaxQP SDP)} \\ \text{subject to} \quad & \operatorname{tr}\left(\tfrac{1}{\|A\|} AX\right) \geq \lambda - \epsilon \\ & \sum_i \left| \langle i|X|i\rangle - \frac{1}{n} \right| \leq \epsilon \\ & \operatorname{tr}(X) = 1,\ X \geq 0. \end{aligned} \tag{3}$$

Here we introduced two additional parameters λ and ϵ. The λ parameter comes from a standard trick to reduce the problem in (2) to a sequence of feasibility problems, as we will

[1] Rewrite the objective function in (1) as $\operatorname{tr}(A|x\rangle\langle x|)$ and note that every matrix $X = |x\rangle\langle x|$ with $x \in \{\pm 1\}^n$ has diagonal entries equal to one and is psd with unit rank. Dropping the (non-convex) rank constraint produces a convex relaxation.

explain later. The ϵ parameter encodes a further relaxation of the constraints of Eq. (2). Let us first discuss the case where $\epsilon = 0$, that is, we have the normalized diagonal constraints $\langle i|X|i\rangle = \frac{1}{n}$.

This renormalization of the original problem pinpoints connections to quantum mechanics: Every feasible point X obeys $\mathrm{tr}\,(X) = 1$ and $X \geq 0$, implying that it describes the state ρ of a n-dimensional quantum system. In turn, such quantum states can be represented approximately by a renormalized matrix exponential $\rho = \exp(-H)/\mathrm{tr}(\exp(-H))$, the *Gibbs state* associated with *Hamiltonian* H. We capitalize on this correspondence by devising a meta algorithm – *Hamiltonian Updates* (HU) – that is inspired by matrix exponentiated gradient updates [TRW05], see also [LRS15, BKL+17, Haz16] for similar approaches. Another key insight is that the diagonal constraints also have a clear quantum mechanical interpretation: the feasible states are those that are indistinguishable from the uniform or maximally mixed state when measured in the computational basis.

This interpretation points the way to another key component to obtaining speedups for MaxQP SDP: by setting $\epsilon > 0$ we further relax the problem and optimize over all states that are *approximately* indistinguishable from the maximally mixed state when measured in the computational basis. This further relaxation will allow us to overcome shortcomings of previous solvers when dealing with SDPs of this form, as it bundles up the n linear constraints in Eq. (3) into one. As we ultimately want to solve Eq. (2) and not (3), a significant part of the technical contribution of this work is to show that this further relaxation is mild. Indeed, we will be able to convert a solution to (3) into one to (2) by only slightly changing the objective value.

This will allow us to solve the relaxed problem up to an ϵ additive error by only imposing a relaxation parameter ϵ^4. As it is the case with this and many other related algorithms to solve SDPs, ensuring that we only require a dimension-independent ϵ^4 precision for the constraints is essential to guarantee speedups. Note that to obtain the same level of precision in the formulation given in (3) would require enforcing that each constraint is satisfied up to an error of order $\mathcal{O}(n^{-1})$.

Although originally designed to exploit the fact that quantum architectures can sometimes create Gibbs states efficiently and inspired by interpreting the problem from the point of view of quantum mechanics, it turns out that this approach also produces faster classical algorithms.

To state our results, we instantiate standard computer science notation. The symbol $\mathcal{O}(\cdot)$ describes limiting function behavior, while $\tilde{\mathcal{O}}(\cdot)$ hides poly-logarithmic factors in the problem dimension and polynomial dependencies on the inverse accuracy $1/\epsilon$. We are working with the adjacency list oracle model, where individual entries and location of nonzero entries of the problem description A can be queried at unit cost. We refer to Section 3.4 for a more detailed discussion.

Theorem I (Hamiltonian Updates: runtime)**.** *Let A be a (real-valued), symmetric $n \times n$ matrix with column sparsity s. Then, the associated MaxQP SDP (2) can be solved up to additive accuracy $n\|A\|\epsilon$ in runtime $\tilde{\mathcal{O}}\left(n^{1.5}\left(\sqrt{s}\right)^{1+o(1)}\epsilon^{-28+o(1)}\exp(1.6\sqrt{12\log(\epsilon^{-1})})\right)$ on a quantum computer and $\tilde{\mathcal{O}}\left(\min\{n^2 s, n^\omega\}\epsilon^{-12}\right)$ on a classical computer.*

Here ω is the matrix multiplication exponent. With some abuse of terminology, the word

"solves" is used with slightly different meanings for the classical and quantum algorithms in the statement above. For the classical algorithm we can indeed output a feasible solution of MaxQP SDP that is $n\epsilon$ close to the optimal target value. In the quantum case, the output is in the form of a quantum state ρ such that $n\rho$ is $\mathcal{O}(n\epsilon)$ close in trace distance to a feasible point and with value that is $n\|A\|\epsilon$ close, what we will call approximately feasible. We emphasize that the quantum algorithm also outputs a classical description of a solution that is approximately feasible in a sense that will be made precise below. The polynomial dependency on inverse accuracy is rather high (e.g. $(1/\epsilon)^{12}$ for the classical algorithm). We expect future work to be able to improve this.

Already the classical runtime improves upon the best known existing results and we refer to Section 2.5 for a detailed comparison. Access to a quantum computer would increase this gap further. However, it is important to point out that Theorem I has an approximation error of order $n\|A\|\epsilon$. In contrast, MMW [AHK05] – the fastest existing algorithm – incurs an error proportional to $\epsilon\|A\|_{\ell_1}$, where $\|A\|_{\ell_1} = \sum_{i,j} |A_{i,j}|$, making a straightforward comparison more difficult. Importantly, the scaling of our algorithm is favorable for generic problem instances and spin glass models, see Section 2.5.

The quantum algorithm outputs a classical description of a Hamiltonian $H^\sharp$ that encodes an approximately optimal, approximately feasible solution $\rho^\sharp = \exp(-H^\sharp)/\operatorname{tr}\left(\exp(-H^\sharp)\right)$ of the renormalized MaxQP SDP (3). This classical output can subsequently be used for randomized rounding for the $\infty \to 1$ norm of a matrix A, $\|A\|_{\infty\to 1} = \max_{x,y\in\{\pm 1\}^n} \langle x|A|y\rangle$.

Theorem II (Rounding)**.** *Suppose that $H^\sharp$ encodes an approximately optimal solution of the renormalized MaxQP SDP (3) with accuracy ϵ^4 for the target matrix*

$$A' = \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix},$$

where A is a $n \times n$ real matrix with at most s nonzero entries per column (column sparsity). Then, there is a classical $\tilde{\mathcal{O}}(ns)$-time randomized rounding procedure that converts $H^\sharp$ into binary vectors $\tilde{x}, \tilde{y} \in \{\pm 1\}^n$ that obey

$$\gamma\left(\|A\|_{\infty\to 1} - \mathcal{O}(n\|A\|\epsilon)\right) \leq \mathbb{E}\left[\langle \tilde{x}|A|\tilde{y}\rangle\right] \leq \|A\|_{\infty\to 1},$$

where $\gamma = \frac{2}{\pi}$ if A is positive semidefinite and $\frac{4}{\pi} - 1$ else.

This result recovers the randomized rounding guarantees of [AN06] in the limit of perfect accuracy ($\epsilon = 0$). However, for $\epsilon > 0$ the error scales with $n\|A\|$. In turn, randomized rounding only provides a multiplicative approximation if $\|A\|_{\infty\to 1}$ is of the same order. This result on the randomized rounding also relies on a detailed analysis of the stability of the rounding procedure w.r.t. to approximate solutions to the problem.

2 Detailed summary of results

We present *Hamiltonian Updates* – a meta-algorithm for solving convex optimization problems over the set of quantum states based on quantum Gibbs sampling – in a more general setting, as we expect it to find applications to other problems. Throughout this work, $\|\cdot\|_{tr}$ and $\|\cdot\|$ denote the trace (Schatten-1) and operator (Schatten-∞) norms, respectively.

2.1 Convex optimization and feasibility problems

SDPs over the set of quantum states are a special instance of a more general class of convex optimization problems. For a bounded, concave function f from the set of symmetric matrices to the real numbers and closed convex sets $\mathcal{C}_1, \ldots, \mathcal{C}_n$, solve

$$\begin{aligned} \text{maximize} \quad & f(X) && \text{(CPOPT)} \\ \text{subject to} \quad & X \in \mathcal{C}_1 \cap \cdots \cap \mathcal{C}_m, \\ & \operatorname{tr}(X) = 1,\ X \geq 0. \end{aligned} \tag{4}$$

The constraint $\operatorname{tr}(X) = 1$ enforces normalization, while $X \geq 0$ is the defining structure constraint of semidefinite programming. Together, they restrict X to the set of n-dimensional quantum states $\mathcal{S}_n = \{X : \operatorname{tr}(X) = 1,\ X \geq 0\}$. We will now specialize to the case $f(A) = \operatorname{tr}(AX)$ for a symmetric matrix A, as this is our main case of interest, but remark that it is simple to generalize the discussion that follows for more general classes. This trace normalization constraint implies fundamental bounds on the optimal value: $|\operatorname{tr}(AX^\sharp)| \leq \|A\|\|X^\sharp\|_{tr} = \|A\|$, according to Matrix Hölder [Bha97, Ex. IV.2.12]. Binary search over potential optimal values $\lambda \in [-\|A\|, \|A\|]$ allows for reducing the convex optimization problem into a sequence of feasibility problems:

$$\begin{aligned} \text{find} \quad & X \in \mathcal{S}_n && \text{(CPFEAS}(\lambda)\text{)} \\ \text{subject to} \quad & \operatorname{tr}(AX) \geq \lambda, \\ & X \in \mathcal{C}_1 \cap \cdots \cap \mathcal{C}_m. \end{aligned} \tag{5}$$

The convergence of binary search is exponential. This ensures that the overhead is benign: a total of $\log(\|A\|/\epsilon)$ queries of CPFEAS(λ) suffices to determine the optimal solution of CPOPT (4) up to accuracy ϵ. In summary:

Fact 2.1. *Binary search reduces the task of solving convex optimization problems* (4) *to the task of solving convex feasibility problems* (5)*.*

2.2 Meta-algorithm for approximately solving convex feasibility problems

We adapt a meta-algorithm developed by Tsuda, Rätsch and Warmuth [TRW05], see also [LRS15, AK16, Haz16, BKL+17] for similar ideas and [Bub15] for an overview of these techniques. All these algorithms, including the variation presented here, can be seen as instances of mirror descent with the mirror map given by the von Neumann entropy with adaptations tailored to the problem at hand. We believe our variation provides a path for also obtaining quantum speedups for nonlinear convex optimizations, so we state it in more detail.

For our algorithm, we require subroutines that allow for testing ϵ-closeness (in trace norm) to each convex set $\mathcal{C}_i$.

Definition 2.1 (ϵ-separation oracle)**.** *Let $\mathcal{C} \subset \mathcal{S}_n$ be a closed, convex subset of quantum states and $\mathcal{C}^* \subset \{X = X^\dagger \in \mathbb{C}^{n\times n} : \|X\| \leq 1\}$ be a closed, convex subset of observables of operator norm at most 1. For $\epsilon > 0$ an ϵ-separation oracle (with respect to $\mathcal{C}^*$) is a subroutine that either accepts a state ρ (in the sense that observables from $\mathcal{C}^*$ cannot distinguish ρ from*

elements of $\mathcal{C}$), or provides a hyperplane P that separates ρ from the convex set using a test from $\mathcal{C}^*$:

$$O_{\mathcal{C},\epsilon}(\rho) = \begin{cases} \textit{accept } \rho \textit{ if } \min_{Y\in\mathcal{C}} \max_{P\in\mathcal{C}^*} \operatorname{tr}(P(\rho - Y)) \leq \epsilon, \\ \textit{else: output } P \in \mathcal{C}^* \textit{ s.t. } \operatorname{tr}(P(\rho - Y)) \geq \frac{\epsilon}{2} \textit{ for all } Y \in \mathcal{C}. \end{cases}$$

We note that the Oracle is well-defined in the sense that if $\min_{Y\in\mathcal{C}} \max_{P\in\mathcal{C}^*} \operatorname{tr}(P(\rho-Y)) > \epsilon$, then there exists P such that $P \in \mathcal{C}^*$ and for all $Y \in \mathcal{C}$

$$\operatorname{tr}(P(\rho - Y)) \geq \frac{\epsilon}{2}.$$

Indeed, by Sion's min-max theorem [Sio58] we have

$$\max_{P\in\mathcal{C}^*} \min_{Y\in\mathcal{C}} \operatorname{tr}\left(P(\rho - Y)\right) = \min_{Y\in\mathcal{C}} \max_{P\in\mathcal{C}^*} \operatorname{tr}\left(P(\rho - Y)\right) > \epsilon.$$

This implies that there even exists a P that separates the state ρ from the set $\mathcal{C}$ by ϵ. Nonetheless, we instantiate the weaker requirement with only $\epsilon/2$ separation. This will be vital to ensure that the algorithm can tolerate errors and/or approximations in the samples from ρ.

By allowing for fine-tuning of $\mathcal{C}^*$ we are able to reduce the number of closeness conditions we need to test. *Hamiltonian Updates* (HU) a general meta-algorithm for approximately solving convex feasibility problems (5) (CPFEAS). The task is to find a state ρ that is ϵ-close to each convex set $\mathcal{C}_i$ with respect to observables in some $\mathcal{C}_i^*$ ($\max_{P_i\in\mathcal{C}_i^*} \min_{Y_i\in\mathcal{C}_i} \operatorname{tr}\left(P_i(\rho - Y_i)\right) \leq \epsilon$) and also obeys $\rho \in \mathcal{S}_n$ ($\rho \geq 0$ and $\operatorname{tr}(\rho) = 1$). A change of variables takes care of positive semidefiniteness and normalization: replace X in problem (5) by a Gibbs state $\rho_H = \exp\left(-H\right)/\operatorname{tr}(\exp(-H))$. At each iteration, we query ϵ-separation oracles. If they all accept, the current iterate is ϵ-close to feasible in the sense that there is a matrix in each C_i that is ϵ close in trace distance to the accepted state, and we are done. Otherwise, we update the matrix exponent to penalize infeasible directions: $H \to H + \frac{\epsilon}{16}P$, where P is a separating hyperplane that witnesses infeasibility. This process is visualized in Figure 1 and we refer to Algorithm 1 for a detailed description.

Algorithm 1 *Meta-Algorithm for approximately solving convex feasibility problems* (5).

Require: Query access to m ϵ-separation oracles $O_{1,\epsilon}(\cdot), \ldots, O_{m,\epsilon}(\cdot)$

function HAMILTONIANUPDATES(T, ϵ)
  $\rho = n^{-1}I$ and $H = 0$ ▷ initialize the maximally mixed state
  for $t = 1, \ldots, T$ **do**
    for $i = 1, \ldots, m$ **do** ▷ Query oracles and check feasibility
      if $O_{i,\epsilon}(\rho) = P$ **then**
        $H \leftarrow H + \frac{\epsilon}{16}P$ ▷ Penalize infeasible direction
        $\rho \leftarrow \exp\left(-H\right)/\operatorname{tr}(\exp(-H))$ ▷ Update quantum state
        break loop
      end if
    end for
    return (ρ, H) and **exit function** ▷ Current iterate is ϵ-feasible
  end for
end function

Theorem 2.1 (HU: convergence)**.** *Algorithm 1 requires at most* $T = \lceil 64 \log(n)/\epsilon^2 \rceil + 1$ *iterations to either certify that* (5) *is infeasible or output a state* ρ *satisfying:*

$$\text{for all } 1 \leq i \leq m : \max_{P_i \in \mathcal{C}_i^*} \min_{Y_i \in \mathcal{C}_i} \operatorname{tr}\left(P_i(\rho - Y_i)\right) \leq \epsilon \tag{6}$$

As it is also the case for the aforementioned variations of the algorithm above, the proof follows from establishing sufficiently large step-wise progress in quantum relative entropy. The quantum relative entropy between *any* feasible state and the initial state $\rho_0 = n^{-1}I$ (maximally mixed state) is bounded by $\log(n)$. Therefore, the algorithm must terminate after sufficiently many iterations. Otherwise, the problem is infeasible. We refer to Section 3.1 for details. Note that, unlike related previous quantum solvers [BKL+17, vAGGdW17, vAG19], our algorithm only considers the primal problem.

Theorem 2.1 has important consequences: The runtime of approximately solving quantum feasibility problems is dominated by the cost of implementing m separation oracles $O_{i,\epsilon}$ and the cost associated with matrix exponentiation. This reduces the task of efficiently solving convex feasibility problems to the quest of efficiently identifying separating hyperplanes and developing fast routines for computing Gibbs states.

The latter point already hints at a genuine quantum advantage: quantum architectures can efficiently prepare (certain) Gibbs states [CS17, Fra18, KBa16, PW09, TOV+09, TOV+09, YAG12, vAGGdW17].

It should be stressed that the approximate feasibility guarantee in (6) is not very strong and a careful choice of the C_i, C_i^* and a careful analysis of the continuity of the problem is usually required to ensure that it gives a good approximation to CPOPT (4).

2.3 Classical and quantum solvers for the renormalized MAXQP SDP

Let us now formulate the renormalized, relaxed problem in Eq. (3) in this framework and discuss the appropriate oracles. For fixed $\lambda \in [-1, 1]$ the (feasibility) MAXQP SDP is equivalent to a quantum feasibility problem:

$$\begin{aligned} \text{find} \quad & \rho \in \mathcal{S}_n \cap \mathcal{A}_\lambda \cap \mathcal{D}_n \\ \text{where} \quad & \mathcal{A}_\lambda = \left\{X : \operatorname{tr}\left(A\|A\|^{-1}X\right) \geq \lambda\right\}, \quad \mathcal{A}_\lambda^* = \left\{-A\|A\|^{-1}\right\} \\ & \mathcal{D}_n = \left\{X : \langle i|X|i\rangle = 1/n, i \in [n]\right\}, \quad \mathcal{D}_n^* = \left\{X : \|X\| \leq 1, X \text{ is diagonal}\right\}. \end{aligned}$$

The set $\mathcal{A}_\lambda$ corresponds to a half-space, while $\mathcal{D}_n$ is an affine subspace with codimension n. Let us now see that the convergence promises of Thm. 2.1 indeed convert to the renormalized, relaxed, feasibility MAXQP SDP, see Eq. (3). Let us start with observing

$$\max_{P \in \mathcal{A}_\lambda^*} \min_{Y \in \mathcal{A}_\lambda} \operatorname{tr}\left(P(\rho - Y)\right) \leq \epsilon \iff -\operatorname{tr}\left(A\|A\|^{-1}(\rho - Y)\right) \leq \epsilon \quad \text{for all } Y \in \mathcal{A}_\lambda. \tag{7}$$

Combined with the defining halfspace condition for $\mathcal{A}_\lambda$, this display asserts $\operatorname{tr}\left(A\|A\|^{-1}\rho\right) \geq \lambda - \epsilon$. We can analyze the oracle for $\mathcal{D}_n$ in a similar fashion. Note that,

$$\max_{P \in \mathcal{D}_n^*} \min_{Y \in \mathcal{D}_n} \operatorname{tr}\left(P(\rho - Y)\right) \leq \epsilon \iff \sum_{i=0}^{n-1} |\langle i|\rho|i\rangle - 1/n| \leq \epsilon. \tag{8}$$

Thus, we indeed obtain Eq. (3) from this formulation up to an error ϵ for the target value. It will be important to ensure that both quantum and classical algorithms work only having access to approximations of the current iteration. The simple structure of both sets readily suggests two separation oracles that take this into account:

$\mathcal{O}_{\mathcal{A}_\lambda}$: compute an approximation $\tilde{a} \in \mathbb{R}$ up to additive error $\frac{\epsilon}{4}$ of $\mathrm{tr}(A\|A\|^{-1}\rho)$. Check if $\tilde{a} \geq \lambda - \frac{3\epsilon}{4}$ and output $P = -A\|A\|^{-1}$ if this is not the case.

