On the average-case complexity of learning output distributions of quantum circuits

Alexander Nietner[1,2], Marios Ioannou[1,2], Ryan Sweke[1,3], Richard Kueng[1,4], Jens Eisert[1,2], Marcel Hinsche[1,2], and Jonas Haferkamp[1,2,5,6]

[1]Author list in pseudorandom order. All authors contributed equally.
[2]Dahlem Center for Complex Quantum Systems, Freie Universität Berlin, Germany
[3]IBM Quantum, Almaden Research Center, San Jose, CA 95120, USA
[4]Institute for Integrated Circuits, Department of Computer Science, Johannes Kepler University Linz, Austria
[5]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02318, USA
[6]Department of Mathematics, Saarland University, Saarbrücken, Germany

In this work, we show that learning the output distributions of brickwork random quantum circuits is average-case hard in the statistical query model. This learning model is widely used as an abstract computational model for most generic learning algorithms. In particular, for brickwork random quantum circuits on n qubits of depth d, we show three main results:

- At super logarithmic circuit depth $d = \omega(\log(n))$, any learning algorithm requires super polynomially many queries to achieve a constant probability of success over the randomly drawn instance.
- There exists a $d = O(n)$, such that any learning algorithm requires $\Omega(2^n)$ queries to achieve a $\Omega(2^{-n})$ probability of success over the randomly drawn instance.
- At infinite circuit depth $d \to \infty$, any learning algorithm requires $2^{2^{\Omega(n)}}$ many queries to achieve a $2^{-2^{O(n)}}$ probability of success over the randomly drawn instance.

As an auxiliary result of independent interest, we show that the output distribution of a brickwork random quantum circuit is constantly far from any fixed distribution in total variation distance with probability $1 - O(2^{-n})$, which confirms a variant of a conjecture by Aaronson and Chen.

Alexander Nietner: a.nietner@fu-berlin.de
Marios Ioannou: marios.ioannou@fu-berlin.de
Marcel Hinsche: m.hinsche@fu-berlin.de

Contents

1 Introduction

Quantum circuits are of central importance in quantum computing and serve as a discrete toy model for the physical world. Understanding the intrinsic properties of quantum circuits is thus of fundamental interest. One such property is quantum circuit complexity, which corresponds to the minimum number of elementary gates necessary to implement a given quantum circuit. Another such property is the complexity of simulating quantum circuits, which is the basis for many quantum advantage proposals. There, one asks what computational resources are required for sampling from the output distribution of a given quantum circuit when applied to a fixed input product state and when measured in the computational basis. In this work, we take the perspective of learning theory, and study the complexity of learning the output distribution of quantum circuits. At a high level, this amounts to the resources required to reproduce samples according to the output distribution of a quantum circuit when given black-box access to the corresponding output distribution.

More specifically, we study the *average case* complexity of learning the output distributions of quantum circuits. This amounts to the cost of learning when the quantum circuit is drawn randomly according to some measure. We note that in the setting of quantum circuit complexity the average-case setting has been the subject of intense work, due to connections between randomly drawn quantum circuits and holographic models in high-energy physics [BCH+21]. Similarly, the established average case hardness of classically simulating quantum circuits has been fundamental to proposals for quantum advantage [HE22]. In our setting, we ask the following question:

What is the complexity of learning generic quantum circuit output distributions?

We answer this question within the *statistical query* (SQ) framework by proving lower bounds on the query complexity required to learn only a fraction of the output distributions of random quantum circuits. We note that this problem is the average case version of the natural quantum extension of learning the output distributions of classical circuits [Kea93]. Additionally, it also has a physical interpretation. Specifically, on a high level, the learning problem we consider corresponds to the setting where an observer living in a world governed by quantum physics is to learn a model of its environment with respect to the outcomes of measurements in a fixed basis.

We stress that while in this work we consider a "model" of the unknown quantum process to be an algorithm for generating new samples from the distribution obtained by measuring the output state in a fixed basis, there are indeed many other notions of both "model" and access to the unknown quantum process one could consider. For example, one could allow access to copies of the output state of the quantum circuit, and ask for an algorithm which is only required to provide estimates of expectation values for some fixed set of observables [Ad17; AA24]. Alternatively, one might even only be interested in testing whether the unknown quantum process satisfies a certain property or not (for example, such as being a Clifford circuit). Each such setting will model a different scenario of interest, however in this work we

consider the setting of learning a model to generate new samples from fixed basis measurements, in light of both the physical and learning-theoretic interpretations discussed above, as well as a variety of connections to quantum machine learning.

More specifically, the output distributions of parameterized local quantum circuits with respect to computational basis measurements are often used as a model class in quantum machine learning approaches to probabilistic modelling, where they are referred to as *quantum circuit Born machines* (QCBMs). As such, understanding the extent to which learning algorithms for QCBMs may offer advantages over purely classical approaches, such as those based on deep neural networks, is currently the subject of much interest [HIN+23; CMD+20]. Our work can be seen as investigating the average-case learnability of the model class of QCBMs, and as such, a rigorous first step towards characterizing the limitations of QCBM based algorithms.

Additionally, a core technical ingredient of our results has ramifications for quantum advantage proposals based on sampling from the output distributions of random quantum circuits. We prove that with overwhelming probability, the output distribution of a random quantum circuit will have at least some constant total-variation distance from the uniform distribution. This resembles a conjecture by Aaronson and Chen [AC17] where they conjecture the same statement with different constants. They required and proved a weaker form of this statement for establishing the complexity theoretic foundations of quantum advantage proposals. We provide more details on the connections between our work and the variety of related works in Section 1.3.

1.1 Set-up

General framework: We say that a class $\mathcal{D}$ of distributions can be learned by an algorithm $\mathcal{A}$ if, when given access to any $P \in \mathcal{D}$, the algorithm returns a description of some close distribution Q. In particular, for an accuracy $\epsilon > 0$ we say that $\mathcal{A}$ ϵ-learns P if it returns a description of a Q which is ϵ-close in total variation distance. An algorithm is said to be a statistical query algorithm if it has access to P only via approximate expectation values instead of individual samples, where the approximation is promised to be within a tolerance τ. This is not only a handy restriction on the algorithm that makes analysis simpler, but also practically inspired since most heuristic algorithms are of this form [FGR+17]. In particular, for $\tau = \Omega(1/\mathrm{poly}(n))$ the statistical query oracle can be simulated from polynomially many samples. Formal definitions are given in Section 2.

Average case complexity: The average case complexity of an algorithm characterizes the cost to achieve a certain probability of success with respect to a measure over the instances[1]. The (deterministic) average case query complexity with respect to μ and β corresponds to the minimal number of queries any algorithm must make

[1]Thus, we work in the Monte Carlo framework of random algorithms. One can likewise characterize the average case complexity by the *expectation* of success, which then corresponds to the Las Vegas framework. The Monte Carlo framework is more general in our case as the task we consider can in general not be verified efficiently.

in order to have at least a success probability of β with respect to $P \sim \mu$. The randomized average case query complexity is defined in the same manner, though introducing another parameter α capturing the success probability over the internal (classical or quantum) randomness of the algorithm.

Quantum and probabilistic algorithms: As explained in more detail in Section F the bounds for deterministic algorithms apply almost exactly to random, i.e., probabilistic or quantum, algorithms. In particular, the randomized average case complexity for ϵ-learning is lower bounded by the deterministic average case complexity of ϵ-learning up to a prefactor of $2(\alpha - 1/2)$ (c.f. Theorem 1 in comparison with Theorem 34), where we denote by α the success probability with respect to the internal randomness of the algorithm. Thus, for the sake of ease of presentation, throughout this work we focus on the deterministic case. We refer to Section F for the extension to probabilistic and quantum algorithms.

Random quantum circuits: The distribution class $\mathcal{D}$ we consider is given by QCBMs, the set of Born distributions corresponding to brickwork quantum circuits at depth d and with gates from a gate set $\mathcal{G}$. In particular, we consider distributions of the form

$$P_U(x) = |\langle x|U|0^n\rangle|^2 , \tag{1}$$

with U being some unitary brickwork circuit composed of gates from the gate-set $\mathcal{G}$. While our main focus is on $\mathcal{G} = \mathrm{U}(4)$, the set of unitary two qubit gates, our techniques will carry over and give similar results for discrete approximations of $\mathrm{U}(4)$.

Average case bounds and their interpretation depend on the choice of the underlying measure. In this work, we consider the measure over distributions P_U that is induced by sampling a random quantum circuit U. This is, each gate is sampled individually from the the uniform measure over the gate set $\mathcal{G}$. For $\mathcal{G} = \mathrm{U}(4)$ this measure is well studied and known as the unitary Haar measure. One practical aspect of this measure is that it corresponds to a natural notion of a generic distribution that would potentially be sampled in the lab by individually sampling each local gate. Other interesting measures would be the uniform distribution over the actual distribution class $\mathcal{D}$. However, in case of $\mathcal{G} = \mathrm{U}(4)$, this class is continuous. Hence, defining the uniform measure would require a suitable discretization, e.g., by means of an ϵ-net. On the other hand, as we will show, the measure induced by random quantum circuits will be bias free in the sense that for any distribution $P \in \mathcal{D}$, the probability of sampling some distribution close to P is exponentially small. Thus, similar to the uniform distribution, the measure we consider does not introduce any bias towards any distribution.

1.2 Our results

The main contribution of this work consists in characterizing the average-case complexity of learning the output-distributions of random quantum circuits for different circuit depths. Here we provide an informal overview of these results.

Our main focus is on the scaling of the average case complexity with respect to both circuit depth d and the success probability over the randomly drawn instance β:

1. **Circuit depth d:** As the circuit depth increases, the expressivity of the set of distributions we consider increases as well. At $d = 1$ all output distributions of local quantum circuit are product distributions and hence, can be learned trivially. At infinite depth, in contrast, local quantum circuit output distributions can represent any distribution. Thus, they are not learnable in the worst case. As such, the complexity of learning must scale with the circuit depth. We we are interested in understanding this dependence.

2. **Probability over instances β:** Intuitively the larger β, the harder we expect the average-case learning task to be. Specifically, setting $\beta = 1$ recovers the worst-case setting, where we require the learning algorithm to succeed on *all* instances. Setting $\beta < 1$ implies that we only require the learning algorithm to succeed on a fraction of instances, which makes the average-case learning problem easier. In this work we investigate how small β can be made, while still guaranteeing hardness of average-case learning.

In addition to the dependency of average-case query-complexity on d and β, we also investigate the dependence on both the tolerance of the statistical queries τ and the desired accuracy ϵ. For ease of presentation, in the informal results below we suppress the dependencies on ϵ and τ and focus on the case where $\tau = \Omega(1/\mathrm{poly}(n))$ and ϵ is a sufficiently small constant. We refer to the formal statements for details.

Finally, we stress again that while the results given below are stated for deterministic algorithms, they immediately translate to both probabilistic and quantum algorithms. This correspondence is sketched in Section 1.1 while the details can be found in Section F. With this in mind, we state the main results of this work as follows:

Informal Theorem 1: *Let ϵ small and n large enough and let $\tau = \Omega(1/\mathrm{poly}(n))$. Let $\mathcal{A}$ be an algorithm for ϵ-learning the output distributions of brickwork random quantum circuits of depth d from q many τ-accurate statistical queries. Then it holds*

1. ***Infinite depth:*** *When $d \to \infty$, $q = 2^{2^{\Omega(n)}}$ queries are necessary for any $\beta > 2\exp(-2^{n-2}/9\pi^3) = 2^{-2^{\Omega(n)}}$ (c.f. Theorem 2).*

2. ***Linear depth:*** *There is a $d' = O(n)$ such that for any $d \geq d'$, $q = \Omega(2^n)$ queries are necessary for any $\beta > 3200 \cdot 2^{-n} = O(2^{-n})$ (c.f. Theorem 6).*

3. ***Sublinear depth:*** *for any $c \log n \leq d \leq c(n + \log n)$, $q = 2^{\Omega(d)} = 2^{\omega(\log(n))}$ queries are necessary for any $\beta > 4/5 + \epsilon + \tau = O(1)$, where $c = 1/\log(5/4)$ (c.f. Theorem 12).*

In particular, we find that the average case problem at superlogarithmic depth is hard with constant probability over the random instance. Moreover, at linear depth we find hardness with probability exponentially close to one. At infinite depth, the problem becomes hard with probability double exponentially close to one.