$\mathcal{O}_{\mathcal{D}_n}$: compute an approximation $\tilde{p} \in \mathbb{R}^n$ of $p(i) = \langle i|\rho|i\rangle$ satisfying $\sum_i |p(i) - \tilde{p}(i)| \leq \frac{\epsilon}{4}$. Check $\sum_i |\tilde{p}(i) - 1/n| \leq \frac{3\epsilon}{4}$ and output

$$P = \sum_{i=1}^{n} (\mathbb{I}\{\tilde{p}(i) > 1/n\} - \mathbb{I}\{\tilde{p}(i) < 1/n\})|i\rangle\langle i| \tag{9}$$

if this is not the case.

Note that the oracles are only defined for quantum states as inputs. Let us briefly check that it satisfies the definitions in 2.1. For $\mathcal{O}_{\mathcal{A}_\lambda}$ we have that if $\tilde{a} \geq \lambda - \frac{3\epsilon}{4}$, then $\mathrm{tr}(A\|A\|^{-1}\rho) \geq \lambda - \epsilon$, as desired. The other case is similar.

For the oracle for $\mathcal{O}_{\mathcal{D}_n}$, let us first assume that we are in the case that $\sum_i |\tilde{p}(i) - 1/n| \geq \frac{3\epsilon}{4}$. Clearly, we have that P defined in Eq. (9) is diagonal and of operator norm at most 1. For ease of notation let $f : [n] \to \{-1, 1\}$ be -1 if $\tilde{p}(i) < 1/n$ and 1 else. By construction we have for any $Y \in \mathcal{O}_{\mathcal{D}_n} \cap \mathcal{S}_n$:

$$\begin{aligned}
&\mathrm{tr}\left(P(\rho - Y)\right) = \sum_i f(i)\left(p(i) - \tfrac{1}{n}\right) \\
&\overset{(1)}{\geq} -\sum_i |p(i) - \tilde{p}(i)| + \sum_i f(i)\left(\tilde{p}(i) - \tfrac{1}{n}\right) = \sum_i |p(i) - \tilde{p}(i)| + \sum_i \left|\tilde{p}(i) - \tfrac{1}{n}\right| \\
&\geq -\frac{\epsilon}{4} + \sum_i |\tilde{p}(i) - \tfrac{1}{n}| \geq \frac{\epsilon}{2},
\end{aligned}$$

where in (1) we used Hölder's inequality. On the other hand, if $\sum_i |\tilde{p}(i) - 1/n| \geq \frac{3\epsilon}{4}$, then a similar argument shows that $\sum_i |p(i) - 1/n| \leq \epsilon$. Thus, we conclude that both oracles are correct.

The key insight to later obtain quantum speedups for the MaxQP SDP is that the second oracle can be interpreted as trying to distinguish the current state from the maximally mixed through computational basis measurements. This view is similar in spirit to [LRS15, Lemma 4.6], although here we focus on using this approach to construct solutions and to show that this notion of approximate feasibility is good enough for the MaxQP SDP.

2.3.1 Classical runtime

For fixed $\rho_H = \exp(-H)/\mathrm{tr}\left(\exp(-H)\right)$ both separation oracles are easy to implement on a classical computer given access to ρ_H. Hence, matrix exponentiation is the only remaining bottleneck. This can be mitigated by truncating the Taylor series for $\exp(-H)$ after $l' = \mathcal{O}(\|H\| + 1/\epsilon)$ many steps. Approximating ρ in this fashion only requires

$$\mathcal{O}(\min\left\{n^2 s, n^\omega\right\} \log(n)\epsilon^{-1})$$

steps and only incurs an error of ϵ in trace distance. Moreover, it is then possible to convert an approximately feasible to point to a strictly feasible one with a similar value, see Section 3.3. The following result becomes an immediate consequence of Fact 2.1 and Theorem 2.1.

Corollary 2.1 (Classical runtime for the MaxQP SDP)**.** *Suppose that A has row-sparsity s. Then, the classical cost of solving the associated (renormalized) MaxQP SDP up to additive error ϵ is $\mathcal{O}(\min\{n^2 s, n^\omega\}\log(n)\epsilon^{-12})$.*

The comparatively poor accuracy scaling with ϵ^{-12} stems largely from the fact that we need to convert an approximately feasible optimal solution into a strictly feasible optimal solution. This conversion is contingent on running Algorithm 1 with accuracy $\tilde{\epsilon} = \epsilon^4 \ll \epsilon$ (see Proposition 3.1 below). The total accuracy scaling $\epsilon^{-12} = \tilde{\epsilon}^{-3}$ results from combining the $\mathcal{O}(\log(n)/\tilde{\epsilon})$-cost for approximating the matrix exponential within a single iteration with the iteration bound $T = O(\log(n)/\tilde{\epsilon}^2)$ from Theorem 3.1.

2.3.2 Quantum runtime

Quantum architectures can efficiently prepare (certain) Gibbs states and are therefore well suited to overcome the main classical bottleneck. In contrast, checking feasibility becomes more challenging, because information about ρ is not accessible directly. Instead, we must prepare multiple copies of ρ and perform quantum mechanical measurements to test feasibility:

- $\mathcal{O}(\epsilon^{-2})$ copies of ρ suffice to ϵ-approximate $\operatorname{tr}(A\|A\|^{-1}\rho)$ via phase estimation.
- $\mathcal{O}(n\epsilon^{-2})$ copies suffice with high probability to estimate the diagonal entries of ρ (up to accuracy ϵ in trace norm) via repeated computational basis measurements.

Combining this with the overall cost of preparing a single Gibbs state implies the following runtime for executing Algorithm 1 on a quantum computer. This result is based on the *sparse oracle input model* and we refer to Sec. 3.4 for details.

Corollary 2.2 (Quantum runtime for the MaxQP SDP)**.** *Suppose that A has row-sparsity s. Then, the quantum cost of solving the MaxQP SDP up to additive error $\epsilon n\|A\|$ is*

$$\tilde{\mathcal{O}}(n^{1.5} s^{0.5+o(1)}\operatorname{poly}(1/\epsilon)).$$

The quantum algorithm also outputs a classical description of the Hamiltonian $H^\sharp$ corresponding to an approximately optimal, approximately feasible Gibbs state and its value. More precisely, it outputs a real number a and a diagonal matrix D such that $H^\sharp = aA + D$ and $n\rho_{H^\sharp}$ is $\mathcal{O}(n\epsilon)$ close in trace distance to a feasible point of MaxQP SDP. Moreover, we have the potential to produce samples from the associated approximately optimal Gibbs state $\rho^\sharp = \exp(-H^\sharp)/\operatorname{tr}\left(\exp(-H^\sharp)\right)$ in sub-linear runtime $\tilde{\mathcal{O}}(\sqrt{n})$ on a quantum computer. In the next section we show that the output of the algorithm is enough to give rise to good randomized roundings.

2.4 Randomized rounding

The renormalized MaxQP SDP (3) arises as a convex relaxation of an important quadratic optimization problem (1). However, the optimal solution $X^\sharp$ is typically not of the form $|x\rangle\langle x|$, with $x \in \{\pm 1\}^n$. Goemans and Williamson [GW95] pioneered randomized rounding techniques that allow for converting $X^\sharp$ into a cut $x^\sharp$ that is close to optimal. However, their rounding techniques rely on the underlying matrix being entrywise positive and a more delicate analysis is required to derive analogous results for broader classes of matrices. We will now follow the analysis of [AN06] to do the randomized rounding for the $\infty \to 1$ norm. First, let us make the connection between this norm and the MaxQP SDP clearer. Let A be a real matrix and define

$$A' = \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix}.$$

It is easy to see that for two binary vectors $x, y \in \{\pm 1\}^n$ we have $\langle x \oplus y|A'|x \oplus y\rangle = 2\langle x|A|y\rangle$ (with a slight abuse of notation, we also use the bra-ket notation for inner products of unnormalized vectors). This immediately shows that $2\|A\|_{1\to\infty} = \max_{z\in\{\pm 1\}^{2n}} \langle z|A'|z\rangle$, which is an instance of MaxQP SDP. We will now show that the rounding procedure is stable, i.e. randomized rounding of an approximately feasible, approximately optimal point, such as the ones outputted by the quantum algorithm, still result in a good binary vector for approximating this norm. We strengthen the stability of the rounding even further by showing that rounding with a truncated Taylor expansion of the solution is still good enough, saving runtime. The rounding procedure is described in Algorithm 2.

Algorithm 2 *Randomized rounding based on optimal Hamiltonian* $H^\sharp$

1: **function** RandomizedRounding($H^\sharp, \epsilon$)
2: Draw a random vector $g \in \mathbb{R}^n$ with i.i.d. $\mathcal{N}(0,1)$ entries.
3: Compute $z = \sum_{k=0}^{l} \frac{(-H^\sharp)^k}{2^k k!} g$ for $l = \mathcal{O}(\|H^\sharp\| + \log(1/\epsilon))$.
4: **output** $x_i = \mathrm{sign}(z_i)$.
5: **end function**

Proposition 2.1. *Let A be a real matrix and $H^\sharp$ be such that $\rho^\sharp = \exp(-H^\sharp)/\mathrm{tr}(\exp(-H^\sharp))$ is an ϵ-approximate solution to the renormalized MaxQPSDP for A' (3) with value $\alpha^\sharp = \mathrm{tr}\left(A'\|A'\|^{-1}\rho^\sharp\right)$. Then, the (random) output $x = (x_1 \oplus x_2) \in \{\pm 1\}^{2n}$ of Algorithm 2 can be computed in $\tilde{\mathcal{O}}(ns)$-time and obeys*

$$\gamma n\|A\| \left(\alpha^\sharp - \mathcal{O}(\epsilon)\right) \leq \mathbb{E}\langle x_1|A|x_2\rangle \leq n\|A\|(\alpha^\sharp + \mathcal{O}(\epsilon)),$$

where $\gamma = 2/\pi$ for A p.s.d. and $4/\pi - 1$ else.

This rounding procedure is fully classical and can be executed in runtime $\tilde{\mathcal{O}}(ns)$. We refer to Sec. 3.5 for details. What is more, it applies to both quantum and classical solutions of the MaxQP SDP. Even the quantum algorithm provides $H^\sharp$ in classical form, while the associated $\rho^\sharp$ is only available as a quantum state. Rounding directly with $\rho^\sharp$ would necessitate a fully quantum rounding technique that, while difficult to implement and analyze, seems

to offer no advantages over the classical Algorithm 2. Thus, it is possible to perform the rounding even with the output of the quantum algorithm. We prove this theorem in two steps. First, we follow the proof technique of [AN06] to show that our relaxed notion of approximately feasible is still good enough to ensure a good rounding in expectation. This shows that our notion of feasibility is strong enough for the problem at hand. The stability of the rounding w.r.t. to truncation of the Taylor series then follows by showing appropriate anticoncentration inequalities for the random vector.

Note that in [AN06] the authors prove that the constant $\frac{2}{\pi}$ in Proposition 2.1 is optimal.

2.5 Comparison to existing work

The MaxQP SDP has already received a lot of attention in the literature. Table 1 contains a runtime comparison between the contributions of this work and the best existing classical results [AHK05, AK16]. It highlights regimes, where we obtain both classical and quantum speedups. In a nutshell, Hamiltonian Updates outperforms state of the art algorithms whenever the target matrix A has both positive and negative off-diagonal entries and the optimal value of the SDP scales as $n\|A\|$. It is worthwhile to explore the following examples.

2.5.1 Quadratic quantum speedups and classical speedups for generic instances and spin glasses

Recall that Hamiltonian Updates can only offer speedups for MaxQP SDP instances where the optimal value scales like $n\|A\|$, as opposed to the $\|A\|_{\ell_1}$ scaling required for MMW. Intuitively speaking, such a scaling should arise whenever A has both positive and negative entries, causing cancellations. In order to formalize this intuition, we show that Hamiltonian Updates offers speedups for generic matrices that have both positive and negative entries, see Appendix A for details. Our main result is as follows. Suppose that A is a random Hermitian matrix with entries

$$A_{ij} = \tau_{ij}(g_{ij} + \lambda), \tag{10}$$

where g_{ij} are independent centered random variables with bounded fourth moment, τ_{ij} is a Bernoulli random variable with parameter p and $\lambda > 0$ is some fixed parameter. This random generative model covers many relevant MaxQP instances. Note that if we set $p = \frac{s}{n}$, the matrix A is $\mathcal{O}(s)$ sparse in expectation. Let us first discuss the (centered) $\lambda = 0$-case in more detail. There, $\mathbb{E}\|A\|_{\ell_1} = \Theta(ns)$, $\|A\|_{\infty\to 1} = \Theta(n\sqrt{s})$, $\mathbb{E}\|A\| = \Theta(\sqrt{s})$ and concrete realizations of A concentrate sharply around these expected values. These concentration arguments are derived in Appendix A and imply that, indeed, $n\|A\|$, and not $\|A\|_{\ell_1}$, provides the right scaling for such generic instances. The scaling for MMW [AHK05] is $\tilde{\mathcal{O}}(\min\{(n/\epsilon)^{2.5}s, n^3\alpha^{-1}\|A\|_{\ell_1}\epsilon^{-3.5}\})$ to achieve an error of $\epsilon\|A\|_{\ell_1}$. Thus, to obtain a multiplicative error for such instances using MMW we need to divide ϵ by $s^{-\frac{1}{2}}$, yielding an expected scaling of $\tilde{\mathcal{O}}(\min\{(n/\epsilon)^{2.5}s^{4.5}, n^3 s^{2.25}\epsilon^{-3.5}\})$. This implies that the runtime of Hamiltonian Updates improves upon MMW [AHK05], both classically and quantumly.

To the best of our knowledge, the quantum implementation of Hamiltonian Updates establishes the first quantum speedup for problems of this type. Corollary 2.2 establishes a nearly quadratic speedup for generic MaxQP SDP instances compared to the current state of the art SDP solvers.

It is worth noting that the random matrix defined in Eq. (10) corresponds to a widely studied model in spin glasses: the *(diluted) Sherrington-Kirkpatrick (SK) model* [Pan13, Tal11]. This problem has received considerable attention in the statistical physics literature. In particular, recent work [Mon19] shows that, under some conjectures, it is possible to approximately solve the quadratic optimization in (1) with high probability in time $\tilde{\mathcal{O}}(n^2)$ for the standard, undiluted SK model ($\tau_{ij} = 1$). This is the same time complexity as our quantum solver, as the target matrix of these instances is dense ($s = \Omega(n)$). To the best of our knowledge, a variation of [Mon19] for the diluted model ($p < 1$) has not yet been discussed.

Furthermore, there is an integrality gap for the SDP relaxation of this problem in the Gaussian setting [KB20, MS16] whenever $\lambda = 0$. As we discuss in more detail in Appendix B, this implies that the value of the problem in the case $\tau_{ij} = 1$ converges to the largest eigenvalue of A in the limit $n \to \infty$. On top of that, [MS16] gives a construction of an approximately optimal feasible point that can be computed in $\mathcal{O}(n^\omega)$ time, where $\omega > 2$ is the exponent of matrix multiplication. Correspondingly, we do not obtain a classical speedup for such instances. Once again we refer to Appendix B for more details and we are not aware of similar results for the diluted model.

Let us now discuss the undiluted case with $\lambda > 1$, as the behaviour of the model is not qualitatively different for $\lambda < 1$ [MS16]. To the best of our knowledge, the exact value of the MaxQP SDP is not known for this setting. But, numerical evidence suggests that there is no integrality gap [MS16, JMRT16], and no constructions of approximately optimal points are known. Thus, we expect that it is this regime, where we obtain both quantum and classical speedups.

2.5.2 Speedups for MaxCut and the hidden partition problem:

Additional structure can substantially reduce the runtime of existing MMW solvers [AK16]. For weighted MaxCut, in particular, A is related to the adjacency matrix of a graph and has exclusiely non-negative entries. This additional structure facilitates the use of powerful dimensionality reduction and sparsification techniques that outperform our algorithm for general graphs. Recently, it was shown that quantum algorithms can speed up spectral graph sparsification techniques [Ad20]. As the sparsification step dominates the complexity of these algorithms, this leads to faster solvers for MaxCut, albeit solving the sparsified SDP on a classical computer. However, these sparsification techniques do not readily apply to general problem instances, where the entries of A can have both positive and negative signs (sign problem). We refer to Appendix C for a more detailed discussion.

More direct speedups do, however, apply for approximating MaxCut in Erdös-Rényi graphs. An Erdös-Rényi graph $G(n, p)$ with n vertices is a random graph in which each edge is present independently at random with probability p. One can show that for such graphs, a random balanced partition of the vertices achieves an expected cut of value $\frac{n^2 p}{2}$. Thus, obtaining a cut up to an approximation of order ϵm, where m is the number of edges, is trivial for random graphs: just sample a random one. In [DMS17] the authors show that

whenever $pn \to \infty$, the MaxCut of such graphs satisfies :

$$\frac{n^2 p}{2} + \left(\frac{n^3 p(1-p)}{2}\right)^{\frac{1}{2}} P_* + o(n^{\frac{3}{2}}), \tag{11}$$

where P_* is the so-called Parisi constant. Thus, obtaining approximations to the MaxCut of such graphs is only interesting whenever we can achieve an error scaling as $\mathcal{O}(n^{\frac{3}{2}}\sqrt{p})$ and the usual Goemans-Wiliamson relaxation is not suitable. In order to address this issue, Montanari et al. [MS16] showed that is advisable to instead solve MaxQP SDP relaxation for the matrix

$$B = A - p\mathbf{1}^T\mathbf{1}, \tag{12}$$

where $\mathbf{1}$ is the all ones vector and A the (random) adjacency matrix of the graph. This then has the value $2n^{\frac{3}{2}}\sqrt{p} + o(\sqrt{p/n})$ with high probability. See [Theorem 1][MS16] for more details. Note that the matrix in Eq. (12) has both negative and positive entries with expected value 0 and bounded variance. We conclude that we are in the same setting as in the spin glasses for this dense instance and, thus, we also obtain speedups compared to MMW. However, once again the recent work [Mon19] shows that, under some conjectures, it is possible to approximately solve the underlying MaxQP for B directly in time $\mathcal{O}(n^2)$ with high probability.

Another relevant random graph model is that of the planted partition, whose distribution we will denote by $G(n, a/n, b/n)$ for parameters $a, b > 0$. This distribution over graphs with n vertices is defined as follows. First, we partition the n vertices into two subsets S_1, S_2 with $|S_1| = n/2$ uniformly at random. Conditional on this partition we pick the edges independently at random with probabilities:

$$\mathbb{P}((i,j) \in E | S_1, S_2) = \begin{cases} \frac{a}{n}, & \text{if } \{i,j\} \subset S_1 \text{ or } \{i,j\} \subset S_1, \\ \frac{b}{n}, & \text{else.} \end{cases}$$

Solving the MaxQP SDP for the target matrix described in Eq. (12) with $p = \frac{a+b}{2}$ is relevant to solving the planted partition problem [MS16] and closely related to the model in Eq. (10) with $\lambda = \frac{a-b}{\sqrt{2(a-b)}}$. We refer to [MS16] for details on this, but roughly speaking the problem is to decide if a graph was sampled from Erdös-Rényi distribution with parameter p or from the planted partition with $p = \frac{a+b}{2}$. Note that also for the planted partition the adjacency matrix in Eq. (12) satisfies the conditions under which we obtain speedups.

In [MS16, Theorem 3] the authors show that for certain parameter ranges of a, b, solving the MaxQP SDP in Eq. (12) and using its value to decide which distribution we sampled from gives rise to a good test for this problem. As for both the planted partition and the Erdös-Rényi model the MaxQP SDP in Eq. (12) can be solved faster with our methods, we obtain a speedup for this problem.

2.5.3 Previous quantum SDP solvers:

Previous quantum SDP solvers [BS17, vAGGdW17, vAG19, BKL+17] with inverse polynomial dependence on the error do not provide speedups for solving the MaxQP SDP in the

worst case, as their complexity depends on a problem specific parameter, the width of the SDP. We refer to the aforementioned references for a precise definition of this parameter and for the complexity of the solvers under different input models. As shown in [vAGGdW17, Theorem 24], the width parameter scales at least linearly in the dimension n for what they call *combinable SDPs* [Definition 23][vAGGdW17]. In a nutshell, these are SDP classes for which direct sum combinations of two instances and constraints yield another valid SDP in the class and we refer to [vAGGdW17] for a precise definition. For our purposes, it suffices to note that the MaxQP SDP class of SDPs is combinable, as shown in [vAGGdW17]. Although the authors only observe that their conditions apply to MaxCut, it is easy to see that their results do not require any assumptions on the sign of entries of A. Thus, their results show that the MaxQP SDP also admits instances with linearly growing width. To the best of our knowledge, the solvers mentioned above have a dependence that is at least quadratic in the width and at least a $n^{\frac{1}{2}}$ dependence on the dimension. Thus, the combination of the term stemming from the width and the dimension already gives a higher complexity than our solver. One reason why we bypass these restrictions is that we do not use the primal-dual approach to solve the SDP from the aforementioned references.

Although this gives an indication as to why our algorithm might be better suited for MAX QP in the worst case, it does not necessarily means that our algorithm outperforms the aforementioned ones on the random instances discussed before on average. In Prop. B.1 of Appendix B we show that for the random model in (10) with $\lambda > 1$ it is indeed the case that the width scales linearly with the dimension with high probability, albeit under the assumption that the problem does not have an intregrality gap. The absence of an integrality gap is supported by the numerics of [MS16, JMRT16]. These results show that previous quantum SDP solvers are likely not to provide a speedup on average for such instances with $\lambda > 1$. On the other hand, we also show that for $\lambda < 1$, the width does not necessarily scale with system size. These results certainly motivate further studies on the width of such randomized instances, also for the random graph models.

Another, and arguably conceptually more interesting, reason why our algorithm outperforms other solvers is how we enforce the diagonal constraint.

Enforcing that each diagonal constraint of the renormalized MAXQP SDP in Eq. (3) is satisfied up to an additive error, i.e.

$$|\langle i|\rho|i\rangle - 1/n| \leq \epsilon$$

would require an error ϵ of order n^{-1} to ensure a solution with a quality comparable to ours. This would translate into a width parameter that scales linearly in n in the worst case. What is even worse, we do not know any better width bounds for special cases of the MaxQP SDP. This severely limits the scope of existing quantum SDP solvers – they do not readily apply and have worse runtimes than available classical algorithms. Let us illustrate this by example. In [KP20], the authors give a quantum SDP solver whose complexity is $\widetilde{O}\left(\frac{n^{2.5}}{\xi^2}\mu\kappa^3 \log\left(\epsilon^{-1}\right)\right)$. Here κ and μ are again problem-specific parameters and ξ is the precision to which each constraint is satisfied. As noted before, a straightforward implementation of the MAXQP SDP requires ξ to be at most of order n^{-1}, which establishes a runtime of order at least $n^{4.5}$ using those methods. Thus, we conclude that all current quantum SDP solvers do not offer speedups over state of the art classical algorithms, see Table 1 for more details.

Algorithm	Runtime	Error	Speedup
This work (Classical)	$\tilde{\mathcal{O}}(\min\{n^2 s, n^\omega\}\epsilon^{-12})$	$\epsilon n\|A\|$	-
This work (Quantum)	$\tilde{\mathcal{O}}(n^{1.5}s^{0.5+o(1)}\epsilon^{-28})$	$\epsilon n\|A\|$	-
MMW [AHK05]	$\tilde{\mathcal{O}}(\min\{(n/\epsilon)^{2.5}s, n^3\alpha^{-1}\|A\|_{\ell_1}\epsilon^{-3.5}\})$	$\epsilon\|A\|_{\ell_1}$	$\|A\|_{\ell_1} \geq n\|A\|$, $\epsilon = \Theta(1)$
Interior Point [LSW15]	$\mathcal{O}(n^{\omega+1}\log(\epsilon^{-1}))$	ϵ	$\epsilon = \Theta(1)$
MMW for MaxCut [AK16] (non-negative entries only)	$\tilde{\mathcal{O}}(ns)$	$\epsilon\|A\|_{\ell_1}$	Erdös-Rényi random graphs

Table 1: comparison of different classical algorithms to solve the original MaxQP SDP (2). The speedup column clarifies in which regimes we obtain speedups and ω denotes the exponent of matrix multiplication. Here α corresponds to the value of MAXQP SDP.

This discussion showcases that our technique to relax the diagonal constraints gives rise to a novel way of enforcing constraints that allows for better control of errors in quantum SDP solvers and could be used for other relevant SDPs. Moreover, the fact that the approximate solution can still be used to obtain good roundings highlights the fact that our notion of approximate feasibility does not render the problem artificially easy.

Finally, we want to point out that subtleties regarding error scaling do not arise for MaxCut. If A is the adjacency matrix of a d-regular graph on n vertices, then $n\|A\| = nd = \|A\|_{\ell_1}$ and the different errors in Table 1 all coincide.

3 Technical details and proofs

3.1 Proof of Theorem 2.1

By construction, Algorithm 1 (Hamiltonian Updates) terminates as soon as it has found a quantum state ρ that is ϵ-close to being feasible. Correctly flagging infeasibility is the more interesting aspect of Theorem 2.1 (convergence to feasible point). Several variations of the statement and proof below can be found in the literature [TRW05, Haz16, AK16, LRS15, Bub15, BKL+17, ACH+19], but we present it for completeness.

Lemma 3.1. *Suppose Algorithm 1 does not terminate after $T = \lceil 64\log(n)/\epsilon^2\rceil + 1$ steps. Then, the feasibility problem (5) is infeasible.*

Proof. By contradiction. Suppose there exists a feasible point ρ^* in the intersection of all $m+1$ sets and we ran the algorithm for T steps. Instantiate the short-hand notation $\rho_t = \rho_{H_t} = \exp(-H_t)/\mathrm{tr}(\exp(-H_t))$ for the t-th state and Hamiltonian in Algorithm 1. Initialization with $H_0 = 0$ and $\rho_0 = I/n$ is crucial, as it implies that the quantum relative entropy between ρ^* and ρ_0 is bounded:

$$S(\rho^*\|\rho_0) = \mathrm{tr}(\rho^*(\log\rho^* - \log\rho_0)) \leq \log(n).$$

We will now show that the relative entropy between successive (infeasible) iterates ρ_{t+1}, ρ_t and the feasible state ρ^* necessarily decreases by a finite amount. Let P_t be the hyperlane that

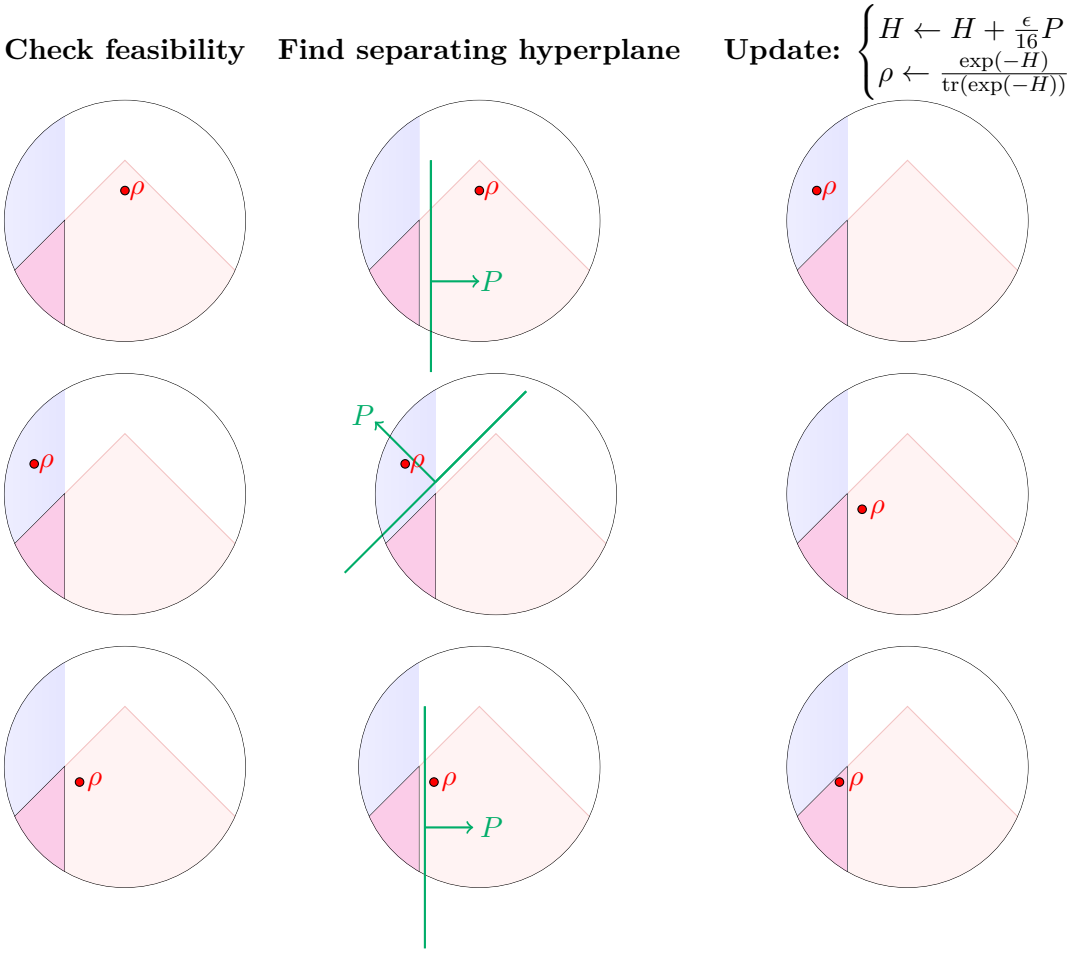


Figure 1: *Caricature of Hamiltonian Update iterations in Algorithm 1:* Schematic illustration of the intersection of three convex sets (i) a halfspace (blue), (ii) a diamond-shaped convex set (red) and (iii) the set of all quantum states (clipped circle). Algorithm 1 (Hamiltonian Updates) approaches a point in the convex intersection (magenta) of all three sets by iteratively checking feasibility (left column), identifying a separating hyperplane (central column) and updating the matrix exponent to penalize infeasible directions (right column).

separates ρ_t from the feasible set providade by the oracle. The update rule $H_{t+1} = H_t + \frac{\epsilon}{16} P_t$ then asserts

$$\begin{aligned} S(\rho^* \| \rho_{t+1}) - S(\rho^* \| \rho_t) &= \operatorname{tr}(\rho^*(H_{t+1} - H_t)) + \log\left(\frac{\operatorname{tr}(\exp(-H_{t+1}))}{\operatorname{tr}(\exp(-H_t))}\right) \\ &= \tfrac{\epsilon}{16} \operatorname{tr}(P_t \rho^*) - \log\left(\frac{\operatorname{tr}\left(\exp\left(-H_{t+1} + \frac{\epsilon}{16} P_t\right)\right)}{\operatorname{tr}(\exp(-H_{t+1}))}\right). \end{aligned} \tag{13}$$

The logarithmic ratio can be bounded using the Peierls-Bogoliubov inequality [AL70, Lemma 1]: $\log(\operatorname{tr}(\exp(F+G))) \geq \operatorname{tr}(F \exp(G))$ provided that $\operatorname{tr}(\exp(G)) = 1$. This implies

$$\begin{aligned} \log\left(\frac{\operatorname{tr}\left(\exp\left(-H_{t+1} + \frac{\epsilon}{16} P_t\right)\right)}{\operatorname{tr}(\exp(-H_{t+1}))}\right) &= \log\left(\operatorname{tr}\left(\exp\left(-H_{t+1} - \log\left(\operatorname{tr}(\exp(-H_{t+1}))\right) I + \tfrac{\epsilon}{16} P_t\right)\right)\right) \\ &\geq \operatorname{tr}\left(\tfrac{\epsilon}{16} P_t \exp\left(-H_{t+1} - \log(\operatorname{tr}(\exp(-H_{t+1}))) I\right)\right) \\ &= \tfrac{\epsilon}{16} \operatorname{tr}(P_t \exp(-H_{t+1}) / \operatorname{tr}(\exp(-H_{t+1}))) = \tfrac{\epsilon}{16} \operatorname{tr}(P_t \rho_{t+1}). \end{aligned} \tag{14}$$

Combining Eq. (13) with Eq. (14) we arrive at

$$S(\rho^* \| \rho_{t+1}) - S(\rho^* \| \rho_t) \leq \tfrac{\epsilon}{16} \operatorname{tr}(P_t(\rho^* - \rho_{t+1})).$$

Next, note that the updates are mild in the sense that ρ_{t+1} and ρ_t are close in trace distance. [BS17, Lem. 16] implies $\|\rho_{t_1} - \rho_t\|_{tr} \leq 2\left(\exp(\tfrac{\epsilon}{16}\|P_t\|) - 1\right) \leq \tfrac{\epsilon}{4}$, because $\|P_t\| \leq 1$ by construction and we can also assume $\tfrac{\epsilon}{16} \leq \log(2)$. Combining these insights with Matrix Hölder [Bha97, Ex. IV.2.12] ensures

$$\begin{aligned} S(\rho^* \| \rho_{t+1}) - S(\rho^* \| \rho_t) &\leq \tfrac{\epsilon}{16} \operatorname{tr}(P_t \rho^*) - \tfrac{\epsilon}{16} \operatorname{tr}(P_t \rho_{t+1}) \\ &= \tfrac{\epsilon}{16} \left(\operatorname{tr}(P_t(\rho_t - \rho_{t+1})) - \operatorname{tr}(P_t(\rho_t - \rho^*))\right) \\ &\leq \tfrac{\epsilon}{16} \left(\|P_t\| \|\rho_t - \rho_{t+1}\|_{tr} - \operatorname{tr}(P_t(\rho_t - \rho^*))\right). \end{aligned}$$

The first contribution is bounded by $\tfrac{\epsilon}{4}\|P_t\| \leq \tfrac{\epsilon}{4}$, while Definition 2.1 ensures $\operatorname{tr}(P_t(\rho_t - \rho^*)) \geq \tfrac{\epsilon}{2}$ (ρ^* is feasible and P_t is an $\tfrac{\epsilon}{2}$-separation oracle for the infeasible point ρ_t). In summary,

$$S(\rho^* \| \rho_{t+1}) - S(\rho^* \| \rho_t) \leq \tfrac{\epsilon}{16}\left(\tfrac{\epsilon}{4} - \tfrac{\epsilon}{2}\right) = -\tfrac{\epsilon^2}{64} \quad \text{for all iterations} \quad t = 0, \ldots, T$$

and we conclude

$$S(\rho^* \| \rho_T) = \sum_{t=0}^{T} \left(S(\rho^* \| \rho_{t+1}) - S(\rho^* \| \rho_t)\right) + S(\rho^* \| \rho_0) \leq -T\tfrac{\epsilon^2}{64} + \log(n).$$

This expression becomes negative as soon as the total number of steps T surpasses $64\log(n)/\epsilon^2$. A contradiction, because the quantum relative entropy is always non-negative. □

3.2 Stability of the relaxed MaxQP SDP

Note that even if Algorithm 1 accepts a candidate point, it does not necessarily mean that this point is exactly feasible. Theorem 2.1 only asserts that this point is ϵ-close to all sets

of interest with respect to a set of observables. For the MaxQP SDP (3), this means that the outputs of the algorithm will only satisfy the diagonal constraints approximately and, in principle, the value of this further relaxed problem could differ significantly from the original value. In the next Proposition we show that this is not the case:

Proposition 3.1. *Let $\alpha_{\epsilon^4} = \operatorname{tr}(A\rho)$ be the value attained by – up to accuracy ϵ^4 – solution ρ to the relaxed MaxQP SDP (3) with input matrix A. Then there is a quantum state $\rho^\sharp$ at trace distance $\mathcal{O}(\epsilon)$ of ρ such that $n\rho^\sharp$ is a feasible point of MaxQP SDP (2). In particular:*

$$\left|\alpha_{\epsilon^4} n\|A\| - \operatorname{tr}\left(n\rho^\sharp A\right)\right| = \mathcal{O}(\epsilon n\|A\|), \tag{15}$$

Moreover, it is possible to construct $\rho^\sharp$ in time $\mathcal{O}(n^2)$ given the entries of ρ.

Proof. Let ρ be a solution to the relaxed MaxQP SDP (3) with relaxation parameter ϵ^4. We will now construct $\rho^\sharp$ such that $n\rho^\sharp$ is an exactly feasible point of the MaxQP SDP (3). These modifications are mild enough to ensure that the associated SDP value will only change by $\mathcal{O}(\epsilon n\|A\|)$. We proceed in two steps: (i) $\rho \mapsto \rho'$: Identify diagonal entries that substantially deviate from $1/n$ in the sense that $|\langle i|\rho|i\rangle - 1/n| > \epsilon^2/n$. Subsequently, replace ρ_{ii} by $1/n$ and set all entries in the i-th row and i-th column to zero. This ensures that ρ' remains positive semidefinite. (ii) $\rho' \to R$: Replace all remaining diagonal entries by $1/n$. This may thwart positive semidefiniteness, but the following convex combination restores this feature:

$$\rho^\sharp = \tfrac{1}{1+\epsilon^2}\left(R + \tfrac{\epsilon^2}{n} I\right).$$

By construction, this matrix is both psd and obeys $\langle i|\rho^\sharp|i\rangle = 1/n$ for all $i \in [n]$. In words: it is a feasible point of the renormalized MaxQP SDP (3).