A natural question concerns the tightness of these statistical query (SQ) lower bounds. We note that a crude SQ upper bound for any distribution learning problem is simply given by the cardinality $\mathcal{N}$ of an ϵ-net over the set of distributions to be learned. This follows from the fact that for any given pair of distributions in the ϵ-net, there is a corresponding optimal distinguishing query function. Such distinguishing queries can be used in a tournament-style SQ algorithm over $\mathcal{N}-1$ rounds to identify the unknown distribution. For brickwork circuits of depth d, one may take $\mathcal{N}$ to be the size of a corresponding ϵ-net over the quantum circuits, giving $\mathcal{N} = \exp[O(nd\log(nd/\epsilon))]$ [ZLK+24]. This can be compared to the above SQ lower bounds: for instance, for linear depth $d = O(n)$ and constant ϵ, we find an SQ upper bound of $q = \exp[O(n^2\log(n))]$ which is larger but comparable to our lower bound of $q = \Omega(2^n)$. We leave it as an open problem to provide more tightly matching upper bounds to the lower bounds proved in informal Theorem 1.

As a side result we show that the output distribution of a random quantum circuit is, with overwhelming probability, at least constantly far in total variation distance from any fixed distribution. This resolves a variant of Aaronson and Chen's [AC17, Conjecture 1], made with the goal of clarifying the hardness of verifying random circuit sampling procedures. As such we believe this auxiliary result to be of independent interest.

Informal Theorem 2: (Informal version of Theorem 36) *There exists some $d' = O(n)$ such that for any depth $d \geq d'$, for any $\epsilon \leq 1/225$ and for any distribution Q we have*

$$\Pr_{U\sim\mu}\left[\mathrm{d_{TV}}(P_U, Q) > \epsilon\right] \geq 1 - O\left(2^{-n}\right), \tag{2}$$

where μ is the measure induced by random brickwork quantum circuits.

1.3 Related work

Our work touches on and combines a variety of well studied fields. In order to provide further context and motivation, we provide below a discussion of relevant related work.

Statistical queries: We work in the statistical query (SQ) framework which was introduced by Kearns [Kea93] as a restriction of Valiants theory of learning [Val84]. Kearns original motivation was the intrinsic robustness of SQ learners with respect to random classification noise. However, the SQ model is also highly relevant in the context of statistical problems [FGR+17], which includes distribution learning, as originally formulated by Kearns et al. [KMR+94]. In this context, SQ algorithms are those which only have access to coarse statistical properties of the data generating distribution. While this is a restriction of the oracle access it turns out, with the famous exception of Gaussian elimination, that almost all known learning algorithms can be recast as SQ algorithm [Kea93; Fel17].

The SQ framework is particularly interesting in the context of variational quantum machine learning, such as QCBM based algorithms. To see this we note that

all current methods for optimizing the QCBM's parameters use noisy evaluations of gradients, or gradient-like quantities, along individual directions. Examples for this include stochastic gradient descent via the parameter shift rule [MNK+18; SBG+19] and simultaneous perturbations and stochastic approximation [Spa98]. Thus, the update in QCBM based algorithms can be directly implemented via statistical queries. The SQ framework has also been used to model variational quantum algorithms beyond QCBMs and prove rigorous lower bounds for them [AK22; Nie23].

Equivalences to other statistical oracles, such as the "honest SQ" oracle and the statistical query oracle with respect to Bernoulli noise are worked out in [FGR+17; Fel17]. Interestingly, SQ learning has been shown to be equivalent to learning with restricted memory [SVW16; FPV18], as well as differentially private learning [DMN+06; KLN+11]. Evolutionary algorithms can be recast as SQ algorithms and in fact were shown to be equivalent to "correlational statistical query" (CSQ) algorithms [Val09; Fel08]. A particularly nice feature of the SQ framework is that it allows for unconditional lower bounds. As such, SQ lower bounds are often taken as evidence for computational hardness if the underlying problem does not admit a linear structure, as is the case for parities. Additionally, many statistical query lower bounds asymptotically match complexity theoretic upper bounds, such as learning DNFs [BFJ+94], learning mixtures of Gaussians [DKS17] and the planted clique problem [FGR+17]. Other results via the SQ framework contain positive results for k-means clustering, principle component analysis and the perceptron algorithm [BDM+05] and manifold estimation [AK21], negative results for learning simple neural networks [CLL22], average case hardness for learning neural networks at super-logarithmic depth [AAK21], as well as lower and upper bounds for optimization and distributional search problems [FGV21; FGR+17; FPV18].

The SQ framework for learning classical objects such as probability distributions as covered in this work has recently been generalized to the quantum statistical query (QSQ) framework for learning quantum objects such as quantum states or processes [AGY20; AHS23; Nie23; WD25]. Recently [Nie23] introduced the evaluation query oracle as an abstraction to unify learning models, which share the key properties of statistical queries, including QSQs, CSQs and parametrized learning algorithms.

Complexity of quantum circuits and computational learning theory: A circuit class is a collection of quantum circuits. A well-established way to assess the complexity of such a circuit class from the viewpoint of classical computing is to study the resources required to classically simulate the circuit class. Examples of circuit classes that have been studied in this regard include Clifford circuits, Clifford+T circuits, matchgate circuits and IQP circuits [Got98; AG04; Val12; TD02; BJS11; BMS17]. Note that in all these examples the description of the circuit class includes a specification of the input state and a fixed measurement basis as the simulatability may crucially depend on these choices. In this work, we analogously assess the complexity of a quantum circuit class from the viewpoint of computational learning theory. Indeed, a long line of research has been aimed at characterizing the complexity of learning *classical* circuit classes, in various learning models [LMN93; Kha93; KMR+94; AGS21]. This complexity is indeed a fundamental property of such circuit classes. Our work provides insight into this fundamental property of

quantum circuits. We stress here that, analogous to the case of classical simulation, we consider learning with respect to a fixed input state and fixed measurement basis. In that sense, our learning model differs from those employed in other works on learning quantum states [Aar07; Mon17] as we require learning the action of the circuit only with respects to measurements in a fixed product basis. This is the key difference between the learning task considered in our work and recent work on learning shallow quantum circuits or their output states [HLB+24; ZLK+24; LL25], respectively.

Heuristic algorithms for distribution learning: Recent years have seen major advances in the development of heuristic neural-network based methods for probabilistic modelling. Generative adversarial networks [GPM+14] and generative transformer models [BMR+20] have managed to achieve impressive results ranging from predicting protein structure to atomic accuracy [JEP+21] to achieving human-level language comprehension [HBM+22]. Learning the underlying model classes is known to be worst case hard [CLL22]. This inspired a series of works in order to better understand these successes from a theoretical perspective (see for example [ABG+14; LSS14; CHM+15; JSA16; Dan17; Sha18; AS23; DV20; AAK21]). Our results can be seen in a similar light for QCBM based algorithms. In particular, our average case hardness results for learning super logarithmic depth quantum circuit distributions is the QCBM equivalent to [AAK21, Contribution 3].

"Far from uniform" property and quantum advantage: In [AC17], the authors propose *heavy output generation* as a particularly natural task for separating classical and quantum computers. More precisely, quantum computers can be used to output sets of bit strings $z_1, \ldots, z_k$ such that more than 2/3 of them have probability larger than the median of the output probabilities. This can be achieved by a subroutine, which uses a conditional probability distribution that samples instances of random quantum circuits as long as necessary for a sufficiently non-uniform probability distribution to appear.

A key result is that far-from-uniform output distributions are not rare for random quantum circuits. We adapted the resulting bound as Theorem 16. Aaronson and Chen moreover conjectured that this far from uniform conjecture not only holds with constant probability but with probability exponentially close to one. This would imply that it suffices to directly sample from a random quantum circuit instance without the subroutine. While we prove that a far from uniform property holds with exponential probability in Theorem 9, the constants in our result do not imply Conjecture 1 in [AC17]. At the same time, it is possible to define a weaker version of heavy output generation, for which our bounds suffice.

From average-case complexity in learning to cryptography: There is a rich correspondence between computational learning theory and cryptography. On the one hand, cryptographic assumptions are often used to prove conditional lower bounds for learning problems [KV94]. On the other hand, the assumed hardness of learning problems can sometimes be used for the construction of cryptographic primitives. For this latter direction, it is well known that the existence of cryptographic primitives such as one-way functions require the existence of learning problems which are

average-case hard [IL90; Bar17]. As concrete examples, Blum, Furst, Kearns and Lipton [BFK+94] have shown that an efficient average case learner for polynomial size circuits in the distribution specific PAC model, would imply the non-existence of one-way functions. This result has been recently extended by Nanashima [Nan21] to give a characterization of auxiliary-input one way functions, based on the hardness of PAC learning polynomial size circuits in a modified average-case variant of PAC learning. In light of these results, it is natural to ask whether one can characterize either classical or quantum cryptographic primitives in terms of the complexity of average-case learners for *quantum* circuit output distributions in the distribution learning setting. While we do not address this question in this work, and while our restriction to the SQ model is a major restriction in this regard, we believe that the connection to cryptography merits further study and hope that the insights gained from our results can be helpful in this regard.

1.4 Proof overview

For ease of presentation, we use the notation $P[\phi]$ to denote $\mathbf{E}_{x\sim P}[\phi(x)]$ and denote by $\mathcal{U}$ the uniform distribution. The starting point of all our results is a lower bound on the average case query complexity in terms of properties of the measure μ. Suppose there is a deterministic algorithm $\mathcal{A}$ that ϵ-learns a β fraction of $\mathcal{D}$ with respect to μ from q many τ-accurate statistical queries. Then, it holds (c.f. Theorem 1)

$$q+1 \geq \frac{\beta - \mathbf{Pr}_{P\sim\mu}\left[\mathrm{d_{TV}}(P,\mathcal{U}) \leq \epsilon+\tau\right]}{\max_{\phi} \mathbf{Pr}_{P\sim\mu}\left[|P[\phi]-\mathcal{U}[\phi]| > \tau\right]}\,, \tag{3}$$

where the max is over all bounded functions $|\phi(x)| \leq 1$.

The above bound is obtained by first reducing a suitable worst-case uniformity test to the average-case learning problem, and then lower bounding the complexity of the uniformity test. Given some set of distributions $\widetilde{\mathcal{D}} \subseteq \mathcal{D}$, we consider the decision problem of testing $\widetilde{\mathcal{D}}$ versus $\mathcal{U}$ defined via:

Given statistical query access to some $P \in \widetilde{\mathcal{D}} \cup \{\mathcal{U}\}$ decide whether "$P = \mathcal{U}$" or "$P \in \widetilde{\mathcal{D}}$".

We then note:

1. An *average-case* learner for $\mathcal{D}$ which succeeds on a β fraction of instances with respect to μ, implies the existence of a *worst-case* learning algorithm for some $\mathcal{D}' \subseteq \mathcal{D}$ with $\mu(\mathcal{D}') = \beta$. This trivially implies a worst-case learning algorithm for $\widetilde{\mathcal{D}} = \mathcal{D}'/B_{\epsilon+\tau}(\mathcal{U})$.
2. A *worst-case* learning algorithm for $\widetilde{\mathcal{D}}$ using at most q queries implies an algorithm for deciding $\widetilde{\mathcal{D}}$ versus $\mathcal{U}$ using $q+1$ queries.