We now show that these reformulations are mild. To this end, let $B = \{i : \ |n\langle i|\rho|i\rangle - 1| > \epsilon^2\} \subset [n]$ be the indices associated with large deviations. Without loss of generality, we can assume that these are the first $|B|$ indices. Then,

$$\begin{aligned}\|\rho' - \rho\|_{tr} =& \left\| \begin{pmatrix} n^{-1}I_B & 0 \\ 0 & \rho_{22} \end{pmatrix} - \begin{pmatrix} \rho_{11} & \rho_{12} \\ \rho_{21} & \rho_{22} \end{pmatrix} \right\|_{tr} = \left\| \begin{pmatrix} n^{-1}I_B - \rho_{11} & -\rho_{12} \\ -\rho_{21} & 0 \end{pmatrix} \right\|_{tr} \\ \leq& \|\rho_{11}\|_{tr} + 2\|\rho_{12}\|_{tr} + \|n^{-1}I_B\|_{tr}. \end{aligned} \tag{16}$$

Next, note that ϵ^4-approximate feasibility implies $\sum_{i=1}^n |\langle i|\rho|i\rangle - 1/n| \leq \epsilon^4$. This, in turn, demands $|B|\frac{\epsilon^2}{n} \leq \epsilon^4$ or, equivalently $|B| \leq n\epsilon^2$. The definition of B moreover asserts

$$\|\rho_{22}\|_{tr} \geq (n - |B|)\tfrac{1-\epsilon^2}{n} \geq (1-\epsilon^2)^2.$$

Moreover, as shown in [Kin03], we have

$$\left\| \begin{bmatrix} \|\rho_{11}\|_{tr} & \|\rho_{12}\|_{tr} \\ \|\rho_{12}^T\|_{tr} & \|\rho_{22}\|_{tr} \end{bmatrix} \right\|_{tr} \leq \left\| \begin{bmatrix} \rho_{11} & \rho_{12} \\ \rho_{12}^T & \rho_2 \end{bmatrix} \right\|_{tr} = \|\rho\|_{tr} = \operatorname{tr}(\rho) = 1.$$

As $\|\cdot\|_{tr} \geq \|\cdot\|_2$ (the Frobenius, or Schatten-2 norm), it follows from the last equation that

$$\|\rho_{11}\|_{tr}^2 + 2\|\rho_{12}\|_{tr}^2 + \|\rho_{22}\|_{tr}^2 \leq 1.$$

And, as $\|\rho_{22}\|_{tr}^2 \geq (1-\epsilon^2)^4$, we conclude $\|\rho_{11}\|_{tr}^2 + 2\|\rho_{12}\|_{tr}^2 = \mathcal{O}(\epsilon^2)$. which in turn implies $\|\rho_{11}\|_{tr} + 2\|\rho_{12}\|_{tr} = \mathcal{O}(\epsilon)$. Inserting this relation into Eq. (16) yields

$$\|\rho' - \rho\|_{tr} = \mathcal{O}(\epsilon). \tag{17}$$

Next, note that we obtain R from ρ' by just replacing all diagonal entries of ρ' by $1/n$. As by construction all the diagonal elements of ρ' are in the range $1/n \pm \epsilon^2/n$, we can write

$$R = \rho' + D,$$

where D is a diagonal matrix whose entries are in the range $[-\epsilon^2/n, \epsilon^2/n]$. Thus, $D + \frac{\epsilon^2}{n} I$ is psd. Normalizing the trace we see that

$$\rho^\sharp = \frac{1}{1+\epsilon^2}\left(\rho' + D + \frac{\epsilon^2}{n} I\right)$$

is psd with diagonal entries $1/n$ and, thus, $n\rho^\#$ is a feasible point of MaxQP SDP (2). We also have that:

$$\|\rho' - \rho^\sharp\|_{tr} = \frac{1}{1+\epsilon^2}\|\epsilon^2\rho' + D + \epsilon^2 \frac{I}{n}\|_{tr} = \mathcal{O}(\epsilon^2). \tag{18}$$

by a triangle inequality. Thus, combining Eq. (18) and Eq. (17) we conclude from another triangle inequality that:

$$\|\rho - \rho^\sharp\|_{tr} = \mathcal{O}(\epsilon).$$

The claim then follows from a (matrix) Hölder inequality:

$$\left|\operatorname{tr}(nA\rho) - \operatorname{tr}\left(nA\rho^\sharp\right)\right| \leq n\|A\|\|\rho - \rho^\sharp\|_{tr} = \mathcal{O}(n\|A\|\epsilon).$$

Note that the proof technique above is constructive and allows us to construct a feasible point from an approximately feasible one in $\mathcal{O}(n^2)$ time by manipulating the entries. □

3.3 Approximately solving the MaxQP SDP on a classical computer

We will now show how to use Hamiltonian Updates (Algorithm 1) to solve the MaxQP SDP (3) on a classical computer. It turns out that the main classical bottleneck is the cost of computing matrix exponentials $\rho = \exp(-H)/\operatorname{tr}(\exp(-H))$. The following result, also observed in [LRS15], asserts that coarse truncations of the matrix exponential already yield accurate approximations.

Lemma 3.2. *Fix a Hermitian $n \times n$ matrix H, an accuracy ϵ and let l be the smallest even number that obeys $(l+1)(\log(l+1)-1) \geq 2\|H\| + \log(n) + \log(1/\epsilon)$. Then, the truncated matrix exponential $T_l = \sum_{k=0}^{l} \frac{1}{k!}(-H)^k$ is guaranteed to obey*

$$\left\| \frac{\exp(-H)}{\operatorname{tr}(\exp(-H))} - \frac{T_l}{\operatorname{tr}(T_l)} \right\|_{tr} \leq \epsilon.$$

Proof. First note, that truncation at an even integer l ensures that T_l is positive semidefinite. This is an immediate consequence of the fact that even-degree Taylor expansions of the (scalar) exponential are non-negative polynomials. In particular, $\|T_l\|_{tr} = \operatorname{tr}(T_l)$. Combine this with $\operatorname{tr}(X) \leq \|X\|_{tr} \leq n\|X\|$ for all Hermitian $n \times n$ matrices to conclude

$$\left\| \frac{\exp(-H)}{\operatorname{tr}(\exp(-H))} - \frac{T_l}{\operatorname{tr}(T_l)} \right\|_{tr} \leq \frac{1}{\operatorname{tr}(\exp(-H))} \|\exp(-H) - T_l\|_{tr} + \frac{|\operatorname{tr}(\exp(-H)) - \operatorname{tr}(T_l)|}{\operatorname{tr}(T_l)\operatorname{tr}(\exp(-H))} \|T_l\|_{tr}$$

$$\leq \frac{2\|\exp(-H) - T_l\|_{tr}}{\operatorname{tr}(\exp(-H))} \leq 2n \exp(\|H\|) \|\exp(-H) - T_l\|,$$

where we have also used $\operatorname{tr}(\exp(-H)) \geq \|\exp(-H)\| \geq \exp(-\|H\|)$. By construction, both $\exp(-H)$ and T_l commute and are diagonal in the same eigenbasis. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of H. Then, Taylor's remainder theorem asserts

$$\|\exp(-H) - T_l\| = \max_{1 \leq i \leq n} \left| \exp(-\lambda_i) - \sum_{k=0}^{l} \tfrac{1}{k!} (-\lambda)^k \right| \leq \frac{\max_i \exp(-\lambda_i)}{(l+1)!} \leq \frac{\exp(\|H\|)}{(l+1)!}.$$

The value of l is chosen such that

$$\frac{2n \exp(2\|H\|)}{(l+1)!} \leq \exp(2\|H\| + \log(2) + \log(n) - 1 - (l+1)(\log(l+1) - 1)) \leq \epsilon,$$

because $(l+1)! \geq \mathrm{e}\left((l+1)/\mathrm{e}\right)^{l+1}$. □

Corollary 3.1. *Given an s sparse, symmetric $n \times n$ matrix A and $\epsilon > 0$, we can solve the MaxQP SDP (3) up to an additive error $\mathcal{O}(\epsilon n \|A\|)$ in time $\tilde{\mathcal{O}}(\min\{n^2 s, n^\omega\}\epsilon^{-12})$ on a classical computer.*

Although the dependency in ϵ for our algorithm is high, we expect that a more refined analysis of the error could improve this significantly. This is because the approximately feasible to feasible conversion behind Proposition 3.1 requires ϵ^4 accuracy.

Proof. As each run of Algorithm 1 takes at most $\tilde{\mathcal{O}}(1)$ iterations, we only need to implement the oracles in time $\tilde{\mathcal{O}}(n^2 s \epsilon^{-1})$ to establish the advertised runtime for an approximate solution. First, note that the operator norm $\|H_t\|$ only grows modestly with the number of iterations $t = 0, \ldots, T$. This readily follows from $H_0 = 0$, and $\|H_{t+1} - H_t\| \leq \frac{\epsilon}{16}\|P_t\| \leq \frac{\epsilon}{16}$. What is more, the maximal number of steps is $T = \lceil 64 \log(n)/\epsilon^2 \rceil$, implying $\|H_t\| \leq 4\log(n)/\epsilon$ for all t.

In turn, Lemma 3.2 implies that computing the Taylor series of $\exp(-H_t)$ up to a term of order $\mathcal{O}(\log(n)/\epsilon)$ suffices to compute a matrix $\tilde{\rho}_t$ that is $\frac{\epsilon}{4}$-close to the true iterate $\rho_t = \exp(-H_t)/\operatorname{tr}(\exp(-H_t))$ in trace distance. Now note that the complexity of multiplying any matrix with H_t is $\mathcal{O}(\min\{n^2 s, n^\omega\})$, as H_t is a linear combination of a diagonal matrix and A. Thus, we conclude that computing $\tilde{\rho}_t$ takes time $\mathcal{O}(n^2 s \log(n)/\epsilon)$. Checking the diagonal constraints then takes time $\mathcal{O}(n)$ and computing $\operatorname{tr}(A\|A\|^{-1}\tilde{\rho}_t)$ takes time $\mathcal{O}(ns)$. This suffices to implement both ϵ-separation oracles and highlights that the runtime is dominated by computing approximations of the matrix exponential.

Finally, we show in Proposition 3.1 that in order to ensure an additive error of order $\mathcal{O}(\epsilon n\|A\|)$ for the MaxQP SDP, it suffices to solve the relaxed one up to an error ϵ^4, from which the claim follows and we can then convert the approximately feasible solution to a feasible solution in time $\mathcal{O}(n^2)$. □

3.4 Approximately solving the MaxQP SDP on a quantum computer

We will now show how to implement ϵ-separation oracles on a quantum computer. As discussed before, implementing the oracle requires us to evaluate diagonals of the Gibbs states $\rho = \exp(-H)/\operatorname{tr}(\exp(-H))$ and the value of $\operatorname{tr}(\rho A\|A\|^{-1})$. These two tasks can be performed easily on a quantum computer given the ability to prepare approximate copies of the quantum state ρ.

Lemma 3.3. *We can implement ϵ-separation oracles for the MaxQP SDP (3) on a quantum computer given access to $\frac{\epsilon}{8}$ approximate $\mathcal{O}(n\epsilon^{-2})$ copies in trace distance of the input state ρ and the ability to measure $\operatorname{tr}(A\rho)\|A\|^{-1}$. Moreover, the classical postprocessing time needed to implement the oracle is $\mathcal{O}(n\epsilon^{-2})$.*

Proof. Let $\tilde{\rho}$ be the approximation to ρ. We implement the oracle by first measuring $\mathcal{O}(n\epsilon^{-2})$ approximate copies $\tilde{\rho}$ of the input ρ in the computational basis. This is enough to ensure that with probability of failure at most $\mathcal{O}(\mathrm{e}^{-cn})$ the resulting empirical distribution of the measurement outcomes, $\hat{p} = \sum_i \hat{p}(i)|i\rangle\langle i|$, satisfies

$$\|\sum_i \langle i|\tilde{\rho}|i\rangle|i\rangle\langle i| - \hat{p}\|_{tr} \leq \frac{\epsilon}{8}.$$

If $\|I/n - \hat{p}\|_{tr} \leq \frac{3\epsilon}{4}$, then the oracle for the diagonal constraints accepts the current state. If not, we output

$$P = \sum_{i=1}^{n} (\mathbb{I}\{\tilde{p}(i) > 1/n\} - \{\tilde{p}(i) < 1/n\})|i\rangle\langle i|.$$

To see that this indeed satisfies the definition of the oracle, note that the empirical distribution $\hat{p}$ is at most $\frac{\epsilon}{4}$ away in total variation distance to the distribution on the diagonals of ρ. This is because we obtain a $\frac{\epsilon}{8}$ contribution from the approximation $\tilde{\rho}$ and $\frac{\epsilon}{8}$ from statistical noise. Thus, if $\|I/n - \hat{p}\|_{tr} \leq \frac{3\epsilon}{4}$,

$$\|\sum_i \langle i|\rho|i\rangle|i\rangle\langle i| - \frac{I}{n}\|_{tr} \leq \epsilon.$$

by a triangle inequality, as desired. A similar argument shows that we also have

$$\operatorname{tr}\left(P\left(\rho - \frac{I}{n}\right)\right) \geq \frac{\epsilon}{2}$$

whenever $\|I/n - \hat{p}\|_{tr} \geq \frac{3\epsilon}{4}$. Indeed, we have:

$$\operatorname{tr}\left(P\left(\rho - \frac{I}{n}\right)\right) = \operatorname{tr}\left(P\left(\hat{p} - \frac{I}{n}\right)\right) + \operatorname{tr}\left(P\left(\tilde{\rho} - \hat{p}\right)\right) + \operatorname{tr}\left(P\left(\rho - \tilde{\rho}\right)\right). \tag{19}$$

By the definition of P we have:

$$\operatorname{tr}\left(P\left(\hat{p}-\frac{I}{n}\right)\right)=\|I/n-\hat{p}\|_{tr}\geq\frac{3\epsilon}{4} \tag{20}$$

and

$$\operatorname{tr}\left(P\left(\tilde{\rho}-\hat{p}\right)\right)+\operatorname{tr}\left(P\left(\rho-\tilde{\rho}\right)\right)\geq-\frac{\epsilon}{8}-\frac{\epsilon}{8}=-\frac{\epsilon}{4}. \tag{21}$$

This step requires a classical postprocessing time of order $\mathcal{O}(n\epsilon^{-2})$. For implementing the second oracle, we simply measure $A\|A\|^{-1}$ directly. A total of $\mathcal{O}(\epsilon^{-2})$ copies of $\tilde{\rho}$ suffice to determine $\operatorname{tr}\left(A\|A\|^{-1}\rho\right)$ up to precision $\frac{\epsilon}{4}$ via phase estimation [NC00]. □

Lemma 3.3 reduces the task of implementing separation oracles to the task of preparing independent copies of a fixed Gibbs state. There are many different proposals for preparing Gibbs states on quantum computers [CS17, Fra18, KBa16, PW09, TOV+09, TOV+09, YAG12, vAGGdW17]. Here, we will follow the algorithm proposed in [PW09]. This approach allows us to reduce the problem of preparing $\rho_H=\exp(-H)/\operatorname{tr}\left(\exp(-H)\right)$ to the task of simulating the Hamiltonian H. More precisely, [PW09, Appendix] highlights that $\tilde{\mathcal{O}}\left(\sqrt{n}\epsilon^{-3}\right)$ invocations of a controlled U, where U satisfies

$$\|U-e^{it_0H}\|\leq\mathcal{O}(\epsilon^3)\quad\text{where}\quad t_0=\pi/(4\|H\|)$$

suffice to produce a state that is $\frac{\epsilon}{8}$ close in trace distance to ρ_H. The probability of failure is constant. We expect that a more refined analysis can lead to a better dependence on the error ϵ. The methods presented in [vAGGdW17] seem like a good starting point for such future improvements. Here, however, we prioritize the scaling in the problem dimension n only.

By construction, the Hamiltonians we wish to simulate are all of the form $H=aA\|A\|^{-1}+bD$, where $a,b=\mathcal{O}(\log(n)\epsilon^{-1})$ and D is a diagonal matrix with bounded operator norm $\|D\|\leq 1$. It follows from [CW12, Theorem 1] that $\tilde{\mathcal{O}}\left(t(a+b)\exp(1.6\sqrt{\log(\log(n)t\epsilon^{-3})})\right)$ separate simulations of aA and bD suffice to simulate H for time t up to an error ϵ^3. Thus, we further reduce the problem of simulating H to simulating A and D separately.

At this point, it is important to specify input models for the matrix A, the problem description of the MaxQP SDP. We will work in the *sparse oracle input model*. That is, we assume to have access to an oracle O_{sparse} that gives us the position of the nonzero entries. Given indices i for a column of A and a number $1\leq j\leq s$, where A is s-sparse, the oracle acts as:

$$O_{\text{sparse}}\left|i,j\right\rangle=\left|i,f(i,j)\right\rangle.$$

Here $f(i,j)$ is the index of the $j-$th nonzero element of the $i-$th column of A. Moreover, we assume that the magnitude of individual entries are accessible by means of another oracle:

$$O_A\left|i,j,z\right\rangle=\left|i,j,z\oplus\left(A_{ij}\|A\|^{-1}\right)\right\rangle,$$

Here, the entry $\left[A\|A\|^{-1}\right]_{ij}$ is represented by a bit string long enough to ensure the desired precision. The results of [Low19] then highlight that it is possible to simulate $\exp(itA\|A\|^{-1})$ in time $\mathcal{O}\left(\left(t\sqrt{s}\right)^{1+o(1)}\epsilon^{o(1)}\right)$.

Let us now turn to the task of simulating diagonal Hamiltonians D. Let O_D be the matrix entry oracle for D. We suppose that it acts on $\mathbb{C}^n \otimes (\mathbb{C}^2)^{\otimes m}$, where m is large enough to represent the diagonal entries to desired precision in binary, as

$$O_D |i, z\rangle \mapsto |i, z \oplus D_{ii}\rangle . \tag{22}$$

It is then possible to simulate $H = D$ for times $t = \tilde{\mathcal{O}}(\epsilon^{-1})$ with $\tilde{\mathcal{O}}(1)$ queries to the oracle O_D and elementary operations [BACS07]. Thus, efficient simulation of e^{-iDt} follows from an efficient implementation of the oracle O_D. The latter can be achieved with a quantum RAM [GLM08]. We consider the quantum RAM model from [Pra14]. There, it is possible to make insertions in time $\tilde{\mathcal{O}}(1)$. Thus, given a classical description of a diagonal matrix D, we may update the quantum RAM in time $\tilde{\mathcal{O}}(n)$. After we have updated the quantum RAM, we may implement the oracle O_D in time $\tilde{\mathcal{O}}(1)$. Combining all these subroutines establishes the second main result of this work.

Corollary 3.2. *Given an s-sparse, symmetric $n \times n$ matrix A (with appropriate oracle access) and $\epsilon > 0$, we can solve the MaxQP SDP (2) up to an additive error $\epsilon n \|A\|$ in time $\tilde{\mathcal{O}}\left(n^{1.5} (\sqrt{s})^{1+o(1)} \epsilon^{-28+o(1)} \exp(1.6\sqrt{12 \log(\epsilon^{-1})})\right)$ on a quantum computer. The output of the quantum algorithm consists of a real number a and a diagonal matrix D such that for $H = a\frac{A}{\|A\|} + D$ we have that $n\rho_H$ is at trace distance $n\epsilon$ to a feasible point of MaxQP SDP (2).*

Proof. As we saw before, the ability to solve the relaxed MaxQP SDP (3) up to precision $\tilde{\epsilon} = \epsilon^4$ is sufficient to ensure an output with the properties above.

It follows from Theorem 3.3 that producing $\tilde{\mathcal{O}}(n\tilde{\epsilon}^{-2})$ copies of Gibbs states suffices to implement the oracle. The results of [PW09] then imply that each copy can be obtained with with $\tilde{\mathcal{O}}(\sqrt{n}\tilde{\epsilon}^{-3})$ Hamiltonian simulation steps, which, as discussed above, can each be done in time

$$\tilde{\mathcal{O}}\left((\sqrt{s})^{1+o(1)} \tilde{\epsilon}^{o(1)} \exp(1.6\sqrt{\log(\log(n)\tilde{\epsilon}^{-1})})\right) =$$
$$\tilde{\mathcal{O}}\left((\sqrt{s})^{1+o(1)} \tilde{\epsilon}^{o(1)} \exp(1.6\sqrt{\log(\tilde{\epsilon}^{-1})})\right).$$

Thus, the cost per iteration of the algorithm is

$$\tilde{\mathcal{O}}\left(n^{1.5} (\sqrt{s})^{1+o(1)} \tilde{\epsilon}^{-5+o(1)} \exp\left[1.6\sqrt{\log(\epsilon^{-1})}\right]\right).$$

As the algorithm requires $\tilde{\mathcal{O}}(\tilde{\epsilon}^{-2})$ iterations and replacing $\tilde{\epsilon} = \mathcal{O}(\epsilon^{-2})$ we obtain the claim. □

3.5 Randomized rounding

As pioneered by the seminal work of Goemans and Williamson [GW95], it is possible to use randomized rounding techniques to obtain an approximate solution to the original quadratic optimization problem for certain instances (1). These solutions are in expectation within a

multiplicative factor of the value of the SDP relaxation (3) and the exact constant depends on the structure of the matrix A . We will explore Rietz's method, as in [AN06], to show that it is possible to perform the rounding on a classical computer to approximate $\|A\|_{\infty\to 1}$ with our approximately feasible solutions to MaxQP SDP and still obtain good approximations.

First, recall that the rounding algorithms usually work by first multiplying a random Gaussian vector by the square root of the solution. The approximate solution is then given by the signs of this random vector. Note that both classical and quantum algorithms output a classical description of the Hamiltonian $H^\sharp$ associated with an approximately optimal, approximately feasible Gibbs state $\rho^\sharp$ to (3). Pseudocode for the rounding algorithm is provided in Algorithm 2. The first important proof ingredient is an adaptation of [AN06, Eq. (4.1)].

Lemma 3.4. *Fix $v, w \in \mathbb{R}^n$ (non-zero) and let $g \in \mathbb{R}^n$ be a random vector with standard normal entries. Then,*

$$\begin{aligned} &\tfrac{\pi}{2}\mathbb{E}\left[\operatorname{sign}(\langle v|g\rangle)\operatorname{sign}(\langle w|g\rangle)\right] \\ =&\langle \tfrac{v}{\|v\|}, \tfrac{w}{\|w\|}\rangle + \mathbb{E}\left[\left(\langle \tfrac{v}{\|v\|}|g\rangle - \sqrt{\tfrac{\pi}{2}}\operatorname{sign}\left(\langle \tfrac{v}{\|v\|}|g\rangle\right)\right)\left(\langle \tfrac{w}{\|w\|}|g\rangle - \sqrt{\tfrac{\pi}{2}}\operatorname{sign}\left(\langle \tfrac{w}{\|w\|}|g\rangle\right)\right)\right]. \end{aligned} \tag{23}$$

Proof. In [AN06, Eq. (4.1)] the authors use rotation invariance to establish this identity for two unit vectors. The claim then follows from observing that the distribution of $\operatorname{sign}(\langle v|g\rangle)\operatorname{sign}(\langle w|g\rangle)$ is invariant under scaling both v and w by non-negative numbers. In particular, $v \mapsto v/\|v\|$ and $w \mapsto w/\|w\|$ does not affect the distribution. □

The next step involves a technical continuity argument.

Lemma 3.5. *Fix $\epsilon > 0$ and let ρ be a quantum state s.t.:*

$$\|\sum_i \langle i|\rho|i\rangle |i\rangle\langle i| - I/n\|_{tr} \leq \epsilon^4$$

Define the set $B = \{i \in [i] : \left|\rho_{ii} - \frac{1}{n}\right| > \frac{\epsilon^2}{n}\}$ and let $\rho_{\bar{B}}$ be the submatrix with indices in the complement $\bar{B}$ of B. Then, the matrix σ with entries $\sigma_{ij} = \frac{\rho_{ij}}{n\sqrt{\rho_{ii}\rho_{jj}}}$ is a quantum state that obeys $\|\rho_{\bar{B}} - \sigma_{\bar{B}}\|_{tr} \leq 3\epsilon$.

Proof. Note that $\sigma_{\bar{B}} = \mathcal{D}\left(\rho_{\bar{B}}\right)$, where $\mathcal{D}$ is the linear map given by $\mathcal{D}(X) = D_{\bar{B}} X D_{\bar{B}}$ and $D_{\bar{B}}$ is a $|\bar{B}| \times |\bar{B}|$ diagonal matrix with entries $\sqrt{n\rho_{ii}}^{-1}$ for $i \in \bar{B}$. This implies

$$\|\rho_{\bar{B}} - \sigma_{\bar{B}}\|_{tr} = \|\left(\mathrm{id} - \mathcal{D}\right)\left(\rho_{\bar{B}}\right)\|_{tr} \leq \|\mathrm{id} - \mathcal{D}\|_{tr\to tr}\|\rho_{\bar{B}}\|_{tr} \leq \|\mathrm{id} - \mathcal{D}\|_{\infty\to\infty},$$

because $\|\rho_{\bar{B}}\|_{tr} \leq \|\rho\|_{tr} = \mathrm{tr}(\rho) = 1$. Duality of norms and the fact that both id and $\mathcal{D}$ are self-adjoint with respect of the Frobenius inner product $\mathrm{tr}\left(X^T Y\right)$ implies $\|\mathrm{id} - \mathcal{D}\|_{\infty\to\infty} = \|\mathrm{id} - \mathcal{D}\|_{tr\to tr}$. This allows us to bound $\|\mathrm{id} - \mathcal{D}\|_{\infty\to\infty}$ instead. By construction, we have that all the entries of $D_{\bar{B}}$ are in $1 \pm \epsilon$. Write $D_{\bar{B}} = I + D_\epsilon$, where D_ϵ is a diagonal matrix with entries that are bounded by ϵ in absolute value. Then,

$$\mathrm{id} - \mathcal{D}(X) = D_\epsilon X + X D_\epsilon + D_\epsilon X D_\epsilon \quad \text{for any matrix} \quad X.$$

Submultiplicativity of the operator norm then implies

$$\|D_\epsilon X + X D_\epsilon + D_\epsilon X D_\epsilon\|_\infty \leq 2\|D_\epsilon\|_\infty \|X\| + \|D_\epsilon\|_\infty^2 \|X\|_\infty \leq 3\epsilon \|X\|_\infty.$$

and, in turn, $\|\mathrm{id} - \mathcal{D}\|_{\infty\to\infty} \leq 3\epsilon$. □

We are now ready to prove the main stability result required for randomized rounding.

Theorem 3.1. *Let $\rho^\sharp$ be an approximately feasible, optimal point of (3) with accuracy $\epsilon^4 > 0$ and input matrix A' with*

$$A' = \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix},$$

where A is a real $n \times n$ matrix. Let $v_1, \ldots, v_{2n}$ be the columns of $\sqrt{\rho^\sharp}$, sample $g \in \mathbb{R}^{2n}$ with i.i.d. Gaussian entries and set $x_i = \mathrm{sign}(\langle v_i | g \rangle)$ and $y = (x_1, \ldots, x_n), z = (x_{n+1}, \ldots, x_{2n})$. Then,

$$\mathrm{tr}\left(\rho^\sharp A\right) n + \mathcal{O}(\epsilon n \|A\|) \geq \sum_{i,j} A_{ij} \mathbb{E}(y_i z_j) \geq (4/\pi - 1)\,\mathrm{tr}\left(\rho^\sharp A\right) n - \mathcal{O}(\epsilon n \|A\|).$$

Proof. The upper bound follows immediately from the fact MaxQP SDP (2) relaxations (renormalized or not) provide upper bounds to the original problem (1). The factors $n\|A\|$ is an artifact of the renormalization (3).

For the lower bound, we once more define $B = \{i \in [i] : |\rho_{ii} - 1/2n| \geq \epsilon^2/2n\} \subset [2n]$. Plugging in v_i and v_j in (23), multiplying both sides by A'_{ij} and summing over i, j implies

$$\frac{\pi}{2} \sum_{i,j} A'_{ij} \mathbb{E}(x_i x_j) = 2n \sum_{i,j} A'_{ij} (\sigma_{ij} + \tau_{ij}) \quad \text{with} \quad \sigma_{ij} = \tfrac{\rho_{ij}}{2n\sqrt{\rho_{ii}\rho_{jj}}} \quad \text{and}$$

$$\tau_{ij} = \mathbb{E}\left[\left(\langle \tfrac{v_i}{\|v_i\|} | g \rangle - \sqrt{\tfrac{\pi}{2}}\,\mathrm{sign}\left(\langle \tfrac{v_i}{\|v_i\|} | g \rangle\right)\right)\left(\langle \tfrac{v_j}{\|v_j\|} | g \rangle - \sqrt{\tfrac{\pi}{2}}\,\mathrm{sign}\left(\langle \tfrac{v_j}{\|v_j\|} | g \rangle\right)\right)\right].$$

Following the same proof strategy as in [AN06, Sec. 4.1], we note that the matrix T defined by $[T]_{ij} = \tau_{ij}$ is a Gram matrix and, thus, psd. Moreover, in [AN06, Sec. 4.1] the author shows that $\tau_{ii} = \frac{\pi}{2} - 1$. These two properties imply that $\left(\frac{\pi}{2} - 1\right)^{-1} (2n)^{-1} T$ is a feasible point of (3). Moreover, because of the structure of the matrix A', we have that

$$|\,\mathrm{tr}\,(TA')\,| \leq \left(\frac{\pi}{2} - 1\right) \mathrm{tr}\left(\rho^\sharp A'\right) n - \mathcal{O}(\epsilon n \|A\|) \tag{24}$$

To see this, consider the block unitary

$$U = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}.$$

Then for any psd matrix X we have that $\mathrm{tr}\left(A' U X U^\dagger\right) = -\,\mathrm{tr}\,(A'X)$ and so $\mathrm{tr}\left(A' U \rho^\sharp U^\dagger\right)$ provides a lower bound to the value over the approximately feasible set . Thus,

$$\frac{\pi}{2} \sum_{i,j} A'_{ij} \mathbb{E}(y_i z_j) = 2n \sum_{i,j} A'_{ij} (\sigma_{ij} + \tau_{ij}) \geq$$

$$2n \sum_{i,j} A'_{ij} \sigma_{ij} - \left(\frac{\pi}{2} - 1\right) \mathrm{tr}\left(\rho^\sharp A'\right) n - \mathcal{O}(\epsilon n \|A\|)$$

We now have to relate $\operatorname{tr}\left(\rho^\sharp A'\right)$ to $\operatorname{tr}(\sigma A')$. To do so, we can argue like in Proposition 3.1 and see that $\operatorname{tr}(\sigma_{11}), \operatorname{tr}(\rho_{11}) = \mathcal{O}(\epsilon^2)$ (these correspond to the $|B| \times |B|$ psd submatrices with entries in B only). As both σ and ρ are states, we conclude

$$\|\rho_{12}\|_{tr}, \|\sigma_{12}\|_{tr} = \mathcal{O}(\epsilon)$$

by reusing the analysis provided in the proof of Proposition 3.1. Thus, it follows from Hölders inequality and Lemma 3.5 that

$$\begin{aligned}\operatorname{tr}(A'(\rho - \sigma)) =& \operatorname{tr}(A'(\rho_{22} - \sigma_{22})) + \operatorname{tr}(A'(\rho_{11} - 2\rho_{12} - \sigma_{11} - 2\sigma_{12})) \\ =& \|A\| \left(\|\sigma_{22} - \rho_{22}\|_{tr} + \|\rho_{11}\|_{tr} + 2\|\rho_{12}\|_{tr} + 2\|\sigma_{11}\|_{tr} + 2\|\sigma_{12}\|_{tr}\right) = \mathcal{O}(\|A\|\epsilon),\end{aligned}$$

from which the claim follows. □

Proposition 3.1 highlights that performing the rounding with approximate solutions to the MaxQP SDP (3) still ensures a good approximate solution in expectation for the $\|A\|_{\infty\to 1}$ norm. In the case of matrices A that are psd it is possible to improve the constant in the rounding and we do not to resort to lifting the problem to a matrix with double the dimension:

Corollary 3.3. *Let $\rho^\sharp$ be an approximately feasible, optimal point of (3) with accuracy $\epsilon^4 > 0$ and psd input matrix A. Let $v_1, \ldots, v_n$ be the columns of $\sqrt{\rho^\sharp}$, sample $g \in \mathbb{R}^n$ with i.i.d. Gaussian entries and set $x_i = \operatorname{sign}(\langle v_i | g \rangle)$. Then,*

$$\operatorname{tr}\left(\rho^\sharp A\right) n + \mathcal{O}(\epsilon n \|A\|) \geq \sum_{i,j} A_{ij} \mathbb{E}(x_i x_j) \geq (2/\pi) \operatorname{tr}\left(\rho^\sharp A\right) n - \mathcal{O}(\epsilon n \|A\|).$$

Proof. The proof follows by following the same proof as above but noting that we may use the estimate $\operatorname{tr}(TA) \geq 0$ instead of (24), as both A and T are psd. Optimality of the constant was shown in [AN06]. □

As Alon and Naor [AN06] also show that for psd matrices A we have

$$\|A\|_{\infty\to 1} = \max_{x \in \{\pm 1\}^n} \langle x | A | x \rangle,$$

i.e. we may restrict to the same vector on the left and right, it follows that Corollary 3.3 gives almost optimal rounding guarantees. These two statements certify that, as longs as $\|A\|_{\infty\to 1} = \Theta(n\|A\|)$, performing the rounding with our approximately feasible solutions gives rise to approximations of the $\infty \to 1$ norm that are almost as good the strictly feasible solutions.

But computing $\sqrt{\rho^\sharp} g = \exp(-H/2) g / \sqrt{\operatorname{tr}(\exp(-H))}$ directly still remains expensive because of matrix exponentiation. We will surpass this bottleneck by truncating the Taylor series of the matrix exponential in a fashion similar to Lemma 3.2. The following standard anti-concentration result for Gaussian random variables will be essential for this argument.

Fact 3.1. *Let X be a $\mathcal{N}(0, \sigma^2)$ random variable. Then $\mathbb{P}(|X| \leq \sigma\epsilon) = \mathcal{O}(\epsilon)$.*

Lemma 3.6. *Let $\rho^\sharp$ with associated Hamiltonian $H^\sharp$ be an approximately optimal solution to the MaxQP SDP (3) with $\|H^\sharp\| = \mathcal{O}(\log(n)/\epsilon)$. Set $S_l = \sum_{k=0}^{l} \frac{1}{k!}(-H^\sharp/2)^k$ with $l = \mathcal{O}(\log(n)/\epsilon)$. Then, a random vector $g \in \mathbb{R}^n$ with standard normal entries obeys*

$$\operatorname{sign}\left[\left(e^{H^\sharp/2}g\right)_i\right] = \operatorname{sign}\left[(S_l g)_i\right] \quad \textit{for all} \quad i \in [n] \quad \textit{such that} \quad \left|\rho^\sharp_{ii} - \tfrac{1}{n}\right| < \tfrac{\epsilon}{n}$$

with probability at least $1 - \mathcal{O}(\epsilon^{-1})$.

Note that the design of Algorithm 1 ensures that optimal Hamiltonians always obey $\|H^\sharp\| = \mathcal{O}(\log(n)/\epsilon)$.

Proof. Define $h = \exp(-H^\sharp/2)g$ and note that this is a Gaussian random vector with covariance matrix $\exp(-H^\sharp)$. Let $B = \{i : |\rho_{ii} - 1/n| > \frac{\epsilon}{n}\} \subset [n]$ denote the set of indices for which ρ_{ii} deviates substantially from $1/n$. Then, every entry of h that is not contained in this index set obeys

$$[h]_i = [\exp(-H/2)g]_i \sim \mathcal{N}\left(0, \tfrac{c}{n}\operatorname{tr}(\exp(-H))\right) \quad \text{with} \quad c \in (1-\epsilon, 1+\epsilon).$$

The assumption $\|H^\sharp\| = \mathcal{O}(\log(n)/\epsilon)$ ensures $\operatorname{tr}(\exp(-H^\sharp))/n \geq n^{-c'/\epsilon - 1}$ for some constant c'. We can combine this with Fact 3.1 (Gaussian anti-concentration) to conclude

$$\mathbb{P}\left[|\,[h]_i\,| \leq n^{-2-c'/(2\epsilon)}\right] = \mathcal{O}(1/n^2) \quad \text{for all} \quad i \in \bar{B} = [n] \setminus B.$$

A union bound then asserts

$$\mathbb{P}\left[\exists i \in \bar{B} : \, |\,[h]_i\,| \leq n^{-2-c'/\epsilon}\right] = \mathcal{O}(1/n).$$

Moreover, it follows from standard concentration arguments that

$$\mathbb{P}\left[n - n^{\frac{1}{4}} \leq \|g\|^2 \leq n + n^{\frac{1}{4}}\right] \geq 1 - 2e^{-\sqrt{n}/8}.$$

Thus, with probability at least $1 - \mathcal{O}(n^{-1})$, we have that $\|g\|^2 \leq n + n^{\frac{1}{4}}$ and $|\,[h]_i\,| \geq n^{-2-c'/\epsilon}$ for every entry $i \in \bar{B}$. Following the same proof strategy as in Lemma 3.2, it is easy to see that picking $l = \mathcal{O}(\epsilon^{-1}\log(n))$ suffices to ensure that

$$\|S_l - \exp(-H/2)\| \leq n^{-4-\frac{c'}{2\epsilon}}$$

Conditioning on the events emphasized above, implies

$$\max_{i\in[n]} |[(\exp(-H/2) - S_l)\, g]_i| \leq \|\,(\exp(-H/2) - S_k)\, g\| \leq \|\exp(-H/2) - S_k\|\|g\| \leq n^{-4-\frac{c'}{2\epsilon}}\|g\|.$$

This in turn ensures $\max_{i\in\bar{B}} |[(\exp(-H/2) - S_l)\, g]_i| \leq n^{-3-\frac{c'}{2\epsilon}}$, which then gives

$$\operatorname{sign}\left([h]_i\right) = \operatorname{sign}\left([(\exp(-H/2)g)]_i\right) = \operatorname{sign}\left([S_k g]_i\right) \quad \text{for all} \quad i \in \bar{B},$$

because conditioning ensures $|[\exp(-H/2)g]_i| \geq n^{-2-\frac{c'}{2\epsilon}}$. □

Combining the statements we just proved we conclude that:

Proposition 3.2 (Restatement of Proposition 2.1)**.** *Let $\epsilon > 0$ and A a real, psd matrix be given. Moreover, let H be the solution Hamiltonian to the relaxed MaxQP SDP (3) with error parameter ϵ^4 and α^* its value. Then, with probability at least $1 - n^{-1}$, the output x of Algorithm 2 satisfies:*

$$n\|A\|(\alpha^* + \mathcal{O}(\epsilon)) \geq \mathbb{E}[\sum_{ij} A_{ij}x_i x_j] \geq \frac{2}{\pi} n\|A\|(\alpha^* - \mathcal{O}(\epsilon)), \tag{25}$$

Proof. It follows from Lemma 3.6 that the output of Algorithm 2 will only differ from the vector obtained by performing the rounding with the approximate solution on a set of size $\mathcal{O}(n\epsilon^2)$ with probability at least $1 - n^{-1}$. This is because, as argued before, by picking ϵ^4 we have at most $\mathcal{O}(\epsilon^2 n)$ diagonal entries that do not satisfy $|\rho_{ii} - 1/n| \leq \epsilon/n$. We will now argue that sign vectors that differ at $\mathcal{O}(n\epsilon^2)$ position can differ in value by at most $\mathcal{O}(\epsilon n\|A\|)$. Let x be the vector obtained by the ideal rounding and x' the one with the truncated Taylor series. Then there exists a vector e with at most $\mathcal{O}(n\epsilon^2)$ nonzero entries bounded by 2 such that $x = x + e$ by our assumption. By Cauchy-Schwarz:

$$|\langle x|A|x\rangle - \langle x'|A|x'\rangle| \leq |\langle e|A|x\rangle| + |\langle x|A|e\rangle| + |\langle e|A|e\rangle| \leq \|A\| \left(2\|x\|\|e\| + \|e\|^2\right).$$

Now, as x is a binary vector, $\|x\| = \sqrt{n}$ and, as e has at most $\mathcal{O}(\epsilon^2 n)$ nonzero entries, it follows that $\|e\| = \mathcal{O}(\epsilon\sqrt{n})$ and we conclude

$$|\langle x|A|x\rangle - \langle x'|A|x'\rangle| = \mathcal{O}(\epsilon n\|A\|)$$

As Theorem 3.1 asserts that performing the rounding with the approximate solution is enough to produce a sign vector that satisfies (25) in expectation, this yields the claim. □

The analogous claim, i.e. that truncating still gives rise to good solutions, clearly also holds in the setting of Proposition 3.1.

Thus, we conclude that the rounding can be performed in time $\tilde{\mathcal{O}}(ns)$ on a classical computer, as multiplying a vector with H takes time $\tilde{\mathcal{O}}(ns)$ and we only need to perform these operations for a total number of steps that is logarithmic in the problem dimension n (but polynomial in inverse accuracy $1/\epsilon$). As $ns \leq n^{1.5}\sqrt{s}$ for $s \leq n$, we conclude that the cost of solving the relaxed MaxQP SDP (3) dominates the cost of rounding.