As such, it is sufficient for us to lower bound the complexity the uniformity test. To this end, we use a counting argument to show that for any measure ν over $\widetilde{\mathcal{D}}$ it

holds that the number of queries necessary to decide $\widetilde{\mathcal{D}}$ versus $\mathcal{U}$ satisfies

$$q \geq \left(\max_{\phi} \mathop{\mathbf{Pr}}_{P \sim \nu} \left[|P[\phi] - \mathcal{U}[\phi]| > \tau \right] \right)^{-1} . \tag{4}$$

We then obtain Equation (3) by considering the measure ν defined by conditioning μ on $\widetilde{\mathcal{D}}$.

Given this, from Equation (3) it is clear that in our context, in order to obtain the desired lower bounds we require:

- An upper bound on the maximal fraction distinguishable from uniform $\mathfrak{f} = \max_{\phi} \mathbf{Pr}_U[|P_U[\phi_i] - \mathcal{U}[\phi_i]| \geq \tau]$.
- An upper bound on the mass of the ϵ-ball around the reference distribution. The complement of this probability is often referred to as the probability of being *far from uniform*.

In this work we give bounds on both quantities for random quantum circuits of various depths. We begin in the limit of infinitely deep circuits, and thus of Haar-random unitaries and then partially derandomize our results using concentration inequalities based on higher moments of the Haar measure. This will allow us to prove results about random quantum circuits of comparably low depth, which are far from the Haar measure but quickly generate the same moments.

1.4.1 Bounding $\mathfrak{f}$

For Haar random unitaries, the concentration of measure phenomenon in the form of Levy's lemma produces tight bounds on tails, which implies bounds on $\mathfrak{f}$.

For linear depth circuits we can use Chebyshev's inequality in order to bound $\mathfrak{f}$

$$\max_{\phi} \mathop{\mathbf{Pr}}_{U} [|P_U[\phi_i] - \mathcal{U}[\phi_i]| \geq \tau] \leq \frac{\mathbf{Var}_U \left[P_U[\phi] \right]}{\tau^2} \tag{5}$$

where we have used $\mathbf{E}_U \left[P_U[\phi] \right] = \mathcal{U}[\phi]$ since for any depth $d \geq 1$ μ is a 1-design. The variance is a second moment and can be exactly computed for the Haar measure via standard symmetry arguments or from the Weingarten calculus.

For sublinear circuits, the distribution over unitaries does not form a unitary design, however, we can still bound the second moments involved in the variance by adapting the statistical physics mapping from [Hun19] in a similar way as in [BCG21]. Many of our bounds will depend on such moment bounds over the Haar measure that we detail in Section C.

1.4.2 Bounding the far from uniform probability

To obtain a lower bound on the far from uniform probability (or equivalently an upper bound on the probability mass of an ϵ-ball around the uniform distribution), we start by writing the total variation distance in terms of the ℓ_1-norm.

In the infinite circuit depth regime we make use of Gaussian integration in order to obtain upper and lower bounds on the expected distance between a random output distribution P_U and the uniform distribution. Then we use Levy's Lemma once again to obtain a bound on the probability of P_U being ϵ far from the uniform distribution.

In the linear depth regime we apply a variant of Berger's inequality for ℓ_p-norms, c.f. Theorem 8, in order to obtain

$$\mathop{\mathbf{Pr}}_{U}\left[\|P_U - \mathcal{U}\|_1 < 2\epsilon\right] \leq \mathop{\mathbf{Pr}}_{U}\left[\frac{\|P_U - \mathcal{U}\|_2^3}{\|P_U - \mathcal{U}\|_4^2} < 2\epsilon\right]. \tag{6}$$

Treating the numerator and denominator separately we can thus apply the union bound and obtain

$$\mathop{\mathbf{Pr}}_{U}\left[\|P_U - \mathcal{U}\|_1 < 2\epsilon\right] \leq \mathop{\mathbf{Pr}}_{U}\left[\|P_U - \mathcal{U}\|_2^3 < 2\epsilon_1\right] + \mathop{\mathbf{Pr}}_{U}\left[\|P_U - \mathcal{U}\|_4^2 > 2\epsilon_2\right], \tag{7}$$

for suitable $\frac{\epsilon_1}{\epsilon_2} \geq \epsilon$. Both terms can be bounded separately with the same strategy. Using Chebyshev's inequality for the random variable $X(p,q) = \|P_U - \mathcal{U}\|_p^q$ we can bound the deviation of $X(p,q)$ from its mean $\mathbf{E}[X(p,q)]$. In particular, due to the variance term in Chebyshev's inequality we find that an (approximate) $2p$-design is sufficient for an exponential concentration of $X(p,q)$. Since $\mathbf{E}[X(2,3)]$ is exponentially small it is thus crucial, that $X(4,2)$ itself is sufficiently small and sharply concentrated such that we can find ϵ_1 and ϵ_2 that cancel to a constant. Again, we confirm this ingredient via a variance bound provided that our measure is induced by an 8-design. We conclude from [BHH16; Haf22] that random linear depth Born distributions are far from uniform.

In the sublinear depth regime we use a result by Aaronson and Chen [AC17] which lower bounds the expected distance between a randomly drawn P_U and the uniform distribution even for $d = 1$. We then use Markov's inequality to translate this to a bound the probability of P_U being ϵ far from the uniform distribution.

1.5 Discussion and future work

In this work we give lower bounds for the average case query complexity of learning the output distributions of random quantum circuits in different depth regimes. In particular, we show that the problem of learning the output distribution of random quantum circuits is hard with constant probability over the instance already at super logarithmic depth. Moreover, we prove that the problem becomes hard with probability exponentially close to one over the instance at linear depth. Our analysis is accompanied by the corresponding results for both Haar random unitary output distributions and Haar random Clifford output distributions. While the former gives hardness with probability doubly exponentially close to 1, the latter only gives hardness with a constant probability over the instance.

There are multiple natural avenues to continue this work:

1. While we prove a strong asymptotic average-case complexity bound for linear

depth circuits, the explicit depth at which our bounds apply comes with a rather large prefactor ($d = 10^{20}n$). This prefactor is likely an artefact of the proof techniques in [BHH16; Haf22], which bound the depth at which unitary designs are well-approximated. There are at least two promising approaches to lowering this constant and consequently making our bound more directly applicable to a practical regime. One could directly compute the moments using the statistical physics mapping from [Hun19]. Alternatively, extensive numerical calculations of finite-size spectral gaps in combination with Knabe bounds might lower the explicit depth at which unitary designs are generated [HH21]. While these calculations are beyond the scope of this paper, we believe that it would be worthwhile to address this issue.

2. Moreover, when relaxing the assumption on ϵ from constant to inverse polynomial and on β from inverse exponential to inverse polynomial, thus making the learning task harder, one can use Markov's inequality instead of Chebyshev's in the far from uniform proof (c.f. Theorem 10). This already improves the asymptotics to hold at $d = 690n$. This might simplify the corresponding statistical physics mapping since it makes use of fourth instead of eighth moments. We expect that such a calculation implies average-case hardness with probability $1 - o(1)$ over the instances for any depth $d = \omega(\log(n))$.

3. In this work, we rule out efficient algorithms at super logarithmic depth. It is thus natural to ask what happens at logarithmic depth and below. For example, are the output distributions of constant depth quantum circuits efficient to learn?

4. Last, we emphasize that all our results hold in the statistical query model. Another prominent model is learning from samples. With the famous exception of parities, most known tight upper bounds for learning can be realized in the statistical query model. We ask whether our average-case complexity results carry over to average-case hardness of learning from samples.

2 Notation and preliminaries

2.1 Statistical query learning

Let $\mathcal{D}$ be a class of distributions over a domain X. For two distributions $P, Q \in \mathcal{D}$ we denote by $\mathrm{d_{TV}}(P, Q) := \frac{1}{2}\sum_{x\in X} |P(x) - Q(x)|$ the total variation distance between them. The open ϵ-ball $B_\epsilon(P)$ around any distribution P over the domain X is given by the set of all distributions Q over X such that $\mathrm{d_{TV}}(P, Q) < \epsilon$. For a distribution P over X and a function $\phi : X \to [-1, 1]$ we use the short hand notation

$$P[\phi] := \mathop{\mathbf{E}}_{x\sim P}[\phi(x)] \tag{8}$$

to refer to the expectation value of ϕ with respect to P. We denote by $\mathcal{D}_X$ the set of distributions over X and by $\mathcal{D}_n$ the set of all distributions over the domain $\{0, 1\}^n$. The uniform distribution is denoted by $\mathcal{U}$.

A well studied model in learning theory is the statistical query learning model. Here we assume that the learner has access to expectation values of functions with respect to the underlying probability distribution. This can be formalized by considering access to a statistical query oracle.

Definition 1: (Statistical query oracle) For $\tau > 0$ and a distribution P over X we denote by $\texttt{Stat}_\tau(P)$ the statistical query (SQ) oracle of P with tolerance τ. When queried with some function $\phi : X \to [-1, 1]$ the oracle returns some v such that $|v - P[\phi]| \leq \tau$.

Remark 1: On immediate consequence of Definition 1 which is useful for the interpretation of our formal result statements is as follows: For any $\zeta < \tau$, any SQ oracle $\texttt{Stat}_\zeta(P)$ is also a valid SQ oracle $\texttt{Stat}_\tau(P)$. Thus, lower bounds for SQ algorithms with respect to ζ trivially imply the same lower bound for SQ algorithms with respect to τ.

A prominent special case is to consider τ to be lower bounded by an inverse polynomial, since this reflects the scenario where the statistical query oracle can be run efficiently with a polynomial number of samples. Up to polynomial corrections, the complexity of the oracle is then given by the complexity of computing the query function ϕ.

In order to learn a distribution it is crucial to fix the representation of the distribution to be learned. Here, a representation can be thought of as an algorithm that specifies the distribution. For the sake of clarity typical representations are generators and evaluators.

- A generator of a probability distribution P is a probabilistic algorithm that produces samples according to $x \sim P$.
- An evaluator of a probability distribution $P \in \mathcal{D}_n$ is an algorithm $\texttt{Eval}_P : \{0, 1\}^n \to [0, 1]$ that outputs the probability amplitude $\texttt{Eval}_P[x] = P(x)$.

We will refer to a representation (e.g. a generator or an evaluator) of a distribution Q as an ϵ-approximate representation of P if $\mathrm{d}_{\mathrm{TV}}(P, Q) < \epsilon$. We would like to stress that, due to fundamental limitations, our results hold for any ϵ-approximate representation even if we allow the corresponding algorithms to be computationally inefficient.

Distribution learning in the statistical query framework is made formal by the following definition.

Problem 1: (ϵ-learning of $\mathcal{D}$ from statistical queries) Let $\epsilon \in (0, 1)$ be an accuracy parameter, $\tau \in (0, 1)$ be the tolerance and let $\mathcal{D}$ be a distribution class. For a fixed representation of the distribution, the task of ϵ-learning $\mathcal{D}$ from statistical queries with tolerance τ is defined as given access to $\texttt{Stat}_\tau$ for any unknown $P \in \mathcal{D}$, output an ϵ-representation of P.

In [HIN+23] the authors have studied the worst case query complexity of Problem 1 for $\mathcal{D}$ being the class of output distributions of local quantum circuits. Here we want to consider the average case query complexity for the same distribution class. This is a strictly easier task as we do not require the learner to succeed for each and every distribution in the class. Rather, we want to characterize the number of statistical queries needed to succeed in solving Problem 1 on a fraction of distributions in $\mathcal{D}$ with respect to a measure μ over the distributions in $\mathcal{D}$. Similarly, when one considers a quantum or probabilistic learner one can ask about the number of statistical queries needed to succeed in the aforementioned task with some fixed probability with respect to the algorithm's randomness. Thus, the randomized average case complexity is defined with respect to a measure μ and the two parameters α and β. By α we denote the success probability of the algorithm and by β the size, with respect to μ, of the fraction on which the algorithm is successful.