4 Conclusion and Outlook

By adapting ideas from [TRW05, Haz16, LRS15, BKL+17], we have provided a general meta-algorithm for approximately solving convex feasibility problems with psd constraints. *Hamiltonian Updates* is an iterative procedure based on a simple change of variables: represent a trace-normalized, positive semidefinite matrix as $X = \exp(-H)/\operatorname{tr}(\exp(-H))$. At each step, infeasible directions are penalized in the matrix exponent until an approximately feasible point is reached. This procedure can be equipped with rigorous convergence guarantees and lends itself to quantum improvements: $X = \exp(-H)\operatorname{tr}(\exp(-H))$ is a *Gibbs state* and

H is the associated *Hamiltonian*. Quantum architectures can produce certain Gibbs states very efficiently.

We have demonstrated the viability of this approach by considering semidefinite programming relaxations of quadratic problems with binary constraints (MaxQP SDP) (2). The motivation for considering this practically important problem class was two-fold: (i) MaxQP SDPs have received considerable attention in the (classical) computer science community. Powerful meta-algorithms, like matrix multiplicative weights [AK16], have been designed to solve these SDPs very quickly. (ii) So far, quantum SDP solvers [BS17, vAGGdW17, vAG19, BKL+17, KP20] have failed to provide speedups for MaxQP SDPs. The quantum runtime associated with these solvers depends on problem-specific parameters that scale particularly poorly for MaxQP SDPs. Moreover, the notions of approximate feasibility championed in these other works are too loose for this class of problem.

The framework developed in this paper has allowed us to address these points. Firstly, we shown that a classical implementation of Hamiltonian Updates already improves upon the best existing results. A runtime of $\tilde{\mathcal{O}}(n^2 s)$ suffices to find an approximately optimal solution. Secondly, we have showed that quantum computers do offer additional speedups. A quantum runtime of $\tilde{\mathcal{O}}(n^{1.5} s^{0.5+o(1)})$ is sufficient. We emphasize that this is the first quantum speedup for MaxQP SDP relaxations. Subsequently, we have devised a classical randomized rounding procedure that converts both quantum and classical solutions into close to optimal solutions of the original quadratic problem.

We note in passing that our algorithm is very robust, in the sense that it only requires the preparation of Gibbs states up to a precision ϵ that can be taken to be constant in the number of qubits. This requirement is combined with other simple tasks like computational basis measurements and the ability to estimate the expectation value of the target matrix on states. Although the subroutines used in this work to perform these tasks certainly require nontrivial quantum circuits, it would be interesting to identify classes of target matrices A for which preparing the corresponding Gibbs state and estimating the expectation values is feasible on near-term devices.

We believe that the framework presented here lends itself to further applications.

One concrete application of Hamiltonian Updates, in particular the idea to treat constraints as the statistics of measurements, is *quantum state tomography*, see e.g. [BCG13] and references therein. Sample-optimal tomography protocols have revealed that classical postprocessing is the main bottleneck for reconstructing density matrices [FGLE12, OW16, KRT17, HHJ+17, GKKT20]. A classical implementation of Hamiltonian Updates allows for optimizing postprocessing costs at the expense of a worse dependence on accuracy [BKF20]. Further improvements are possible by executing the algorithm on a quantum computer, giving a quantum speedup for quantum state tomography.

Another promising and practically relevant application is *binary matrix factorization*. A recent line of works [KT19a, KT19b] reduces this problem to a sequence of SDPs. Importantly, each SDP corresponds to a MAXQP SDP (2) with a random rank-one objective $A = |a\rangle\langle a|$ and an additional affine constraint $\mathrm{tr}\,(PX) = n$. Here, P is a fixed low-rank orthoprojector. This application, however, is likely going to be more demanding in terms of approximation accuracy. Hence, improving the runtime scaling in inverse accuracy will constitute an important first step that is of independent interest.

5 Acknowledgments

We would like to thank Aram Harrow for inspiring discussions. Our gratitude extends, in particular, to Ronald de Wolf, András Gilyén and Joran van Apeldoorn who provided valuable feedback regarding an earlier version of this draft. Finally, we would like to thank the anonymous reviewers for in-depth comments and suggestions. D.S.F. would like to thank the hospitality of Caltech's Institute for Quantum Information and Matter, where the main ideas in this paper were conceived during a visit. F.G.L.S.B and R.K. acknowledge funding provided by the Institute for Quantum Information and Matter, an NSF Physics Frontiers Center (NSF Grant PHY-1733907), as well as financial support from Samsung. R.K's work is also supported by the Office of Naval Research (Award N00014-17-1-2146) and the Army Research Office (Award W911NF121054). D.S.F. acknowledges financial support from VILLUM FONDEN via the QMATH Centre of Excellence (Grant no. 10059), the graduate program TopMath of the Elite Network of Bavaria, the TopMath Graduate Center of TUM Graduate School at Technische Universität München and by the Technische Universität München – Institute for Advanced Study, funded by the German Excellence Initiative and the European Union Seventh Framework Programme under grant agreement no. 291763.

References

[ACH+19] S. Aaronson, X. Chen, E. Hazan, S. Kale, and A. Nayak. Online learning of quantum states. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124019, 2019.

[Ad20] S. Apers and R. de Wolf. Quantum speedup for graph sparsification, cut approximation and laplacian solving. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 637–648, 2020.

[AFdlVKK03] N. Alon, W. Fernandez de la Vega, R. Kannan, and M. Karpinski. Random sampling and approximation of MAX-CSPs. volume 67, pages 212–243. 2003.

[AHK05] S. Arora, E. Hazan, and S. Kale. Fast algorithms for approximate semidefinite programming using the multiplicative weights update method. In *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)*, pages 339–348, 2005.

[AK16] S. Arora and S. Kale. A combinatorial, primal-dual approach to semidefinite programs. *J. ACM*, 63(2):Art. 12, 35, 2016.

[AL70] H. Araki and E. H. Lieb. Entropy inequalities. *Comm. Math. Phys.*, 18:160–170, 1970.

[AN06] N. Alon and A. Naor. Approximating the cut-norm via Grothendieck's inequality. *SIAM J. Comput.*, 35(4):787–803, 2006.

[BACS07] D. W. Berry, G. Ahokas, R. Cleve, and B. C. Sanders. Efficient quantum algorithms for simulating sparse Hamiltonians. *Comm. Math. Phys.*, 270(2):359–371, 2007.

[BCG13] K. Banaszek, M. Cramer, and D. Gross. Focus on quantum tomography. *New Journal of Physics*, 15(12):125020, 2013.

[Bha97] R. Bhatia. *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997.

[BKF20] F. G. Brandão, R. Kueng, and D. S. França. Fast and robust quantum state tomography from few basis measurements. *arXiv preprint arXiv:2009.08216*, 2020.

[BKL+17] F. G. Brandão, A. Kalev, T. Li, C. Y.-Y. Lin, K. M. Svore, and X. Wu. Exponential quantum speed-ups for semidefinite programming with applications to quantum learning. *arXiv preprint arXiv:1710.02581*, 2017.

[BM05] S. Burer and R. D. C. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Math. Program.*, 103(3, Ser. A):427–444, 2005.

[BS17] F. G. S. L. Brandao and K. M. Svore. Quantum speed-ups for solving semidefinite programs. In *58th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2017*, pages 415–426. IEEE Computer Soc., Los Alamitos, CA, 2017.

[Bub15] S. Bubeck. Convex Optimization: Algorithms and Complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[BV04] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004.

[BVB16] N. Boumal, V. Voroninski, and A. Bandeira. The non-convex Burer-Monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems 29*, pages 2757–2765. Curran Associates, Inc., 2016.

[CS17] A. N. Chowdhury and R. D. Somma. Quantum algorithms for Gibbs sampling and hitting-time estimation. *Quantum Inf. Comput.*, 17(1-2):41–64, 2017.

[CW04] M. Charikar and A. Wirth. Maximizing Quadratic Programs: Extending Grothendieck's Inequality. In *45th Annual IEEE Symposium on Foundations of Computer Science*, pages 54–60. IEEE, 2004.

[CW12] A. M. Childs and N. Wiebe. Hamiltonian simulation using linear combinations of unitary operations. *Quantum Inf. Comput.*, 12(11-12):901–924, 2012.

[DMS17] A. Dembo, A. Montanari, and S. Sen. Extremal Cuts of Sparse Random Graphs. *The Annals of Probability*, 45(2):1190–1217, March 2017. arXiv: 1503.03923.

[FGLE12] S. T. Flammia, D. Gross, Y.-K. Liu, and J. Eisert. Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators. *New Journal of Physics*, 14(9):095022, 2012.

[FK99] A. Frieze and R. Kannan. Quick approximation to matrices and applications. *Combinatorica*, 19(2):175–220, 1999.

[Fra18] D. S. França. Perfect sampling for quantum Gibbs states. *Quantum Inf. Comput.*, 18(5-6):361–388, 2018.

[Git13] A. Gittens. *Topics in Randomized Numerical Linear Algebra*. ProQuest LLC, Ann Arbor, MI, 2013. Thesis (Ph.D.)–California Institute of Technology.

[GKKT20] M. Guţă, J. Kahn, R. Kueng, and J. A. Tropp. Fast state tomography with optimal error bounds. *Journal of Physics A: Mathematical and Theoretical*, 53(20):204001, May 2020.

[GLM08] V. Giovannetti, S. Lloyd, and L. Maccone. Quantum random access memory. *Phys. Rev. Lett.*, 100(16):160501, 4, 2008.

[GW95] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. Assoc. Comput. Mach.*, 42(6):1115–1145, 1995.

[Haz16] E. Hazan. *Introduction to Online Convex Optimization*. now Publishers Inc, 2016.

[HHJ+17] J. Haah, A. W. Harrow, Z. Ji, X. Wu, and N. Yu. Sample-optimal tomography of quantum states. *IEEE Transactions on Information Theory*, 63(9):5628–5641, 2017.

[JMRT16] A. Javanmard, A. Montanari, and F. Ricci-Tersenghi. Phase transitions in semidefinite relaxations. *Proceedings of the National Academy of Sciences*, 113(16):E2218–E2223, mar 2016.

[KB20] D. Kunisky and A. S. Bandeira. A tight degree 4 sum-of-squares lower bound for the Sherrington–Kirkpatrick Hamiltonian. *Mathematical Programming*, November 2020.

[KBa16] M. J. Kastoryano and F. G. S. L. Brandão. Quantum Gibbs samplers: the commuting case. *Comm. Math. Phys.*, 344(3):915–957, 2016.

[Kin03] C. King. Inequalities for trace norms of 2×2 block matrices. *Comm. Math. Phys.*, 242(3):531–545, 2003.

[KLP+16] R. Kyng, Y. T. Lee, R. Peng, S. Sachdeva, and D. A. Spielman. Sparsified Cholesky and multigrid solvers for connection Laplacians. In *STOC'16—Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, pages 842–850. ACM, New York, 2016.

[Kol98] T. G. Kolda. *Limited-memory matrix methods with applications*. PhD thesis, University of Michigan, 1998.

[KP20] I. Kerenidis and A. Prakash. A Quantum Interior Point Method for LPs and SDPs. *ACM Transactions on Quantum Computing*, 1(1):1–32, December 2020.

[KRT17] R. Kueng, H. Rauhut, and U. Terstiege. Low rank matrix recovery from rank one measurements. *Appl. Comput. Harmon. Anal.*, 42(1):88–116, 2017.

[KT19a] R. Kueng and J. A. Tropp. Binary component decomposition Part I: the positive-semidefinite case. *arXiv preprint arXiv:1907.13603*, 2019.

[KT19b] R. Kueng and J. A. Tropp. Binary component decomposition Part II: the asymmetric case. *arXiv preprint arXiv:1907.13602*, 2019.

[KY13] A. Knowles and J. Yin. The isotropic semicircle law and deformation of Wigner matrices. *Communications on Pure and Applied Mathematics*, 66(11):1663–1749, apr 2013.

[Lat05] R. Latala. Some estimates of norms of random matrices. *Proc. Amer. Math. Soc.*, 133(5):1273–1282, 2005.

[Low19] G. H. Low. Hamiltonian simulation with nearly optimal dependence on spectral norm. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing - STOC 2019*, pages 491–502, New York, New York, USA, 2019. ACM Press.

[LRS15] J. R. Lee, P. Raghavendra, and D. Steurer. Lower bounds on the size of semidefinite programming relaxations. In *STOC'15—Proceedings of the 2015*

ACM Symposium on Theory of Computing, pages 567–576. ACM, New York, 2015.

[LSW15] Y. T. Lee, A. Sidford, and S. C.-W. Wong. A faster cutting plane method and its implications for combinatorial and convex optimization. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science—FOCS 2015*, pages 1049–1065. IEEE Computer Soc., Los Alamitos, CA, 2015.

[MMMO17] S. Mei, T. Misiakiewicz, A. Montanari, and R. I. Oliveira. Solving sdps for synchronization and maxcut problems via the Grothendieck inequality. *arXiv preprint arXiv:1703.08729*, 2017.

[Mon19] A. Montanari. Optimization of the Sherrington-Kirkpatrick hamiltonian. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1417–1433, 2019.

[MS16] A. Montanari and S. Sen. Semidefinite programs on sparse random graphs and their application to community detection. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 814–827, Cambridge MA USA, June 2016. ACM.

[NC00] M. A. Nielsen and I. L. Chuang. *Quantum computation and quantum information*. Cambridge University Press, Cambridge, 2000.

[Nik09] V. Nikiforov. Cut-norms and spectra of matrices. *arXiv preprint arXiv:0912.0336*, 2009.

[OP83] D. O'Leary and S. Peleg. Digital image compression by outer product expansion. *IEEE Transactions on Communications*, 31(3):441–444, 1983.

[OW16] R. O'Donnell and J. Wright. Efficient quantum tomography. In *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*, STOC '16, pages 899–912, New York, NY, USA, 2016. ACM.

[Pan13] D. Panchenko. *The Sherrington-Kirkpatrick model*. Springer Monographs in Mathematics. Springer, New York, 2013.

[Pra14] A. Prakash. *Quantum algorithms for linear algebra and machine learning*. PhD thesis, University of California, Berkeley, 2014.

[PW09] D. Poulin and P. Wocjan. Sampling from the thermal quantum Gibbs state and evaluating partition functions with a quantum computer. *Phys. Rev. Lett.*, 103(22):220502, 4, 2009.

[RV18] E. Rebrova and R. Vershynin. Norms of random matrices: local and global problems. *Adv. Math.*, 324:40–83, 2018.

[Sio58] M. Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, March 1958.

[ST11] D. A. Spielman and S.-H. Teng. Spectral sparsification of graphs. *SIAM J. Comput.*, 40(4):981–1025, 2011.

[Tal11] M. Talagrand. The Diluted SK Model and the K-Sat Problem. In *Mean Field Models for Spin Glasses*, pages 325–395. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[TOV+09] K. Temme, T. J. Osborne, K. G. Vollbrecht, D. Poulin, and F. Verstraete. Quantum metropolis sampling. *Nature*, 471:87,2011, 2009.

[TRW05] K. Tsuda, G. Rätsch, and M. K. Warmuth. Matrix exponentiated gradient

updates for on-line learning and Bregman projection. *J. Mach. Learn. Res.*, 6:995–1018, 2005.

[TYUC17] J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher. Practical sketching algorithms for low-rank matrix approximation. *SIAM J. Matrix Anal. Appl.*, 38(4):1454–1485, 2017.

[vAG19] J. van Apeldoorn and A. Gilyén. Improvements in Quantum SDP-Solving with Applications. In C. Baier, I. Chatzigiannakis, P. Flocchini, and S. Leonardi, editors, *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*, volume 132 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 99:1–99:15, Dagstuhl, Germany, 2019. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[vAGGdW17] J. van Apeldoorn, A. Gilyén, S. Gribling, and R. de Wolf. Quantum SDP-solvers: better upper and lower bounds. In *58th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2017*, pages 403–414. IEEE Computer Soc., Los Alamitos, CA, 2017.

[YAG12] M.-H. Yung and A. Aspuru-Guzik. A quantum–quantum metropolis algorithm. *Proceedings of the National Academy of Sciences*, 109(3):754–759, 2012.

A Norms of random matrices

There is an interesting discrepancy in the error scaling between the methods presented here and existing ones by Arora et al. [AHK05]: $\|A\|_{\ell_1}$ (existing work) vs $n\|A\|$ (here). The following fundamental relations relate these norms [Nik09]:

$$\|A\|_{\infty\to 1} \leq n\|A\|, \quad \|A\|_{\infty\to 1} \leq \|A\|_{\ell_1}, \quad \|A\| \leq \|A\|_{\ell_1} \leq n\sqrt{\operatorname{rank}(A)}\|A\|.$$

All inequalities are tight up to constants. The above inequalities highlight that it is a priori not clear what the correct scaling for errors approximating the cut norm should be. The goal of this section will be to show that for random matrices A with independent, standardized entries that have bounded fourth moment $n\|A\|$ reproduces the correct error behavior, while $\|A\|_{\ell_1}$ does not.

Proposition A.1 (Cut norm of random matrices)**.** *Let A be a $n \times n$ random matrix whose entries are sampled independently from a real-valued distribution α that obeys $\mathbb{E}[\alpha] = 0$, $\mathbb{E}[\alpha^2] = 1$ and $\mathbb{E}[\alpha^4] = \mathcal{O}(1)$. Then,*

$$\mathbb{E}[\|A\|_{\ell_1}] = \Theta(n^2), \quad \mathbb{E}[\|A\|_{\infty\to 1}] = \Theta(n^{1.5}), \quad \mathbb{E}[\|A\|] = \mathcal{O}\left(\sqrt{n}\right).$$

Proof. We refer to Latala's work for the third claim [Lat05]. A key ingredient for establishing the second claim is [Git13, Corollary 3.10]:

$$\tfrac{1}{\sqrt{2}}\mathbb{E}\left(\|A\|_{\text{col}}\right) \leq \mathbb{E}\left(\|A\|_{\infty\to 1}\right) \leq 4\mathbb{E}\left(\|A\|_{\text{col}}\right),$$

where $\|A\|_{\text{col}} = \sum_i \sqrt{\sum_j [A]_{ij}^2}$ is the sum of the Euclidean norms of the columns of A. Now, note that the entries of A are i.i.d. copies of the random variable α. In turn, the expected

column norm of A is just n times the expected Euclidean norm of the random vector $a = (a_1, \ldots, a_n)^T$, where each a_i is an independent copy of α. Jensen's inequality then asserts

$$\mathbb{E}\left[\|a\|_2\right] \leq \left(\mathbb{E}\left[\sum_{i=1}^{n} a_i^2\right]\right)^{1/2} = \sqrt{n\mathbb{E}\left[\alpha^2\right]} = \sqrt{n},$$

while a matching lower bound follows from $\sqrt{x} \geq \frac{1}{2}(1 + x - (x-1)^2)$. Indeed, define $y = \|a\|_2^2/n = \frac{1}{n}\sum_{i=1}^{n} a_i^2$ and note that this new random variable obeys $\mathbb{E}[y] = 1$ and $\mathbb{E}[(y-1)^2] = \mathcal{O}(1/n)$ by assumption. This ensures a matching lower bound:

$$\mathbb{E}\left[\|a\|_2\right] = \sqrt{n}\mathbb{E}\left[\sqrt{y}\right] \geq \frac{\sqrt{n}}{2}\left(1 + \mathbb{E}\left[y\right] - \mathbb{E}\left[(y-1)^2\right]\right) = \Omega(\sqrt{n}),$$

This ensures $\mathbb{E}\left[\|A\|_{\infty\to 1}\right] = n\mathbb{E}\left[\|a\|_2\right] = \Theta(n^{3/2})$ and establishes the second claim.