Definition 2: (Average case complexity) Let $\mathcal{D}$ be a class of distributions, μ a probability measure over $\mathcal{D}$ and $\alpha, \beta \in (0,1)$. The deterministic average case query complexity of Problem 1 is defined as the minimal number q of queries any learning algorithm $\mathcal{A}$ must make in order to achieve

$$\mathop{\mathbf{Pr}}_{P\sim\mu}\left[\text{“}\mathcal{A}^{\mathtt{Stat}(P)}\ \epsilon\text{-learns } P \text{ from } q \text{ queries”}\right] \geq \beta\,. \tag{9}$$

Likewise, the randomized average case query complexity is defined as the minimal number q of queries any random learning algorithm $\mathcal{A}$ must make in order to achieve

$$\mathop{\mathbf{Pr}}_{P\sim\mu}\left[\mathop{\mathbf{Pr}}_{\mathcal{A}}\left[\text{“}\mathcal{A}^{\mathtt{Stat}(P)}\ \epsilon\text{-learns } P \text{ from } q \text{ queries”}\right] \geq \beta\right] \geq \alpha\,, \tag{10}$$

where $\mathbf{Pr}_{\mathcal{A}}$ denotes the probability over the internal randomness of $\mathcal{A}$.

The deterministic and randomized average case complexities in the framework of statistical query learning are closely related. As advertised in Section 1.1, one can directly translate our lower bounds for deterministic learning to lower bounds of randomized learning by means of a global prefactor $2(\alpha - 1/2)$. Thus, for the sake of ease we focus on the deterministic average case complexity throughout the main text and refer to Section F for the details about random algorithms.

As discussed in Section 1.4 the average case query complexity of deterministic algorithms for learning in the SQ framework can now be lower bounded as follows.

Lemma 1: (Deterministic average case complexity) *Suppose there is a deterministic algorithm $\mathcal{A}$ that ϵ-learns a β fraction of $\mathcal{D}$ with respect to μ from q many τ-accurate statistical queries. Then for any Q it holds*

$$q + 1 \geq \frac{\beta - \mathbf{Pr}_{P\sim\mu}\left[\mathrm{d_{TV}}(P,Q) \leq \epsilon + \tau\right]}{\max_\phi \mathbf{Pr}_{P\sim\mu}\left[|P[\phi] - Q[\phi]| > \tau\right]}\,, \tag{11}$$

where again, the max *is over all functions $\phi : X \to [-1,1]$.*

We refer to Section E for the proof of Theorem 1. To provide a simplified expression we make the following remark.

Remark 2: Note that without loss of generality we can take $\tau \leq \epsilon$ which leads to the bound

$$q + 1 \geq \frac{\beta - \mathbf{Pr}_{P\sim\mu}\left[\mathrm{d_{TV}}(P,Q) \leq 2\epsilon\right]}{\max_\phi \mathbf{Pr}_{P\sim\mu}\left[|P[\phi] - Q[\phi]| > \tau\right]}. \tag{12}$$

To see why we can do so, consider instead the case $\tau > \epsilon$. Given P, Q such that $\tau > \mathrm{d_{TV}}(P,Q) > \epsilon$ we can see that these distributions are indistinguishable with respect to τ-accurate queries and thus there cannot exist an ϵ-learner.

From Theorem 1 it is clear that a crucial figure of merit is the fraction of distributions that can be distinguished from a single query. Following [Fel17] we define.

Definition 3: (Maximally distinguishable fraction) Let $\mathcal{D}$ be a distribution class over the domain X and let μ be some probability measure over $\mathcal{D}$. The maximally distinguishable fraction with tolerance parameter τ and with respect to the measure μ and the reference distribution Q is defined as

$$\mathrm{frac}(\mu, Q, \tau) := \max_\phi \mathop{\mathbf{Pr}}_{P\sim\mu}\left[|P[\phi] - Q[\phi]| > \tau\right], \tag{13}$$

where the maximum is over all functions $\phi : X \to [-1, 1]$.

In the special case that the reference distribution is the uniform distribution, as is the case in the remainder of this paper, $\mathcal{U}$ we will refer to this by the short hand

$$\mathfrak{f} = \mathrm{frac}(\mu, \mathcal{U}, \tau), \tag{14}$$

where the measure μ and the tolerance τ will be clear from context.

Summary: A lower bound for average case query complexity for learning in the statistical query model is determined by:

- The size of the $\epsilon+\tau$-ball of any fixed reference distribution $\mathbf{Pr}_{P\sim\mu}[\mathrm{d_{TV}}(P,Q) \leq \epsilon + \tau]$.
- The maximally distinguishable fraction with respect to the same reference distribution $\mathrm{frac}(\mu, Q, \tau)$.

Note 1: For the sake of ease of presentation we give the derivation of bounds on the weights of the ball around the reference distribution in terms of ϵ. We then translate the corresponding result to Theorem 1 substituting ϵ by $\epsilon + \tau$.

2.2 Random quantum circuits

Given some n-qubit unitary $U \in \mathrm{U}(2^n)$, we denote by $P_U(x) = |\langle x|U|0^n\rangle|^2$ the quantum circuit output, or Born distribution. We denote by μ_U the unitary Haar

measure, or simply the uniform measure, over $\mathrm{U}(D)$. Similarly, we denote by μ_S the spherical Haar measure, or likewise the uniform measure, over the complex unit sphere $\mathbb{S}^{D-1}$, where in both cases the dimensionality D will be clear from context.

Definition 4: (Brickwork architecture) An n-qubit brickwork quantum circuit of depth d (with periodic boundary conditions) is a quantum circuit that is of the form

$$\begin{aligned} U = (U_{2,3}^{(d)} \otimes \cdots \otimes U_{n,1}^{(d)}) \cdot (U_{1,2}^{(d-1)} \otimes \cdots \otimes U_{n-1,n}^{(d-1)}) \cdots \\ \cdots (U_{2,3}^{(2)} \otimes \cdots \otimes U_{n,1}^{(2)}) \cdot (U_{1,2}^{(1)} \otimes \cdots \otimes U_{n-1,n}^{(1)}) \end{aligned} \tag{15}$$

where $U_{i,j}^{(k)} \in \mathrm{U}(4)$ is the unitary in the k'th layer acting on neighboring qubits i and j. For the sake of ease we have assumed d and n to be even.

While we give definitions and analysis only for periodic boundary conditions, we note that all our results will carry over to open boundary conditions at the price of slightly worse prefactors.

Definition 5: (Random quantum circuits) A random brickwork quantum circuit of depth d on n qubits is formed by drawing $\lfloor n/2 \rfloor \cdot d$ many 2-qubit unitaries $U_{i,j}^{(k)}$ i.i.d. Haar randomly and contracting them. We denote the resulting probability distribution on $\mathrm{U}(2^n)$ by μ_C, where n and d will be clear from context.

Given a two-qubit gate set $\mathcal{G} \subseteq \mathrm{U}(4)$. We denote by $\mathcal{D}_{\mathcal{G}}(n, d)$ the set of Born distributions which can be realized by brickwork quantum circuits on n qubits of depth d.

2.3 Unitary designs

Unitaries generated by random quantum circuits quickly mimic Haar random unitaries for many practical purposes. The reason for this is that they generate unitary t-designs. These are "evenly" spread probability distributions over the unitary group that have the same t'th moments as the Haar measure [Dan05; GAE07]. This is often expressed in terms of t-fold twirls: Let ν be a probability measure on the unitary group $\mathrm{U}(D)$. Then we define for any matrix $A \in \mathbb{C}^{D^t \times D^t}$.

$$\Phi^{(t)}(\nu)(A) := \int U^{\otimes t} A (U^\dagger)^{\otimes t} \mathrm{d}\nu(U)\,. \tag{16}$$

We call ν an approximate unitary t-design if, for μ_U being the Haar measure.

$$\Phi^{(t)}(\nu) \approx \Phi^{(t)}(\mu_U)\,. \tag{17}$$

We provide a detailed definition in Appendix B. Moreover, see Appendix B for the relation to state designs and bounds on the generation of approximate designs by random quantum circuits.

3 Haar random unitaries

In this section, we bound the two key quantities, namely $\mathfrak{f}$ and the far from uniform probability, for random quantum circuits of infinite depth, corresponding to Haar random unitaries. Plugging these bounds into Theorem 1, we obtain the following lower bound on the average case query complexity.

Theorem 2: (Formal version of infinite depth part of Informal Theorem 1) *Let $\tau > 0$, $\epsilon \leq 1/e - 2^{-n/2-1} - \tau$ and set $\xi = 1/e - 2^{-n/2-1} - \epsilon - \tau$. Any algorithm that succeeds in ϵ-learning a β fraction of the output distributions of infinitely deep random brickwork quantum circuits requires q many τ-accurate statistical queries, with*

$$q + 1 \geq \frac{\beta - 2\exp\left(-\frac{2^{n+2}\xi^2}{9\pi^3}\right)}{2\exp\left(-\frac{2^n\tau^2}{9\pi^3}\right)}\,. \tag{18}$$

Remark 3: Note that, for any

$$\tau \geq 2^{-n/4} \quad \text{and any} \quad \epsilon \leq \frac{1}{e} - 2^{-n/2-1} - 2^{-n/4+2} - \tau$$

which corresponds to $\xi \geq 2^{-n/4+2}$, we find by Theorem 2 the query complexity for learning any fraction $\beta > 2\exp\left(-2^{n/2+4}/9\pi^3\right) = 2^{-2^{\Omega(n)}}$ requires $q = 2^{2^{\Omega(n)}}$ many queries. In words: learning a doubly exponentially small fraction takes doubly exponentially many, inverse exponentially accurate statistical queries.

In the case of infinitely deep circuits, it is known that the distribution of circuit unitaries converges to the unique, rotationally invariant Haar measure on the full unitary group in $D = 2^n$ dimensions μ_U. If we apply such a Haar-random unitary to a designated (arbitrary) pure starting state, say $|\psi_0\rangle = |0, \ldots, 0\rangle$, we obtain a pure state $|\psi\rangle$ that is sampled from the spherical Haar measure μ_S, i.e. uniformly from the set of all pure states in D dimensions:

$$|\psi\rangle = U|\psi_0\rangle \overset{unif}{\sim} \left\{|u\rangle \in \mathbb{C}^D : \ \langle u|u\rangle = 1\right\} \subset \mathbb{C}^D \quad \text{(and } D = 2^n \text{ for } n \text{ qubits)}.$$

Such (Haar) uniform distributions of pure states have two remarkable features: (i) we can use powerful frameworks like Weingarten calculus and Gaussian integration to compute complicated expectation values and (ii) concentration of measure (Levy's lemma) asserts that concrete realizations concentrate very sharply around this expected behavior. These two features can be combined into a powerful strategy to obtain very sharp bounds on deviation probabilities, like the two essential ingredients in Theorem 1:

$$\underset{U\sim\mu_U}{\mathbf{Pr}}\left[\mathrm{d_{TV}}(P_U, \mathcal{U}) \geq \epsilon\right] \quad \text{and} \quad \underset{U\sim\mu_U}{\mathbf{Pr}}\left[|P_U[\phi] - \mathcal{U}[\phi]| > \tau\right]$$

for any fixed function $\phi : \{0,1\}^n \to [-1,1]$. To apply this formalism, our strategy will be to first reformulate the arguments in both probabilities as functions in the

(Haar) random pure state $|\psi\rangle = U|\psi_0\rangle$. Then we will use Levy's lemma to obtain bounds on both probabilities. Let's focus on this last step as it applies to both probabilities.