The first claim follows from the fact that the fourth-moment bound $\mathbb{E}\left[\alpha^4\right] = \mathcal{O}(1)$ demands $\mathbb{E}\left[|\alpha|\right] = \Theta(1)$. Combine this with i.i.d. entries of the random matrix A to conclude

$$\mathbb{E}\left[\|A\|_{\ell_1}\right] = n^2\mathbb{E}\left[|\alpha|\right] = \Theta(n^2).$$

□

In the case of random matrices with Gaussian entries, such as in the case of the SK-model, we have also have exponential concentration around these expectation values, as shown in [Pan13, Theorem 1.2].

Another family of random matrices for which we expect that $n\|\cdot\|$ provides the correct error scaling for cut norms are matrices of the form $B = A^*A$, where A again has i.i.d. entries of mean 0 and unit variance. Indeed, in [RV18] the authors show that

$$\mathbb{E}\left(\|A\|_{\infty\to 2}\right) \leq \mathcal{O}\left(\sqrt{n}\mathbb{E}\left(\|A\|_{2\to\infty}\right)\right).$$

with high probability. One can combine these recent results with more standard relations, like $\|A\|_{2\to\infty}^2 = \|B\|_{1\to\infty}$, $\|A\|_{2\to\infty} \leq \|A\|$ and $\|B\| = \|A\|^2$. This asserts $\mathbb{E}\left[\|B\|_{1\to\infty}\right] \leq n\mathbb{E}\left[\|B\|\right] = \mathcal{O}(n^2)$, while $\mathbb{E}\left[\|B\|_{\ell_1}\right] = \Omega(n^{2.5})$.

B Random instances of the MaxQP SDP

The following random instances of the MaxQP SDP have received significant attention in recent literature [MS16, KB20, JMRT16]: we define the Gaussian orthogonal ensemble (GOE) to be the random matrix distribution over symmetric $n \times n$ matrices A with i.i.d. normal entries $A_{ii} \sim \mathcal{N}(0, 2/n)$ on the diagonal and $A_{ij} \sim \mathcal{N}(0, 1/n)$ for $i < j$. For a parameter $\lambda \geq 0$, we also define the deformed GOE as $B(\lambda) = (\lambda/n)\mathbf{1}\mathbf{1}^T + A$, where $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^n$ is the vector of ones and A is sampled form the GOE.

This random matrix model is intimately connected to the Sherrington-Kirkpatric model and determining the MaxCut of random graphs. We will instantiate the notation from [MS16] and write $\text{MaxQP}(B(\lambda))$ for the optimal value of the MaxQP SDP with input $B(\lambda)$. In [MS16], Montanari ans Sen showed the following interesting result.

Theorem B.1 (Theorem 5 of [MS16]). *Fix $\lambda \geq 0$ and sample $B(\lambda)$ from the associated deformed GOE.*

1. *If $0 \leq \lambda \leq 1$, then for any $\epsilon > 0$, we have $\textsc{MaxQP}(B(\lambda))/n \in [2-\epsilon, 2+\epsilon]$ with probability converging to 1 as $n \to \infty$.*
2. *Else if $\lambda > 1$, then there exists a constant $\Delta(\lambda) > 0$ such that $\textsc{MaxQP}(B(\lambda))/n \geq 2 + \Delta(\lambda)$ with probability converging to 1 as $n \to \infty$.*

This seemingly abstract theorem has profound implications to our work. To appreciate them, it is worth noting that for $0 \leq \lambda \leq 1$, it is known that the maximal eigenvalue of $B(\lambda)$ is also contained in the interval $[2-\epsilon, 2+\epsilon]$ with high probability [KY13]. Thus, the optimal value of MaxQP is comparable in size to the re-scaled largest eigenvalue $n\|B(\lambda)\|$. Moreover, it also follows that for these instances $\|B(\lambda)\|_{\ell_1} = \Omega(\textsc{MaxQP}(B(\lambda))\sqrt{n}$ in expectation.

On the other hand, if $\lambda > 1$, the largest eigenvalue of the matrix and $B(\lambda)$ is given by $\lambda + \lambda^{-1}$ [KY13]. We see that both the largest eigenvalue and the value of $\textsc{MaxQP}(B(\lambda))$ go through a phase transition at $\lambda = 1$.

Let us now focus on the case $\lambda < 1$. Note that the dual of the MaxQP SDP with target matrix A is given by optimizing over $y \in \mathbb{R}^{n+1}$ as follows:

$$\begin{aligned} \text{minimize} \quad & n y_0 + \sum_{i=1}^{n} y_i \\ \text{subject to} \quad & y_0 I + \operatorname{diag}(y) \geq A, \; y_i \geq 0, \end{aligned} \qquad \text{(DUAL MaxQP SDP)} \tag{26}$$

where $\operatorname{diag}(y) = \operatorname{diag}(y_1, \ldots, y_n)$ denotes the diagonal matrix with entries y_i for $1 \leq i \leq n$. The additional dual variable y_0 arises from also incorporating the redundant constraint $\operatorname{tr}(X) \leq n$ in the associated primal SDP (3). This choice is motivated by the observation that previous quantum SDP solvers [BKL+17, vAGGdW17, vAG19] actually output approximately optimal solutions for the dual SDP with this redundant constraint.

Theorem B.1 implies that we can always find a trivial feasible point that is approximately optimal and sparse for Eq. (26). Indeed, for any $\epsilon > 0$, $\gamma = (2+\epsilon)e_0$ will be feasible with probability 1 in the limit $n \to \infty$. We conclude that for $\epsilon > 0$ fixed and in the regime $\lambda < 1$, solving the dual problem is trivial. So, existing solvers that output a feasible, approximately optimal solution [BKL+17, vAGGdW17, vAG19] are of little practical interest for the problem at hand. In stark contrast, feasible and approximately optimal solutions of the primal problem are still relevant, because they can be used to perform the rounding.

It is also important to note that the proof of [MS16, Theorem 5] is constructive. Indeed, let P_δ denote the the projector onto the range of the best rank-δn approximation of $B(\lambda)$, that is, the subspace spanned by the the eigenvectors corresponding to the largest δn eigenvalues of $B(\lambda)$. Moreover, let D with $(D)_{ii} = (P_\delta)_{ii}$ be the restriction of this projector to the main diagonal. By construction, $(D)_{ii} = (P_\delta)_{ii} > 0$ almost surely for all $1 \leq i \leq n$. And, in turn, $X = D^{-\frac{1}{2}} P_\delta D^{-\frac{1}{2}}$ must be a feasible point of the primal MaxQP SDP (2). Montanari and Sen then show that $\operatorname{tr}(B(\lambda)X) \geq 2 - \epsilon$ for some $\epsilon = \Omega(\delta)$. This establishes the first part of Thm. B.1. In summary, diagonalizing $B(\lambda)$ and computing X is sufficient to obtain an approximately feasible primal solution. Suppressing the error dependence on ϵ, diagonalizing $B(\lambda)$ takes $\mathcal{O}(n^\omega)$ time, while our classical algorithm to solve MaxQP SDP takes the same

up to polylogarithmic factors. The quantum runtime, however, is of order $\tilde{\mathcal{O}}(n^{2+o(1)})$ only. Thus, we see that for $\epsilon = \Theta(1)$ we obtain a quantum speedup as soon as the exponent of matrix multiplication obeys $\omega > 2$ (which is widely believed).

Let us now discuss the regime where $\lambda > 1$. To the best of our knowledge, the limit value of $\textsc{MaxQP}(B(\lambda))/n$ has not been identified yet. Ref. [MS16], however, shows that it must be strictly larger than 2 by constructing a sequence of feasible points that continues to saturate such a lower bound. However, in contrast to before ($\lambda < 1$), it is not clear that this sequence of feasible points is approximately optimal. In fact, it is not even known if $\textsc{MaxQP}(B(\lambda))/n < \lambda + \lambda^{-1}$.' But numerical evidence in favor of this behavior is provided in [JMRT16]. That is, the optimal value of the SDP is strictly smaller than the trivial eigenvalue upper bound. Thus, if it is indeed the case that $\textsc{MaxQP}(B(\lambda))/n+\mu < \lambda+\lambda^{-1}$ for some $\mu > 0$ as $n \to \infty$, then the dual SDP must be nontrivial to solve. Still, a direct solution of the primal problem is arguably more relevant, because it can be used to perform randomized rounding. Nevertheless, we will now argue that previous quantum solvers [BKL+17, vAGGdW17, vAG19] do not give rise to a speedup for the dual problem assuming that $\textsc{MaxQP}(B(\lambda))/n + \mu < \lambda + \lambda^{-1}$.

Before we move on, we once more emphasize that our algorithm considers the primal problem only. This is in stark contrast to existing quantum SDP solvers that address both primal and dual problem. This fully primal approach has the advantage that the runtime of the algorithm does not depend on problem-specific parameters like the $\|\cdot\|_{\ell_1}$ norm of approximately optimal dual solutions, as mentioned in Sec. 2.5. We will now show that this becomes a real advantage for solving the MaxQP SDP for $B(\lambda)$ in the regime $\lambda > 1$.

Proposition B.1. *Let $y^\sharp$ be a δ-approximately optimal solution to* dual *MaxQP SDP* (26) *for $B(\lambda)$ with $\lambda > 1$. Assume, moreover, that there exists a $\mu > 0$ such that that*

$$\textit{MaxQP}(B(\lambda))/n + \mu \leq \|B(\lambda)\|$$

with high probability. Then, with high probablity, every δ-optimal dual solution satisfies

$$\left\|y^\sharp\right\|_{\ell_1} = \Omega(n) \quad \textit{as long as } \delta < \mu. \tag{27}$$

Proof. Let x^* be the value achieved by $y^\sharp$ and set $\eta = \mu - \delta > 0$. If we condition on the event $x^* + \eta \leq \|B(\lambda)\|$, the feasibility constraint in the dual SDP (26) enforces $y_0 + \eta \leq \|B(\lambda)\|$. To see this, note that the value of the dual SDP is clearly monotonically increasing on the other entries, which are all positive. We will now show that in order for the matrix inequality

$$y_0 I + \operatorname{diag}(y^\sharp) \geq B(\lambda) \tag{28}$$

to hold, then the vector $y^\sharp \in \mathbb{R}^n$ must have large ℓ_1 norm. In order to do this, we will resort to an approximate leading eigenvector construction by [MS16]. This construction will have the desirable property that it is not too "spiky". In turn, this approximate leading eigenvector will have a small overlap with each entry of the diagonal matrix $\operatorname{diag}(y)$.

We will make extensive use of the results of [MS16], so we will also follow their notation and normalizations for this proof. Define u_1 to be the eigenvector corresponding to the largest

eigenvalue of $B(\lambda)/n$. Moreover, define the "capping" function $R(x)$ as

$$R(x) = \begin{cases} -1, & \text{if } x < -1 \\ x, & \text{if } -1 \leq x \leq 1 \\ 1, & \text{if } x > 1 \end{cases}$$

For some $\epsilon > 0$, Montanari and Sen then define the vector φ componentwise as $\varphi_i = R(\epsilon\sqrt{n}u_{1,i})$. In [MS16, Lemma G.2], they then establish

$$\left|\frac{1}{n}\operatorname{tr}\left(|\varphi\rangle\langle\varphi|B(\lambda)\right) - \epsilon^2\|B(\lambda)\|\right| = \mathcal{O}(\epsilon^4) \quad \text{with high probability.} \tag{29}$$

On top of that, in Eq. (163) they show that:

$$\frac{1}{n}\left\|\epsilon\sqrt{n}u_1 - \varphi\right\|_2^2 = \mathcal{O}(\epsilon^6). \tag{30}$$

We can now use the vector φ to probe positive semidefiniteness in Eq. (28):

$$\operatorname{tr}\left(|\varphi\rangle\langle\varphi|B(\lambda)\right) \leq \operatorname{tr}\left(|\varphi\rangle\langle\varphi|(y_0 I + \operatorname{diag}(y)\right). \tag{31}$$

Let us start by estimating the left hand side of this scalar inequality. Combining Eq. (30) with a reverse triangle inequality yields

$$\operatorname{tr}\left(|\varphi\rangle\langle\varphi|y_0 I\right) \leq y_0 n(\epsilon^2 + \epsilon^4).$$

Furthermore, by construction, the entries of φ squared to at most 1. Thus,

$$\operatorname{tr}\left(|\varphi\rangle\langle\varphi|\operatorname{diag}(y)\right) = \sum_{i=1}^{n} |\varphi|^2 y_i \leq \sum_{i=1}^{n} y_i$$

and we can combine both arguments to obtain an upper bound on the r.h.s. of Eq. (31). We can also lower-bound the l.h.s. Eq. (29) asserts

$$\operatorname{tr}\left(|\varphi\rangle\langle\varphi|B(\lambda)\right) \geq (\epsilon^2 + C\epsilon^4)n,$$

for some constant universal $C > 0$. Putting these inequalities together we conclude

$$y_0 n(\epsilon^2 + \epsilon^4) + \sum_{i=1}^{n} y_i \geq \epsilon^2\|B(\lambda)\|n - C\epsilon^4 n.$$

Dividing the inequality by $\epsilon^2 n$ and using the fact that our conditioning guarantees $y_0 + \eta \leq \|B(\lambda)\|$ we conclude

$$(\|B(\lambda)\| - \eta)(1 + \epsilon^2) + (n\epsilon^2)^{-1}\sum_{i=1}^{n} y_i \geq \|B(\lambda)\| - C\epsilon^4.$$

Rearranging the terms produces

$$n^{-1}\sum_{i=1}^{n} y_i \geq \epsilon^2(\eta - \epsilon^2\|B(\lambda)\| - C\epsilon^4).$$

Thus, we can pick ϵ small and n large enough to require that the right-hand side of the inequality above is of constant order (recall that $\|B(\lambda)\|$ does not depend on n). In contrast, the average $n^{-1}\sum_{i=1}^{n} y_i$ is of constant order if and only if $\|y\|_{\ell_1} = \Omega(n)$. □

As the methods of [BKL+17, vAGGdW17, vAG19] have a superquadratic dependency on $\|y\|_{\ell_1}$ for approximately optimal solutions, we conclude that their performance is worse than our algorithm for instances of MaxQP SDP with $B(\lambda)$ for $\lambda > 1$ with high probability. Of course, this only holds provided that the typical value of the SDP is a constant fraction away from the spectrum of $B(\lambda)$, as indicated by numerical evidence.

C Comparison to previous work and techniques for further improvement

This section is devoted to giving a brief overview over some promising proposals for speeding up SDP solvers for problems with a similar structure. The main message is that these unfortunately do not immediately apply to the general MaxQP SDP setting, especially for random signed matrices.

The main classical bottleneck behind Algorithm 1 is computing matrix exponentials. Dimension reduction techniques, like Johnson-Lindenstrauss, can sometimes considerably speed up this process, see e.g. [AK16]. There, Arora and Kale apply this idea to solve the MaxCut SDP up to a multiplicative error of $\mathcal{O}(\epsilon nd)$ in time $\tilde{\mathcal{O}}(nd)$ for a d regular graph on n vertices. Moreover, sparsification techniques [Ad20] can be used to bring this complexity down to $\tilde{\mathcal{O}}(n)$ in the adjacency list model and $\tilde{\mathcal{O}}(\min(nd, n^{1.5}d^{-1}))$ in the adjacency matrix input model. Note that the MaxCut SDP is just an instance of the MaxQP SDP, as both have the same constraints. The only difference is that the MaxCut SDP has the additional structure that the target matrix is the weighted adjacency matrix of a graph and, thus, has positive entries. The extra assumption of non-negative entries is a key ingredient behind the fastest approximate MaxCUT SDP solvers which would outperform the main results of this work. It is therefore worthwhile to discuss why these ideas do not readily extend to more general problem instances.

First, note that the fact that the entries of the target matrix has positive entries is crucial for the soundness of the oracle presented in [AK16, Theorem 5.2]. This already rules out the possibility of directly applying their methods to MAXQP if the matrix A has negative entries. The second crucial observation of [AK16] is that it is possible to rewrite the MaxCut SDP as:

$$\begin{aligned} \text{minimize} \quad & \sum_{i,j} [A]_{ij} \|v_i - v_j\|^2 \\ \text{subject to} \quad & \|v_i\|^2 = 1,\ v_i \in \mathbb{R}^n,\ i \in [n] \end{aligned} \tag{32}$$

In this reformulation, the vectors v_i correspond to columns of a Cholesky-decomposition associated with feasible points: $[X]_{ij} = \langle v_i | v_j \rangle$. Next, recall the following variation of the polarization identity:

$$\langle u|v\rangle = \frac{1}{2}\left(\|u\|^2 + \|v\|^2 - \|u - v\|^2\right).$$

This allows us to rewrite the original objective function as

$$\operatorname{tr}(AX) = \sum_{i,j} [A]_{ij} \langle v_i | v_j \rangle = \frac{1}{2} \sum_{i,j} [A]_{ij} \left(\|v_i\|^2 + \|v_j\|^2 - \|v_i - v_j\|^2\right).$$

Feasibility of X then demands $1 = \langle i|X|i\rangle = \langle v_i|v_i\rangle = \|v_i\|^2$ and we, thus, only need to optimize over $\|v_i - v_j\|^2$. Subsequently, Arora and Kale apply dimensionality reduction techniques to compute approximate vectors v_i', v_j' that satisfy:

$$\left|\|v_i - v_j\|^2 - \|v_i' - v_j'\|^2\right| \leq \epsilon\|v_i - v_j\|^2. \tag{33}$$

in time $\mathcal{O}(ns)$. A priori, similar techniques can be applied to the more general MaxQP SDP (3). However, sign problems can substantially affect the approximation error. Pointwise estimates like the one in (33) only suffice to estimate $\mathrm{tr}\,(XA)$ up to an error of order $\mathcal{O}(\epsilon\|A\|_{\ell_1})$. This is fine for matrices with non-negative entries, where this error scaling is comparable to the size of the optimal SDP solution. Matrix entries with different signs, however, may lead to cancellations that result in a much smaller size of the optimal SDP solution. In summary: adapting the ideas of Arora and Kale [AK16] is advisable in situations where the problem matrix obeys $\|A\|_{\ell_1} = \Theta(n\|A\|)$. This ensures a correct error behavior and dimension reduction allows for reducing the classical runtime to $\tilde{\mathcal{O}}(ns)$.

Another important technique for complexity reduction in SDPs is sparsification. Once again, one seminal example is MaxCut, where spectral sparsification methods can be used to reduce the complexity [ST11, KLP+16]. Here, the idea is to find a (usually random) sparser matrix B that has approximately the same cut value as A and then run the algorithm on B instead. Unfortunately, once again signed matrix entries render this approach problematic. Up to our knowledge, the best current sparsification results available for the $\infty \to 1$ norm are those of [Git13, Chapter 3]. There, the author shows in Corollary 3.9 that if we let B be a random matrix with independent random entries s.t. $\mathbb{E}(B_{ij}) = A_{ij}$, then

$$\mathbb{E}\left[\|A - B\|_{\infty\to 1}\right] \leq 2\sum_i \sqrt{\sum_j \mathrm{Var}[B_{ij}]}.$$

A necessary pre-requisite for accurate sparsification using the aforementioned result is therefore

$$2\sum_i \sqrt{\sum_j \mathrm{Var}[B_{ij}]} = \mathcal{O}(\epsilon n\|A\|)$$

It seems unlikely that it is possible to obtain good and general sparsification bounds from this result in our setting. To see why this is the case, note that in order for B to be sparse in expectation, we require that $\mathbb{P}(B_{ij} = 0) = p_{ij}$ for suitably large p_{ij}. This will result in a matrix that has, in expectation, $\sum_{ij}(1 - p_{i,j})$ nonzero entries. To make sure that the number of nonzero entries is not $\mathcal{O}(n^2)$, we need to set many $1 - p_{ij} = o(1)$. Now note that $\mathbb{P}[B_{ij} = 0] = p_{ij}$ and $\mathbb{E}[B_{ij}] = A_{ij}$ necessarily enforce $\mathbb{E}[(B_{ij}^2)] \geq \frac{p_{ij}}{1-p_{ij}}A_{ij}^2$. Thus, we see that we expect this technique to only work in the regime where A has many columns with entries that are $o(1)$ and can be neglected with high probability. Roughly speaking, this corresponds to the regime in which $\|A\|_{\mathrm{col}} \ll n\|A\|$. It is then easy to see tha the random matrices considered before do not satisfy this and, thus, we do not expect that those instances can be sparsified.