Levy's lemma asserts that every reasonably well-behaved function concentrates very sharply around its expectation value if we choose a random vector uniformly from a (real- or complex-valued) unit sphere $\mathbb{S}^{D-1}$ in $D \gg 1$ dimensions. Note that Haar-random state $|\psi\rangle = U|\psi_0\rangle$ with $U \sim \mu_U$ meet this sampling requirement by definition. The well-behavedness of functions is measured by their Lipschitz constant. A function $f : \mathbb{S}^{D-1} \to \mathbb{R}$ is Lipschitz with respect to the ℓ_2-norm in $\mathbb{C}^D$ with constant $L \geq 0$ if

$$|f(|\psi\rangle) - f(|\phi\rangle)| \leq L \, \||\psi\rangle - |\phi\rangle\|_2 \quad \text{for all } |\psi\rangle, |\phi\rangle \in \mathbb{S}^{D-1}.$$

Here is a variant of Levi's Lemma (concentration of measure) that directly applies to pure quantum states in D dimensions. It readily follows from identifying the complex unit sphere $\mathbb{S}^{D-1} \subset \mathbb{C}^D$ with a real-valued unit sphere in $2D$ dimensions isometric embedding, see, e.g. [BCH+21, proof of Proposition 29].

Theorem 3: (Levy's lemma for Haar-random pure states) *Let $f : \mathbb{S}^{D-1} \to \mathbb{R}$ be a function from D-dimensional pure states to the real numbers that is Lipschitz with Lipschitz constant L. Then,*

$$\Pr_{|\psi\rangle \sim \mu_S} \left[\left| f(|\psi\rangle) - \mathop{\mathbf{E}}_{|\psi\rangle \sim \mu_S} [f(|\psi\rangle)] \right| > \tau \right] \leq 2 \exp\left(-\frac{4D\tau^2}{9\pi^3 L^2} \right) \quad \textit{for any } \tau > 0.$$

In words, Levy's lemma suppresses the probability of a τ-deviation from the expectation value. This bound diminishes exponentially in Hilbert space dimension $D = 2^n$, i.e. doubly exponentially in qubit size n. In the following two sections we will use theorem 3 to obtain bounds on the maximally distinguishable fraction and the probability of being far from uniform. We do so by explicitly upper bounding the Lipschitz constants of the involved functions and computing the Haar expectation values.

3.1 Maximally distinguishable fraction

Let us start with $\mathbf{Pr}_{U \sim \mu_U}[|P_U[\phi] - \mathcal{U}[\phi]| > \tau]$. We will use the short-hand notation $x = (x_1, \ldots, x_n) \in \{0,1\}^n$ to enumerate all possible outcome strings of n parallel computational basis measurements:

$$P_U[\phi] = \sum_{x \in \{0,1\}^n} \phi(x_0, \ldots, x_n) \, |\langle x|\psi\rangle|^2 = \langle\psi| \left(\sum_{x \in \{0,1\}^n} \phi(x) \, |x\rangle\langle x| \right) |\psi\rangle = \langle\psi|\Phi|\psi\rangle,$$

where we have introduced the diagonal matrix $\Phi = \sum_x \phi(x)|x\rangle\langle x| \in \mathbb{C}^{D \times D}$ whose spectral norm obeys $\|\Phi\|_\infty = \max_{x \in \{0,1\}^n} |\phi(x)| \leq 1$ regardless of the underlying function in question. This is a very simple and highly structured quadratic form in $|\psi\rangle = U|\psi_0\rangle$. Its expectation value over all Haar-random states produces the

uniform distribution $\mathcal{U}[\phi]$, in formulas

$$\mathop{\mathbf{E}}_{U\sim\mu_U}[P_U[\phi]] = \sum_{x\in\{0,1\}^n} \frac{1}{2^n}\phi(x) = \mathcal{U}[\phi]\,. \tag{19}$$

To see this, note that the uniform average over all pure states in D dimensions produces the maximally mixed state, i.e. $\mathbf{E}_{|\psi\rangle\sim\mu_S}[|\psi\rangle\langle\psi|] = \mathbb{1}/D$. Linearity of the expectation value then ensures

$$\begin{aligned}\mathop{\mathbf{E}}_{U\sim\mu_U}[P_U[\phi]] &= \mathop{\mathbf{E}}_{|\psi\rangle\sim\mu_S}[\langle\psi|\Phi|\psi\rangle] = \mathrm{Tr}\left(\mathop{\mathbf{E}}_{|\psi\rangle\sim\mu_S}[|\psi\rangle\langle\psi|]\ \Phi\right) = \mathrm{tr}\left(\mathbb{1}/D\ \Phi\right) = \frac{1}{D}\,\mathrm{tr}\left(\Phi\right)\\ &= \sum_x \frac{1}{2^n}\phi(x) = \mathcal{U}[\phi],\end{aligned}$$

as claimed. We also provide an alternative derivation based on Gaussian integration in the Appendix (Theorem 18). We can now obtain an explicit bound on the maximally distinguishable fraction by theorem 3.

Theorem 4: *Consider n-qubit Haar-random unitaries $U\sim\mu_U$ ($D=2^n$) and fix a function $\phi:\{0,1\}^n\to[-1,1]$. Then,*

$$\mathop{\mathbf{Pr}}_{U\sim\mu_U}[|P_U[\phi]-\mathcal{U}[\phi]|>\tau] \le 2\exp\left(-\frac{2^n\tau^2}{9\pi^3}\right) \quad \textit{for any } \tau>0.$$

In words: for each fixed function ϕ, its evaluation $P_U[\phi]$ concentrates very sharply doubly-exponentially in qubit size n around the uniform average. Hence, it is extremely unlikely to distinguish P_U from $\mathcal{U}$ with only a single ϕ.

Proof. Rewrite $P_U[\phi] = \langle\psi|\Phi|\psi\rangle =: f_\Phi(|\psi\rangle)$ with $|\psi\rangle = U|\psi_0\rangle$ (Haar random state) and diagonal matrix Φ that obeys $\|\Phi\|_\infty \le 1$. This reformulation highlights that the quadratic form function $f_\Phi:\mathbb{S}^{D-1}\to\mathbb{R}$ is Lipschitz with constant $L\le 2\|\Phi\|_\infty\le 2$, we refer to Theorem 22 in the appendix for a detailed statement and proof. Next, recall from Equation (19) that $\mathbf{E}_{|\psi\rangle\sim\mu_S}[f_\Phi(|\psi\rangle)] = \mathbf{E}_{U\sim\mu_U}[P_U[\phi]] = \mathcal{U}[\phi]$ and apply Theorem 3 to deduce the claim. □

3.2 Far from uniform probability

Moving now to the quantity $\mathbf{Pr}_{U\sim\mu_U}[\mathrm{d_{TV}}(P_U,\mathcal{U})\ge\epsilon]$, we note that the expectation value required is a bit more involved by comparison. Let us start by reformulating the total variation distance between P_U and $\mathcal{U}$ as

$$\begin{aligned}\mathrm{d_{TV}}(P_U,\mathcal{U}) = \frac{1}{2}\sum_{x\in\{0,1\}^n}|P_U(x)-\mathcal{U}(x)| &= \frac{1}{2}\sum_{x\in\{0,1\}^n}\left||\langle x\mid U|\psi_0\rangle|^2-\frac{1}{D}\right|\\ &= \frac{1}{2D}\sum_{x\in\{0,1\}^n}\left||D\langle x\mid\psi\rangle|^2-1\right|,\end{aligned}$$

where $D = 2^n$ and $|\psi\rangle = U|\psi_0\rangle$. Linearity of the expectation value and unitary invariance of the spherical Haar measure then implies

$$\begin{aligned}\mathop{\mathbf{E}}_{U\sim\mu_U}[\mathrm{d_{TV}}(P_U,\mathcal{U})] &= \frac{1}{2D}\sum_{x\in\{0,1\}^n}\mathop{\mathbf{E}}_{|\psi\rangle\sim\mu_S}\Big[\Big|D\,|\langle x|\psi\rangle|^2-1\Big|\Big]\\ &= \frac{1}{2}\mathop{\mathbf{E}}_{|\psi\rangle\sim\mu_S}\Big[\Big|D\,|\langle 0,\ldots,0|\psi\rangle|^2-1\Big|\Big].\end{aligned}$$

The expression on the right hand side is not a polynomial in the overlap $|\langle 0,\ldots,0|\psi\rangle|^2$ which prevents us from using Weingarten calculus to compute it. Another averaging technique, known as Gaussian integration, does the job, however. The key idea is to view the uniform expectation over pure states as a uniform integral over all points that are contained in the complex-valued unit sphere in D dimensions. Up to a normalization factor (scaling), this integral can then be re-cast as an expectation value over the directional degrees of freedom in a 2^n-dimensional complex-valued Gaussian random vector with independent entries $g_j + \mathrm{i}h_j$, $1 \leq j \leq 2^n$ and $g_j, h_j \overset{iid}{\sim} \mathcal{N}(0,1)$. A detailed argument is provided in the appendix and yields

$$\frac{1}{\mathrm{e}} - \frac{1}{2^{n/2+1}} \leq \mathop{\mathbf{E}}_{U\sim\mu_U}[\mathrm{d_{TV}}(P_U,\mathcal{U})] \leq \frac{1}{\mathrm{e}} + \frac{1}{2^{n/2+1}}, \tag{20}$$

where e denotes Euler's constant. Note that the approximation errors on the left and right decay exponentially in qubit size n. This is the content of Theorem 19 in the appendix and the proof uses a precise version of the approximate identity

$$\begin{aligned}&\frac{1}{2}\mathop{\mathbf{E}}_{|\psi\rangle\sim\mu_S}\Big[\Big|D|\langle 0,\ldots,0|\psi\rangle|^2-1\Big|\Big]\\ &\quad\approx \frac{1}{4}\mathop{\mathbf{E}}_{g_j,h_j\overset{iid}{\sim}\mathcal{N}(0,1)}\Big[\Big||g_1+\mathrm{i}h_1|^2-2\Big|\Big] = \frac{1}{4}\mathop{\mathbf{E}}_{g_1,h_1\overset{iid}{\sim}\mathcal{N}(0,1)}\Big[\Big|g_1^2+h_1^2-2\Big|\Big]\\ &\quad= \frac{1}{4}\iint_{\infty}^{\infty}\Big|g_1^2+h_1^2-2\Big|\,\frac{\exp\left(-(g_1^2+h_1^2)/2\right)}{2\pi}\mathrm{d}g_1\mathrm{d}g_2 = \frac{1}{\mathrm{e}}.\end{aligned}$$

The final equality follows from switching into polar coordinates and solving the resulting integral analytically – this is why the technique is called Gaussian integration. The exponentially small offsets $\pm 1/2^{n/2+1}$ in Rel. (20) bound the approximation error that is incurred in the first step of this argument. As before, we can now obtain a bound on the probability of a Haar randomly drawn unitary giving rise to a distribution that is far from uniform by use of Theorem 3.

Theorem 5: *Consider n-qubit Haar-random unitaries $U \sim \mu_U$ ($D = 2^n$). Then, the TV distance between P_U and the uniform distribution $\mathcal{U}$ is guaranteed to obey*

$$\mathop{\mathbf{Pr}}_{U\sim\mu_U}\left[\left|\mathrm{d_{TV}}(P_U,\mathcal{U}) - \frac{1}{\mathrm{e}}\right| \geq \xi + \frac{1}{2^{n/2+1}}\right] \leq 2\exp\left(-\frac{2^{n+2}\xi^2}{9\pi^3}\right) \quad \textit{for any } \xi > 0.$$

In words: this TV distance concentrates very sharply (doubly-exponentially in qubit size n) around the remarkable value $1/\mathrm{e} \geq 0.367$. The exponentially small in n addi-

tive correction term $1/2^{n/2+1}$ is a consequence of the slight mismatches in Rel. (20). Note, however, that this does not qualitatively change the concentration statement. The mismatch is of the same order as the smallest ξ for which the exponential tail bound still provides meaningful results: $2^{n+2}\xi^2 \gtrsim 1$ requires $\xi \gtrsim 1/2^{n/2+1}$.

Proof. The proof is conceptually very similar to the proof of Theorem 4. We first recast the expectation over Haar-random unitaries U as an expectation value over Haar-random state $|\psi\rangle = U|\psi_0\rangle$. The function $\mathrm{d_{TV}}(P_U,\mathcal{U})$ in question becomes $g(|\psi\rangle) = (2D)^{-1}\sum_{x\in\{0,1\}^n}\left|D\,|\langle x|\psi\rangle|^2 - 1\right|$ and can be shown to be Lipschitz with constant $L \leq 1$. Again, we refer to Theorem 23 in the appendix for a precise statement and proof. Next, we use Rel. (20) to infer

$$\mathbf{Pr}\left[\left|\mathrm{d_{TV}}(P_U,\mathcal{U}) - \frac{1}{\mathrm{e}}\right| \geq \xi + \frac{1}{2^{n/2+1}}\right] \leq \mathbf{Pr}\left[\left|\mathrm{d_{TV}}(P_U,\mathcal{U}) - \mathop{\mathbf{E}}_{U\sim\mu_U}\left[\mathrm{d_{TV}}(P_U,\mathcal{U})\right]\right| \geq \xi\right]$$

and apply Levy's Lemma (Theorem 3) with Lipschitz constant $L = 1$ to the right hand side of this display. □

4 Random quantum circuits of linear depth

In this section, we bound the two key quantities, namely $\mathfrak{f}$ and the far from uniform probability, for random quantum circuits of linear depth. We will find that the strong convergence of these circuit ensembles to unitary t-designs suffices to show exponentially small upper bounds on both quantities.

Plugging these bounds into Theorem 1, we obtain the following lower bound on the average case query complexity.

Theorem 6: (Formal version of linear depth part of Informal Theorem 1) *Let $\tau > 0$. Further, let the circuit depth be $d \geq 1.2 \times 10^{20} n$, let $\epsilon \leq 1/150 - \tau$ and let n be large enough. Then, the average case query complexity q of ϵ-learning any β-fraction of brickwork random quantum circuit output distributions is lower bounded by*

$$q + 1 \geq \left(\beta - 3200 \times 2^{-n}\right) 2^{n-2}\tau^2\,. \tag{21}$$

Remark 4: Note that for any $\tau \geq 2^{-n/4}$ and, say any $\epsilon \leq 1/160$ by Theorem 6 learning a fraction $\beta > 3200\cdot 2{-}n = O(2^{-n})$ requires at least $q = 2^{\Omega(n)}$ many queries.

Moreover, in the practically inspired regime $1/\mathrm{poly}(n) \leq \tau \leq 1/150 - \epsilon$, we obtain $q = \Omega(2^n)$.

4.1 Maximally distinguishable fraction

We begin with bounding the maximally distinguishable fraction.

Lemma 7: *Let $\delta > 0$, $n \geq 2$ and $d \geq 3.2((2 + \ln(2))n + \ln(n) + \ln(1/\delta))$. Then*

for all $\phi : \{0,1\}^n \to [-1,1]$ *it holds*

$$\Pr_{U\sim\mu_C}\left[|P_U[\phi] - \mathcal{U}[\phi]| > \tau\right] \leq \frac{(2+\delta)}{2^n\tau^2}. \tag{22}$$

Proof. First, we show the result for an exact unitary 2-design. Using the first moment from Equation (90), we find that

$$\mathop{\mathbf{E}}_{U\sim\mu_U}\left[P_U[\phi]\right] = \sum_{x\in\{0,1\}^n}\left(\mathop{\mathbf{E}}_{U\sim\mu_U}\left[P_U(x)\right]\phi(x)\right) = \mathcal{U}[\phi]. \tag{23}$$

Thus, by Chebyshev's inequality, for any $\tau > 0$,

$$\Pr_{U\sim\mu_U}\left[|P_U[\phi] - \mathcal{U}[\phi]| > \tau\right] \leq \frac{\mathbf{Var}\left[P_U[\phi]\right]}{\tau^2}. \tag{24}$$

The variance is given by

$$\mathop{\mathbf{Var}}_{U\sim\mu_U}\left[P_U[\phi]\right] = \mathop{\mathbf{E}}_{U\sim\mu_U}\left[P_U[\phi]^2\right] - \left(\mathop{\mathbf{E}}_{U\sim\mu_U}\left[P_U[\phi]\right]\right)^2 \tag{25}$$

$$= \sum_x\sum_y \phi(x)\phi(y)\left(\mathop{\mathbf{E}}_{U\sim\mu_U}\left[P_U(x)P_U(y)\right] - \frac{1}{2^{2n}}\right). \tag{26}$$

Inserting the second moment from Equation (91) and bounding $\phi(x)\phi(y) \leq 1$, we find

$$\mathop{\mathbf{Var}}_{U\sim\mu_U}\left[P_U[\phi]\right] = \sum_x\sum_y \phi(x)\phi(y)\left(\frac{1}{2^n(2^n+1)}[1+\delta_{x,y}] - \frac{1}{2^{2n}}\right) \leq \frac{1}{2^{n-1}} = O(2^{-n}) \tag{27}$$

which holds for any exact unitary 2-design.

Let us now turn back to random quantum circuits. At depth $d \geq 3.2((2+\ln(2))n + \ln(n) + \ln(1/\delta))$, the measure μ_C forms an $\delta\cdot 2^{-n}$-approximate 2-design [HH21]. Notice that by Hölder's inequality and Equation (83), it holds

$$\begin{aligned}
&\mathop{\mathbf{E}}_{U\sim\mu_C}\left[P_U(x)P_U(y)\right] - \mathop{\mathbf{E}}_{U\sim\mu_U}\left[P_U(x)P_U(y)\right] \\
&\quad\leq \left|\mathrm{Tr}\left[|x\rangle\langle x|\otimes|y\rangle\langle y|\left(\mathop{\mathbf{E}}_{U\sim\mu_C}\left[U|0^n\rangle\langle 0^n|U^\dagger\right]^{\otimes 2} - \mathop{\mathbf{E}}_{U\sim\mu_U}\left[U|0^n\rangle\langle 0^n|U^\dagger\right]^{\otimes 2}\right)\right]\right| \\
&\quad\leq \left\|\mathop{\mathbf{E}}_{U\sim\mu_C}\left[U|0^n\rangle\langle 0^n|U^\dagger\right]^{\otimes 2} - \mathop{\mathbf{E}}_{U\sim\mu_U}\left[U|0^n\rangle\langle 0^n|U^\dagger\right]^{\otimes 2}\right\|_1 \\
&\quad\leq \frac{\delta}{2^{3n}}.
\end{aligned} \tag{28}$$

Then, as in Equation (27), we arrive at

$$\begin{aligned}\underset{U\sim\mu_C}{\mathbf{Var}}[P_U[\phi]] &= \sum_x \sum_y \phi(x)\phi(y)\left(\underset{U\sim\mu_C}{\mathbf{E}}[P_U(x)P_U(y)] - \frac{1}{2^{2n}}\right) \\ &\leq \sum_x \sum_y \phi(x)\phi(y)\left(\underset{U\sim\mu_U}{\mathbf{E}}[P_U(x)P_U(y)] + \frac{\delta}{2^{3n}} - \frac{1}{2^{2n}}\right) \\ &\leq \frac{1}{2^{n-1}} + \frac{\delta}{2^n} = \frac{2+\delta}{2^n}\,,\end{aligned} \tag{29}$$

which completes the proof. □

4.2 Far from uniformity via unitary designs

In this section, we will use higher moments to show a far from uniform property that holds with probability $1 - \exp(-\Omega(n))$. Notice, that third moments cannot suffice to prove such a statement as the Clifford group is a 3-design but a constant fraction (≈ 0.4) of stabilizer states yield uniform output distributions as shown in Section D.

A standard technique for lower bounding expectation values of ℓ_1-norms is Berger's inequality [Ber97]

$$\mathbf{E}\left[|S|\right] \geq \frac{\left(\mathbf{E}\left[S^2\right]\right)^{\frac{3}{2}}}{\left(\mathbf{E}\left[S^4\right]\right)^{\frac{1}{2}}}, \tag{30}$$

for a random variable S. However, applying this to the expected total variation distance yields another constant lower bound on the expectation value, which is not sufficient to prove a bound with probability $1 - o(1)$. Instead, we will use that, very similarly, we have:

Lemma 8:

$$\|f\|_1 \geq \frac{\|f\|_2^3}{\|f\|_4^2} \tag{31}$$

for functions $f : \{0,1\}^n \to \mathbb{R}$.

Proof. This follows from Hölder's inequality

$$\|f\|_2^2 = \langle f^a, f^b\rangle \leq \left(\sum_x f^{pa}(x)\right)^{\frac{1}{p}} \left(\sum_x f^{qb}(x)\right)^{\frac{1}{q}}, \tag{32}$$

for $a + b = 2$ and $\frac{1}{p} + \frac{1}{q} = 1$. Choosing $a = \frac{4}{3}$, $b = \frac{2}{3}$, $p = 3$ and $q = \frac{3}{2}$, yields the result. □

We will prove concentration inequalities both for the numerator and the denominator using eighth moments and then apply a union bound to show that both scale as desired with high probability. This will allow us to prove a qualitative version of the conjecture by Aaronson and Chen [AC17] in the affirmative.

Theorem 9: *Let μ be a $\frac{1}{1000}2^{-10n}$-approximate unitary 8-design and $n \geq 2$, then:*

$$\mathop{\mathbf{Pr}}_{U\sim\mu}\left[\mathrm{d}_{\mathrm{TV}}\left(P_U,\mathcal{U}\right) \geq \frac{1}{150}\right] \geq 1 - 3200 \times 2^{-n}. \tag{33}$$

Applications of the 8-design property are rare and we pose it as an open problem whether Theorem 9 can be further derandomized, i.e., whether a similarly strong bound can be proved based on lower order moments.

We know that 3-designs are not sufficient as the Clifford group is one. Can the same scaling be shown using only the (approximate) 4-design property? To this end, we prove the following theorem, which does not yield exponential concentration, but a weaker trade-off between the probability and the total variation distance. This property therefore still allows for meaningful average-case hardness statements with probability $1 - o(1)$, which separates 4-designs from the Clifford group. Moreover, far better constants are known for the generation of approximate 4-designs [HH21].

Theorem 10: *Let μ be a $\frac{1}{1000}2^{-10n}$-approximate unitary 4-design and $n \geq 2$, then for any $c > 0$ we have:*

$$\mathop{\mathbf{Pr}}_{U\sim\mu}\left[\mathrm{d}_{\mathrm{TV}}\left(P_U,\mathcal{U}\right) \geq \frac{1}{20\sqrt{c+18}}\right] \geq 1 - 100 \times 2^{-n} - \frac{25}{c}. \tag{34}$$

Before we prove Theorem 9 and Theorem 10, we state the following corollary of Theorem 9 which is due to the bounds from [HH21]:

Corollary 11: *Denote by μ_C the distribution on $\mathrm{U}(2^n)$ obtained from brickwork random quantum circuits of depth d. For $d \geq 1.2 \times 10^{20} n$ and $n \geq 2$, we have*

$$\mathop{\mathbf{Pr}}_{U\sim\mu_C}\left[\mathrm{d}_{\mathrm{TV}}\left(P_U,\mathcal{U}\right) \geq \frac{1}{150}\right] \geq 1 - 3200 \times 2^{-n}. \tag{35}$$

Proof of Theorem 9. Applying Theorem 8 to the function $f(x) = P_U(x) - \mathcal{U}(x)$ we get

$$\left\|P_U - \mathcal{U}\right\|_1 \geq \frac{\left\|P_U - \mathcal{U}\right\|_2^3}{\left\|P_U - \mathcal{U}\right\|_4^2}. \tag{36}$$

The proof strategy is to show concentration inequalities for the numerator and the denominator, independently. Then, we apply the union bound to show that both events are realized simultaneously with high probability.