Last, but not least, we emphasize that it easy to construct examples where the error term $\|A\|_{\ell_1}$ conveys the right scaling, not $n\|A\|$. A concrete example are extremely sparse matrices, where all but $s \ll n$ of the entries are zero.

Bibliography

[AA11] S. Aaronson and A. Arkhipov. The computational complexity of linear optics. In *STOC'11—Proceedings of the 43rd ACM Symposium on Theory of Computing*, pages 333–342. ACM, New York, 2011.

[AAB+19] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. S. L. Brandao, D. A. Buell, B. Burkett, Y. Chen, Z. Chen, B. Chiaro, R. Collins, W. Courtney, A. Dunsworth, E. Farhi, B. Foxen, A. Fowler, C. Gidney, M. Giustina, R. Graff, K. Guerin, S. Habegger, M. P. Harrigan, M. J. Hartmann, A. Ho, M. Hoffmann, T. Huang, T. S. Humble, S. V. Isakov, E. Jeffrey, Z. Jiang, D. Kafri, K. Kechedzhi, J. Kelly, P. V. Klimov, S. Knysh, A. Korotkov, F. Kostritsa, D. Landhuis, M. Lindmark, E. Lucero, D. Lyakh, S. Mandrà, J. R. McClean, M. McEwen, A. Megrant, X. Mi, K. Michielsen, M. Mohseni, J. Mutus, O. Naaman, M. Neeley, C. Neill, M. Y. Niu, E. Ostby, A. Petukhov, J. C. Platt, C. Quintana, E. G. Rieffel, P. Roushan, N. C. Rubin, D. Sank, K. J. Satzinger, V. Smelyanskiy, K. J. Sung, M. D. Trevithick, A. Vainsencher, B. Villalonga, T. White, Z. J. Yao, P. Yeh, A. Zalcman, H. Neven, and J. M. Martinis. Quantum supremacy using a programmable superconducting processor. *Nature*, **574**(7779):505–510, 2019.

[AB09] S. Arora and B. Barak. *Computational Complexity - A Modern Approach*. Cambridge University Press, 2009.

[AS08] N. Alon and J. H. Spencer. *The Probabilistic Method, Third Edition*. Wiley-Interscience series in discrete mathematics and optimization. Wiley, 2008.

[ASZ+21] A. Abbas, D. Sutter, C. Zoufal, A. Lucchi, A. Figalli, and S. Woerner. The power of quantum neural networks. *Nat. Comput. Sci.*, **1**(6):403–409, 2021.

[Bar02] A. I. Barvinok. *A course in convexity*, volume 54 of *Graduate studies in mathematics*. American Mathematical Society, 2002.

[BBC+95] A. Barenco, C. H. Bennett, R. Cleve, D. P. DiVincenzo, N. Margolus, P. Shor, T. Sleator, J. A. Smolin, and H. Weinfurter. Elementary gates for quantum computation. *Phys. Rev. A*, **52**:3457–3467, 1995.

[BBC+19] S. Bravyi, D. Browne, P. Calpin, E. Campbell, D. Gosset, and M. Howard. Simulation of quantum circuits by low-rank stabilizer decompositions. *Quantum*, **3**:181, 2019.

[BCHJ+21] F. G. Brandão, W. Chemissany, N. Hunter-Jones, R. Kueng, and J. Preskill. Models of quantum complexity growth. *PRX Quantum*, **2**:030316, 2021.

[BCJ+19] A. Bohrdt, C. S. Chiu, G. Ji, M. Xu, D. Greif, M. Greiner, E. Demler, F. Grusdt, and M. Knap. Classifying snapshots of the doped hubbard model with machine learning. *Nat. Phys.*, **15**(9):921–924, 2019.

[BGM21] S. Bravyi, D. Gosset, and R. Movassagh. Classical algorithms for quantum mean values. *Nat. Phys.*, **17**(3):337–341, 2021.

[BJS11] M. J. Bremner, R. Jozsa, and D. J. Shepherd. Classical simulation of commuting quantum computations implies collapse of the polynomial hierarchy. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, **467**(2126):459–472, 2011.

[BK02] S. B. Bravyi and A. Y. Kitaev. Fermionic quantum computation. *Ann. Phys.*, **298**(1):210 – 226, 2002.

[BK21] L. Bittel and M. Kliesch. Training variational quantum algorithms is np-hard–even for logarithmically many qubits and free fermionic systems. *preprint arXiv:2101.07267*, 2021.

[BKF19] F. G. L. Brandao, R. Kueng, and D. S. França. Faster quantum and classical sdp approximations for quadratic binary optimization. *preprint arXiv:1909.04613*, 2019.

[BKW21] L. Burgholzer, R. Kueng, and R. Wille. Random stimuli generation for the verification of quantum circuits. In *ASPDAC '21: 26th Asia and South Pacific Design Automation Conference, Tokyo, Japan, January 18-21, 2021*, pages 767–772. ACM, 2021.

[BMBO20] X. Bonet-Monroig, R. Babbush, and T. E. O'Brien. Nearly optimal measurement scheduling for partial tomography of quantum states. *Phys. Rev. X*, **10**:031064, 2020.

[BS17] F. G. Brandão and K. M. Svore. Quantum speed-ups for solving semidefinite programs. In *58th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2017*, pages 415–426. IEEE Computer Soc., Los Alamitos, CA, 2017.

[BV04] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004.

[BWK20] L. Burgholzer, R. Wille, and R. Kueng. Characteristics of reversible circuits for error detection. *preprint arXiv:2012.02037*, 2020.

[CAB+20] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, et al. Variational quantum algorithms. *preprint arXiv:2012.09265*, 2020.

[CC20] N. J. Coble and M. Coudron. Quasi-polynomial time approximation of output probabilities of constant-depth, geometrically-local quantum circuits. *preprint arXiv:2012.05460*, 2020.

[CGW10] X. Chen, Z. C. Gu, and X. G. Wen. Local unitary transformation, long-range quantum entanglement, wave function renormalization, and topological order. *Phys. Rev. B*, **82**:155138, 2010.

[CHKT20] C.-F. Chen, H.-Y. Huang, R. Kueng, and J. A. Tropp. Quantum simulation via randomized product formulas: Low gate complexity with accuracy guarantees. *preprint arXiv:2008.11751*, 2020.

[CRT06] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, **52**(2):489–509, 2006.

[CS17] A. N. Chowdhury and R. D. Somma. Quantum algorithms for Gibbs sampling and hitting-time estimation. *Quantum Inf. Comput.*, **17**(1-2):41–64, 2017.

[CSV13] E. J. Candès, T. Strohmer, and V. Voroninski. PhaseLift: exact and stable signal recovery from magnitude measurements via convex programming. *Comm. Pure Appl. Math.*, **66**(8):1241–1274, 2013.

[CT17] G. Carleo and M. Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, **355**(6325):602–606, 2017.

[Don06] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, **52**(4):1289–1306, 2006.

[EHW+20] J. Eisert, D. Hangleiter, N. Walk, I. Roth, D. Markham, R. Parekh, U. Chabaud, and E. Kashefi. Quantum certification and benchmarking. *Nat. Rev. Phys.*, **2**(7):382–390, 2020.

[EKH+20] A. Elben, R. Kueng, H.-Y. R. Huang, R. van Bijnen, C. Kokail, M. Dalmonte, P. Calabrese, B. Kraus, J. Preskill, P. Zoller, and B. Vermersch. Mixed-state entanglement from local randomized measurements. *Phys. Rev. Lett.*, **125**:200501, 2020.

[FBK21] D. S. França, F. G. S. L. Brandão, and R. Kueng. Fast and robust quantum state tomography from few basis measurements. In M. Hsieh, editor, *16th Conference on the Theory of Quantum Computation, Communication and Cryptography, TQC 2021,*

July 5-8, 2021, Virtual Conference, volume 197 of *LIPIcs*, pages 7:1–7:13. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021.

[Fey82] R. P. Feynman. Simulating physics with computers. volume 21, pages 467–488. 1981/82. Physics of computation, Part II (Dedham, Mass., 1981).

[FGG14] E. Farhi, J. Goldstone, and S. Gutmann. A quantum approximate optimization algorithm. *preprint arXiv:1411.4028*, 2014.

[FSK+21] P. K. Faehrmann, M. Steudtner, R. Kueng, M. Kieferova, and J. Eisert. Randomizing multi-product formulas for improved hamiltonian simulation. *preprint arXiv:2101.07808*, 2021.

[GAN14] I. M. Georgescu, S. Ashhab, and F. Nori. Quantum simulation. *Rev. Mod. Phys.*, **86**:153–185, 2014.

[GH62] R. E. Gomory and T. C. Hu. An application of generalized linear programming to network flows. *J. Soc. Indust. Appl. Math.*, **10**:260–283, 1962.

[GKFW21] T. Grurl, R. Kueng, J. Fuß, and R. Wille. Stochastic quantum circuit simulation using decision diagrams. In *Design, Automation & Test in Europe Conference & Exhibition, DATE 2021, Grenoble, France, February 1-5, 2021*, pages 194–199. IEEE, 2021.

[GKK17] D. Gross, F. Krahmer, and R. Kueng. Improved recovery guarantees for phase retrieval from coded diffraction patterns. *Appl. Comput. Harmon. Anal.*, **42**(1):37–64, 2017.

[GKKT20] M. Guţă, J. Kahn, R. Kueng, and J. A. Tropp. Fast state tomography with optimal error bounds. *J. Phys. A*, **53**(20):204001, 28, 2020.

[GLT18] A. Gilyén, S. Lloyd, and E. Tang. Quantum-inspired low-rank stochastic regression with logarithmic dependence on the dimension. *preprint arXiv:1811.04909*, 2018.

[Gro97] L. K. Grover. Quantum mechanics helps in searching for a needle in a haystack. *Phys. Rev. Lett.*, **79**:325–328, 1997.

[Gro11] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory*, **57**(3):1548–1566, 2011.

[GW95] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. Assoc. Comput. Mach.*, **42**(6):1115–1145, 1995.

[Had21] C. Hadfield. Adaptive pauli shadows for energy estimation. *preprint arXiv:2105.12207*, 2021.

[HBM+21] H.-Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, and J. R. McClean. Power of data in quantum machine learning. *Nat. Commun.*, **12**(1):2631, 2021.

[HBRM20] C. Hadfield, S. Bravyi, R. Raymond, and A. Mezzacapo. Measurements of quantum Hamiltonians with locally-biased classical shadows. *preprint arXiv:2006.15788*, 2020.

[HCT+19] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, **567**(7747):209–212, 2019.

[HH13] D. M. Harris and S. L. Harris. 2 - combinational logic design. In D. M. Harris and S. L. Harris, editors, *Digital Design and Computer Architecture (Second Edition)*, pages 54–106. Morgan Kaufmann, Boston, second edition edition, 2013.

[HHL09] A. W. Harrow, A. Hassidim, and S. Lloyd. Quantum algorithm for linear systems of equations. *Phys. Rev. Lett.*, **103**:150502, 2009.

[HHR+21] S. Hillmich, C. Hadfield, R. Raymond, A. Mezzacapo, and R. Wille. Decision diagrams for quantum measurements with shallow circuits. *preprint arXiv:2105.06932*, 2021.

[HKMW21] S. Hillmich, R. Kueng, I. L. Markov, and R. Wille. As accurate as needed, as efficient as possible: Approximations in dd-based quantum circuit simulation. In *Design, Automation & Test in Europe Conference & Exhibition, DATE 2021, Grenoble, France, February 1-5, 2021*, pages 188–193. IEEE, 2021.

[HKP20] H.-Y. Huang, R. Kueng, and J. Preskill. Predicting many properties of a quantum system from very few measurements. *Nat. Phys.*, **16**:1050—1057, 2020.

[HKP21a] H.-Y. Huang, R. Kueng, and J. Preskill. Efficient estimation of pauli observables by derandomization. *Phys. Rev. Lett.*, **127**:030503, 2021.

[HKP21b] H.-Y. Huang, R. Kueng, and J. Preskill. Information-theoretic bounds on quantum advantage in machine learning. *Phys. Rev. Lett.*, **126**:190505, 2021.

[HKT+21] H.-Y. Huang, R. Kueng, G. Torlai, V. V. Albert, and J. Preskill. Provably efficient machine learning for quantum many-body problems. *preprint arXiv:2106.12627*, 2021.

[Jay57] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev. (2)*, **106**:620–630, 1957.

[JKM19] P. Jung, R. Kueng, and D. Mixon. Derandomizing compressed sensing with combinatorial design. *Front. Appl. Math.*, **5**:26, 2019.

[JW28] P. Jordan and E. Wigner. Über das paulische Äquivalenzverbot. *Zeitschrift für Physik*, **47**(9):631–651, 1928.

[KKEG19] M. Kliesch, R. Kueng, J. Eisert, and D. Gross. Guaranteed recovery of quantum processes from few measurements. *Quantum*, **3**:171, 2019.

[KLDF16] R. Kueng, D. M. Long, A. C. Doherty, and S. T. Flammia. Comparing experiments to the fault-tolerance threshold. *Phys. Rev. Lett.*, **117**:170502, 2016.

[KMV19] R. Kueng, D. G. Mixon, and S. Villar. Fair redistricting is hard. *Theoret. Comput. Sci.*, **791**:28–35, 2019.

[KMvB$^+$19] C. Kokail, C. Maier, R. van Bijnen, T. Brydges, M. K. Joshi, P. Jurcevic, C. A. Muschik, P. Silvi, R. Blatt, C. F. Roos, et al. Self-verifying variational quantum simulation of lattice models. *Nature*, **569**(7756):355–360, 2019.

[KP20] I. Kerenidis and A. Prakash. Quantum gradient descent for linear systems and least squares. *Phys. Rev. A*, **101**:022316, 2020.

[KRT17] R. Kueng, H. Rauhut, and U. Terstiege. Low rank matrix recovery from rank one measurements. *Appl. Comput. Harmon. Anal.*, **42**(1):88–116, 2017.

[KSV02] A. Y. Kitaev, A. H. Shen, and M. N. Vyalyi. *Classical and Quantum Computation*, volume 47 of *Graduate studies in mathematics*. American Mathematical Society, 2002.

[KT19] R. Kueng and J. A. Tropp. Binary component decomposition part II: The asymmetric case. *preprint arXiv:1907.13602*, 2019.

[KT21] R. Kueng and J. A. Tropp. Binary component decomposition part I: The positive-semidefinite case. *SIAM J. Math. Data Sci.*, **3**(2):544–572, 2021.

[Kue19] R. Kueng. Quantum and classical information processing with tensors (lecture notes), Spring 2019. Caltech course notes: `https://iqim.caltech.edu/classes`.

[LMR13] S. Lloyd, M. Mohseni, and P. Rebentrost. Quantum algorithms for supervised and unsupervised machine learning. *preprint arXiv:1307.0411*, 2013.

[LMR14] S. Lloyd, M. Mohseni, and P. Rebentrost. Quantum principal component analysis. *Nat. Phys.*, **10**(9):631–633, 2014.

[MOS14] R. Mansini, W. Ogryczak, and M. G. Speranza. Twenty years of linear programming based portfolio optimization. *European J. Oper. Res.*, **234**(2):518–535, 2014.

[MR95] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[NC00] M. A. Nielsen and I. L. Chuang. *Quantum computation and quantum information*. Cambridge University Press, Cambridge, 2000.

[PMS+14] A. Peruzzo, J. McClean, P. Shadbolt, M.-H. Yung, X.-Q. Zhou, P. J. Love, A. Aspuru-Guzik, and J. L. O'Brien. A variational eigenvalue solver on a photonic quantum processor. *Nature Communications*, **5**(1):4213, 2014.

[Pre18] J. Preskill. Quantum Computing in the NISQ era and beyond. *Quantum*, **2**:79, 2018.

[RFK+18] I. Roth, A. Flinth, R. Kueng, J. Eisert, and G. Wunder. Hierarchical restricted isometry property for kronecker product measurements. In *56th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2018, Monticello, IL, USA, October 2-5, 2018*, pages 632–638. IEEE, 2018.

[RFP10] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, **52**(3):471–501, 2010.

[RKK+18] I. Roth, R. Kueng, S. Kimmel, Y.-K. Liu, D. Gross, J. Eisert, and M. Kliesch. Recovering quantum gates from few average gate fidelities. *Phys. Rev. Lett.*, **121**:170502, 2018.

[RSML18] P. Rebentrost, A. Steffens, I. Marvian, and S. Lloyd. Quantum singular-value decomposition of nonsparse low-rank matrices. *Phys. Rev. A*, **97**:012327, 2018.

[Sho94] P. W. Shor. Algorithms for quantum computation: discrete logarithms and factoring. In *35th Annual Symposium on Foundations of Computer Science (Santa Fe, NM, 1994)*, pages 124–134. IEEE Comput. Soc. Press, Los Alamitos, CA, 1994.

[SS14] D. Stanford and L. Susskind. Complexity and Shock Wave Geometries. *Phys. Rev.*, **D90**:126007, 2014.

[Sus16a] L. Susskind. Computational complexity and black hole horizons. *Fortschr. Phys.*, **64**(1):24–43, 2016.

[Sus16b] L. Susskind. Entanglement is not enough. *Fortsch. Phys.*, **64**:49, 2016.

[Tan18] E. Tang. Quantum-inspired classical algorithms for principal component analysis and supervised clustering. *preprint arXiv:1811.00414*, 2018.

[Tan19] E. Tang. A quantum-inspired classical algorithm for recommendation systems. In *STOC'19—Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 217–228. ACM, New York, 2019.

[TOV+11] K. Temme, T. J. Osborne, K. G. Vollbrecht, D. Poulin, and F. Verstraete. Quantum metropolis sampling. *Nature*, **471**(7336):87–90, 2011.

[vAGGdW17a] J. van Apeldoorn, A. Gilyén, S. Gribling, and R. de Wolf. Quantum SDP-solvers: better upper and lower bounds. In *58th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2017*, pages 403–414. IEEE Computer Soc., Los Alamitos, CA, 2017.

[vAGGdW17b] J. van Apeldoorn, A. Gilyén, S. Gribling, and R. de Wolf. Quantum SDP-solvers: better upper and lower bounds. In *58th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2017*, pages 403–414. IEEE Computer Soc., Los Alamitos, CA, 2017.

[Vid03] G. Vidal. Efficient classical simulation of slightly entangled quantum computations. *Phys. Rev. Lett.*, **91**:147902, 2003.

[vNLH17] E. P. L. van Nieuwenburg, Y.-H. Liu, and S. D. Huber. Learning phase transitions by confusion. *Nat. Phys.*, **13**(5):435–439, 2017.

[Wat18] J. Watrous. *The Theory of Quantum Information*. Cambridge University Press, 2018.

[WHB21] R. Wille, S. Hillmich, and L. Burgholzer. Tools for quantum computing based on decision diagrams. *preprint arXiv:2108.07027*, 2021.

[ZW19] A. Zulehner and R. Wille. Advanced simulation of quantum computations. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.*, **38**(5):848–859, 2019.

Don’t mind your make-up, you’d better make your mind up.

Frank Zappa