We apply Chebyshev's inequality to the *collision probability* $Z := \sum_x P_U(x)^2$ in order to estimate $\left\|P_U - \mathcal{U}\right\|_2^2 = Z - 1/D$. We will use the notation $D = 2^n$. In the following, we will make an error of size $10^{-3}D^{-10}$, compared to the Haar value when evaluating monomials $\mathbf{E}\left[P_U(x_1)^{\lambda_1}\ldots P_U(x_k)^{\lambda_k}\right]$ with $\sum_k \lambda_k \leq 8$. In the following calculations, we denote by $E_i \in \mathbb{R}$ with $i = 1, 2, 3, 4$ error terms with $|E_i| \leq 10^{-3}D^{-10}$. We have chosen this error such that it is negligible for any of the

below calculations. The reader may ignore it, but needs to keep in mind that it mildly affects the constants.

Using Theorem 25, we compute the first and second moment of Z by

$$\mathop{\mathbf{E}}_{U\sim\mu}[Z] = \frac{2}{D+1} + DE_1 \tag{37}$$

and

$$\begin{aligned}
\mathop{\mathbf{E}}_{U\sim\mu}\left[Z^2\right] &= \mathop{\mathbf{E}}_{U\sim\mu}\left[\sum_{x\neq y} P_U(x)^2 P_U(y)^2 + \sum_x P_U(x)^4\right] \\
&= \frac{4(D-1)}{(D+1)(D+2)(D+3)} + \frac{24}{(D+1)(D+2)(D+3)} + D^2E_2 \\
&= \frac{4}{(D+1)(D+2)} + \frac{8}{(D+1)(D+2)(D+3)} + D^2E_2 \\
&\leq 4D^{-2} + 8D^{-3}\,.
\end{aligned} \tag{38}$$

We readily compute the variance σ^2 of Z:

$$\begin{aligned}
\sigma^2 &= \mathop{\mathbf{E}}_{U\sim\mu}\left[Z^2\right] - \left(\mathop{\mathbf{E}}_{U\sim\mu}[Z]\right)^2 \\
&\leq 4D^{-2} + 8D^{-3} - 4\frac{1}{(D+1)^2} + 2D|E_1| + D^2|E_1|^2 \\
&\leq 17\times D^{-3}\,,
\end{aligned} \tag{39}$$

where we have used in the last inequality that $1-x^2 < 1$ for $x > 0$ implies $\frac{1}{1+x} > 1-x$ and hence

$$\left(\frac{1}{D+1}\right)^2 = \left(\frac{1}{1+D^{-1}}\right)^2 D^{-2} \geq \left(1-D^{-1}\right)^2 D^{-2} \geq (1-2\times D^{-1})D^{-2}\,, \tag{40}$$

where again the last step is Bernoulli's inequality. Plugging this into Chebyshev's inequality, we find

$$\begin{aligned}
\mathop{\mathbf{Pr}}_{U\sim\mu}\left[\left|\|P_U - \mathcal{U}\|_2^2 - \left(\frac{D-1}{D+1}\right)D^{-1} - DE_1\right| \geq k\right] &= \mathop{\mathbf{Pr}}_{U\sim\mu}\left[\left|Z - \frac{2}{D+1} - DE_1\right| \geq k\right] \\
&= \frac{\sigma^2}{k^2} \leq \frac{17D^{-3}}{k^2}
\end{aligned} \tag{41}$$

for the probability of Z being outside an interval of radius k centered at its mean

Choosing $k = \frac{1}{2}\frac{D-1}{D+1} \cdot D^{-1} - DE_1$, this implies

$$\begin{aligned}
\mathop{\mathbf{Pr}}_{U\sim\mu}\left[\|P_U - \mathcal{U}\|_2^2 \leq \frac{1}{2}\frac{D-1}{D+1}D^{-1}\right] &\leq 17 \times \left(\frac{D-1}{D+1}D^{-1} - DE_1\right)^{-2} \times D^{-3} \\
&\leq 18 \times \left(\frac{D+1}{D-1}\right)^2 D^{-1} \\
&\leq 18 \times \left(\frac{2^1+1}{2^1-1}\right)^2 D^{-1} \\
&\leq 50D^{-1}\,.
\end{aligned} \tag{42}$$

Here we used in the second inequality that $\frac{1}{1-x} \leq 1 + 2x$ for $0 \leq x \leq \frac{1}{2}$, which, after multiplying it with $1 - x$ is equivalent to $0 \leq x(1 - 2x)$. This bound will be sufficient to bound the numerator of Equation (36).

Next, we will find a concentration inequality for the denominator of Equation (36) by essentially the same strategy. We find

$$\begin{aligned}
\|P_U - \mathcal{U}\|_4^4 &= \sum_x \left(P_U(x) - D^{-1}\right)^4 \\
&= \sum_x \left(P_U(x)^4 - 4P_U(x)^3D^{-1} + 6P_U(x)^2D^{-2} - 4P_U(x)D^{-3} + D^{-4}\right) \\
&\leq \underbrace{\sum_x P_U(x)^4}_{=:X} + 6D^{-2}\underbrace{\sum_x P_U(x)^2}_{=Z}\,,
\end{aligned} \tag{43}$$

where we have used $\sum_x (D^{-4} - 4P_U(x)D^{-3}) \leq 0$.

We will prove probability inequalities for the two terms in Equation (43) independently. In fact, the second term can be handled similarly to Equation (42). We apply Equation (41) with $k = \frac{1}{D+1} - DE_1$ to obtain

$$\mathop{\mathbf{Pr}}_{U\sim\mu}\left[Z \geq \frac{3}{D+1} \times D^{-1}\right] \leq 18(D+1)^2D^{-3} \leq 30D^{-1}. \tag{44}$$

For the first term in Equation (43) we use Theorem 25 to compute the first

$$\begin{aligned}
\mathop{\mathbf{E}}_{U\sim\mu}[X] = \sum_x \mathop{\mathbf{E}}_{U\sim\mu}\left[P_U(x)^4\right] &= \frac{4!}{D\cdots(D+3)} + DE_3 \\
&= \frac{24}{(D+1)(D+2)(D+3)} + DE_3\,,
\end{aligned} \tag{45}$$

and second moments

$$\begin{aligned}\mathop{\mathbf{E}}_{U\sim\mu}\left[X^2\right] &= \sum_x \mathop{\mathbf{E}}_{U\sim\mu}\left[P_U(x)^8\right] + \sum_{x\neq y}\mathop{\mathbf{E}}_{U\sim\mu}\left[P_U(x)^4 P_U(y)^4\right] \\ &= \frac{8!}{(D+1)\cdots(D+7)} + \frac{24^2(D-1)}{(D+1)\cdots(D+7)} + D^2E_4 \\ &= \frac{8! - 2\times 24^2}{(D+1)\cdots(D+7)} + \frac{24^2}{(D+2)\cdots(D+7)} + D^2E_4\,.\end{aligned} \tag{46}$$

For the variance this implies

$$\begin{aligned}\sigma_X^2 &\leq \frac{8!}{(D+1)\cdots(D+7)} + \frac{24^2}{(D+2)\cdots(D+7)} + D^2E_4 \\ &\quad - \frac{24^2}{(D+1)^2(D+2)^2(D+3)^2} + 2\frac{24^2}{(D+1)^2(D+2)^2(D+3)^2}D|E_3| + D^2E_3^2 \\ &\leq 41000D^{-7} + 24^2\underbrace{\left(\frac{1}{(D+2)\cdots(D+7)} - \frac{1}{(D+1)^2(D+2)^2(D+3)^2}\right)}_{<0} \\ &\leq 41000D^{-7}.\end{aligned} \tag{47}$$

Via Chebyshev's inequality, we obtain

$$\mathop{\mathbf{Pr}}_{U\sim\mu}\left[\left|X - \frac{24}{(D+1)(D+2)(D+3)} + DE_3\right| \geq k\right] \leq \frac{41000D^{-7}}{k^2}. \tag{48}$$

Choosing $k = \frac{12}{(D+1)(D+2)(D+3)} - DE_3$, this implies

$$\mathop{\mathbf{Pr}}_{U\sim\mu}\left[X \geq \frac{36}{(D+1)(D+2)(D+3)}\right] \leq \frac{41001}{12^2}D^{-7}(D+1)^2(D+2)^2(D+3)^2 \leq 3100D^{-1}\,, \tag{49}$$

for $n \geq 2$.

We can now put these bounds together via a union bound applied twice: For any

$\epsilon_1, \epsilon_2 > 0$ such that $\frac{\epsilon_1}{\epsilon_2} \geq \epsilon$, we find

$$\begin{aligned}
\Pr_{U\sim\mu}\left[\|P_U - \mathcal{U}\|_1 \leq \epsilon\right] &\leq \Pr_{U\sim\mu}\left[\frac{\|P_U - \mathcal{U}\|_2^3}{\|P_U - \mathcal{U}\|_4^2} \leq \epsilon\right] \\
&\leq \Pr_{U\sim\mu}\left[\left(\|P_U - \mathcal{U}\|_2^3 \leq \epsilon_1\right) \vee \left(\|P_U - \mathcal{U}\|_4^2 \geq \epsilon_2\right)\right] \\
&\leq \Pr_{U\sim\mu}\left[\|P_U - \mathcal{U}\|_2^3 \leq \epsilon_1\right] + \Pr_{U\sim\mu}\left[\|P_U - \mathcal{U}\|_4^2 \geq \epsilon_2\right] \\
&= \Pr_{U\sim\mu}\left[\|P_U - \mathcal{U}\|_2^2 \leq \epsilon_1^{\frac{2}{3}}\right] + \Pr_{U\sim\mu}\left[\|P_U - \mathcal{U}\|_4^4 \geq \epsilon_2^2\right] \\
&\leq \Pr_{U\sim\mu}\left[\|P_U - \mathcal{U}\|_2^2 \leq \epsilon_1^{\frac{2}{3}}\right] + \Pr_{U\sim\mu}\left[X + 6D^2(Z + 2^{-n}) \geq \epsilon_2^2\right] \\
&\leq \Pr_{U\sim\mu}\left[\|P_U - \mathcal{U}\|_2^2 \leq \epsilon_1^{\frac{2}{3}}\right] + \Pr_{U\sim\mu}\left[X \geq a_1\right] + \Pr_{U\sim\mu}\left[6D^{-2}Z \geq a_2\right],
\end{aligned} \tag{50}$$

with $a_1 + a_2 = \epsilon_2^2$. These probabilities are bounded in Equations (42), (44) and (49). Choosing $\epsilon_1^{\frac{2}{3}} = \frac{1}{2}\frac{D-1}{D+1}D^{-1} \geq 10^{-\frac{2}{3}}D^{-1}$ (for $n \geq 2$), $a_1 = 36 \times D^{-3}$ and $a_2 = 18 \times D^{-3}$ implies $\epsilon_2^2 = 54 \times D^{-3}$. Plugging this into Equation (50) yields Theorem 9, where $\mathrm{d_{TV}}\left(P_U, \mathcal{U}\right) = \frac{1}{2}||P_U - \mathcal{U}||_1$. □

Proof of Theorem 10. The proof of Theorem 10 follows analogously to the proof of Theorem 9. However, instead of proving concentration of the random variable $X = \sum_x P_U(x)^4$ to bound $\mathbf{Pr}[X \geq a_1]$ as in Equation (50), we instead apply Markov's inequality, to get

$$\Pr_{U\sim\mu}[X \geq a_1] \leq \frac{\mathbf{E}_{U\sim\mu}[X]}{a_1} \leq \frac{24}{(D+1)(D+2)(D+3)} \cdot \frac{1}{a_1} + DE_3 \leq \frac{25D^{-3}}{a_1}. \tag{51}$$

Therefore, the result follows by choosing $a_1 = cD^{-3}$ for $c > 0$. □

5 Random quantum circuits of sub-linear depth

In this section, we bound the two key quantities, namely $\mathfrak{f}$ and the far from uniform probability, for random quantum circuits in the regime of sub-linear depth. This regime of circuit depth includes the shortest circuits for which we can still show super-polynomial query complexity lower bounds and hence hardness of learning. Note that in this regime, the random circuit ensembles that we consider do not yet form even a unitary 2-design requiring different techniques to obtain these bounds. Plugging these bounds into Theorem 1, we obtain the following lower bound on the average case query complexity.

Theorem 12: (Formal version of sub-linear depth part of Informal Theorem 1) *Let* $c = 1/\log(5/4)$. *Further, let* $c\log n \leq d \leq c(n + \log n)$, $\tau > 0$ *and* $\epsilon \leq 1/4 - \tau$. *Then, for sufficiently large* n, *the average case query complexity* q *of* ϵ*-learning any*

β-fraction of brickwork random quantum circuits is lower bounded by

$$q+1 \geq \frac{(\beta - 3/4 - \epsilon - \tau)\,\tau^2}{3n} \cdot \left(\frac{4}{5}\right)^d , \tag{52}$$

Moreover, if $d > c(n + \log n)$ then it holds

$$q+1 \geq (\beta - 3/4 - \epsilon - \tau)\, 2^{n-2}\tau^2 . \tag{53}$$

Remark 5: Note that Theorem 12 implies that in the practically relevant regime of $\tau = 1/\mathrm{poly}(n)$, learning circuits of depth $\omega(\log(n))$ to some constant precision requires a super-polynomial number of queries. In particular, already for any constant fraction β slightly greater than $3/4$ there exists no efficient statistical query algorithm for any accuracy $\tau = \Omega(1/\mathrm{poly}(n))$.

5.1 Maximally distinguishable fraction via restricted depth moments

We begin with bounding the maximally distinguishable fraction.

Lemma 13: *For all $d \geq \frac{\log n}{\log 5/4}$ and for all $\phi : \{0,1\}^n \to [-1,1]$ it holds*

$$\mathop{\mathbf{Pr}}_{U\sim\mu_C}\left[|P_U[\phi] - \mathcal{U}[\phi]| > \tau\right] \leq \frac{1}{\tau^2}\left[n\left(\frac{4}{5}\right)^d\left(1+\frac{1}{2^n}\right)+\frac{1}{2^n}\right] . \tag{54}$$

Remark 6: Note that Equation (54) can be simplified as follows, which lead to the two cases in Theorem 12:

$$\mathop{\mathbf{Pr}}_{U\sim\mu_C}\left[|P_U[\phi] - \mathcal{U}[\phi]| > \tau\right] \leq \begin{cases} \frac{3n}{\tau^2}\cdot\left(\frac{4}{5}\right)^d , & \text{for } d \leq \frac{n+\log(n)}{\log(5/4)} \\ \frac{3}{2^n\tau^2} , & \text{for } d > \frac{n+\log(n)}{\log(5/4)} . \end{cases} \tag{55}$$

Proof of Theorem 13: The proof strategy is essentially identical to the first part of the proof of Theorem 7 bounding the same quantity for linear depth circuits. However, one replaces the moments over the Haar measure with the moments of restricted depth models obtained in Theorem 26. In particular, the second moment in Equation (26) gets replaced by the one given in Equation (93). The moments in Theorem 26 hold for restricted depth circuits in 1D. They are obtained via a mapping to a statistical mechanics model [Hun19] also used in [BCG21] to bound a similar quantity, for more details see Section C. □

5.2 Far from uniformity for constant-depth circuits

A direct approach to bound $\mathbf{Pr}\left[\mathrm{d_{TV}}(P_U,\mathcal{U}) \geq \epsilon\right]$ is to lower bound the expectation $\mathbf{E}\left[\mathrm{d_{TV}}(P_U,\mathcal{U})\right]$. Then, using Markov's inequality, a far-from-uniform-bound follows immediately as made explicit by the following lemma:

Lemma 14: *For any random variable* $0 \leq Y \leq 1$ *and any* $0 < \epsilon < 1$ *we have*

$$\mathbf{Pr}\left(Y \geq \epsilon\right) \geq \frac{\mathbf{E}\left[Y\right] - \epsilon}{1 - \epsilon}\,. \tag{56}$$

In [AC17], Aaronson and Chen show the following lower bound on $\mathbf{E}_{U\sim\mu_C}\left[\mathrm{d_{TV}}(P_U, \mathcal{U})\right]$.

Lemma 15: (Section 3.5, [AC17]) *For any* $n \geq 2, d \geq 1$, *it holds that*

$$\mathop{\mathbf{E}}_{U\sim\mu_C}\left[\mathrm{d_{TV}}(P_U, \mathcal{U})\right] \geq \frac{1}{4}\,. \tag{57}$$

Curiously, their proof only takes into account the randomness in drawing the very last two-qubit gate which is why the bound holds already at depth $d = 1$. By Theorem 14, we immediately find:

Corollary 16: *For any* $n \geq 2, d \geq 1, \epsilon \in [0, 1/4]$, *it holds that*

$$\mathop{\mathbf{Pr}}_{U\sim\mu_C}\left[\mathrm{d_{TV}}(P_U, \mathcal{U}) \geq \epsilon\right] \geq \frac{\frac{1}{4} - \epsilon}{1 - \epsilon} \geq \frac{1}{4} - \epsilon\,. \tag{58}$$

Acknowledgements

We thank Yihui Quek, Dominik Hangleiter, Jean-Pierre Seifert, Scott Aaronson, and Lijie Chen for many helpful discussions and insights. J. H. acknowledges funding from the Harvard Quantum Initiative. The work of R. K. is supported by the SFB BeyondC (Grant No. F7107- N38), the Project QuantumReady (FFG 896217) and the BMWK (ProvideQ). The Berlin team thanks the BMWK (PlanQK, EniQmA), the BMBF (Hybrid), and the Munich Quantum Valley (K-8). This work has also been funded by the Deutsche Forschungsgemeinschaft (DFG) under Germany's Excellence Strategy – The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689) as well as CRC 183 (B1).

A Omitted proofs for Haar random unitaries

In this appendix section, we provide the technical details for studying the Haar random case. We first introduce Gaussian integration – a powerful way to compute expectation values of functions over uniformly random pure states Haar random states. This technique is complementary to the more widely used Weingarten formalism. The two approaches have different strengths and weaknesses. One core advantage of Gaussian integration is that it can be readily applied to functions that are not well approximated by homogeneous polynomials. The expected TV distance between a Haar random outcome distribution and the uniform distribution is one such function which features prominently in this work.

Subsequently, we will also provide tight bounds on the Lipschitz constants of these

functions. This is the only missing ingredient to show exponentially strong concentration around the previously computed expectation values by means of Levy's lemma.

We emphasize that this appendix section covers the extreme case of global Haar-random unitaries. Such evolutions scramble information across all subsystems and we therefore present our results directly in terms of total Hilbert space dimension D, an abstract computational basis $\{|1\rangle, \dots, |D\rangle\}$ and global (pure) states $|\psi\rangle = U|\psi_0\rangle \in \mathbb{C}^D$, where $|\psi_0\rangle$ is a designated (simple) starting state, e.g., a product state. For n-qubit systems, this would amount to setting $D = 2^n$ and equating the d-th basis state $|d\rangle$ with the n-bit string $\llcorner d \lrcorner = x_1(d) \dots x_n(d) \in \{0,1\}^n$ that corresponds to the binary representation of $d \in \mathbb{N}$.

A.1 Haar random state averages via Gaussian integration

Gaussian integration is a standard technique in mathematical signal processing and data science that allows us to recast an expectation value over the unit sphere as an expectation value over standard Gaussian random vectors. The latter has the distinct advantage that individual vector components become statistically independent from each other and follow the very simple normal distribution. These features are very useful for computing non-polynomial functions that only depend on comparatively few vector entries. We will showcase this technique based on two exemplary expectation values that feature prominently in this work:

(i) the function value $P_U[\phi] = \sum_{d=1}^{D} \phi(d) \left|\langle d|U|\psi_0\rangle\right|^2$ averaged uniformly over all unitaries U, see Theorem 18 below and

(ii) the expected TV distance between the outcome distribution of a Haar random pure state $|\psi\rangle = U|\psi_0\rangle$ and the uniform distribution, see Theorem 19 below.

At the heart of both computations is the following powerful meta-theorem, known as Gaussian integration.

Theorem 17: (Gaussian integration) *Let $f : \mathbb{C}^D \to \mathbb{C}$ be a homogeneous function with even degree $2k \in 2\mathbb{N}$, i.e., $f(a\mathbf{x}) = |a|^{2k} f(\mathbf{x})$ for $\mathbf{x} \in \mathbb{C}^D$ and $a \in \mathbb{C}$. Then, we can rewrite the Haar expectation value as*

$$\mathop{\mathbf{E}}_{|\psi\rangle\sim\mu_S}[f(|\psi\rangle)] = \frac{1}{k!2^k}\binom{D+k-1}{k}^{-1} \mathop{\mathbf{E}}_{g_i,h_i \overset{iid}{\sim} \mathcal{N}(0,1)} \left[f\left(g_1 + \mathrm{i}h_1, \dots, g_D + \mathrm{i}h_D\right)\right], \quad (59)$$

where $g_i, h_i \overset{iid}{\sim} \mathcal{N}(0,1)$ denote independent standard Gaussian random variables ($\mu = 0$, $\sigma^2 = 1$).

In words: the left hand side features the Haar average over all pure states (complex unit vectors), while the right hand side features an expectation value over $2D$ statistically independent standard Gaussian random variables $g_j, h_j \sim \mathcal{N}(0,1)$ with probability density function $\exp\left(-x^2/2\right)/\sqrt{2\pi}$. This reformulation can be extremely

helpful in concrete calculations, because the individual vector components on the right hand side are all statistically independent.

Proof. Let us start by considering the complex-valued standard Gaussian vector $\mathbf{g} + \mathrm{i}\mathbf{h} \in \mathbb{C}^D$ with $\mathbf{g} = (g_1, \ldots, g_D), \mathbf{h} = (h_1, \ldots, h_D) \in \mathbb{R}^D$ and $g_j, h_j \overset{iid}{\sim} \mathcal{N}(0,1)$. Switching to polar coordinates allows us to rewrite this vector as

$$\mathbf{g} + \mathrm{i}\mathbf{h} = r(\mathbf{g}, \mathbf{h}) |\widehat{\mathbf{g} + \mathrm{i}\mathbf{h}}\rangle \tag{60}$$

with direction $|\widehat{\mathbf{g} + \mathrm{i}\mathbf{h}}\rangle \in \mathbb{S}^{D-1}$ and radius

$$r(\mathbf{g}, \mathbf{h})^2 = \sum_{j=1}^{D} \left(g_j^2 + h_j^2\right) \tag{61}$$

We can now use homogeneity of the function $f : \mathbb{C}^D \to \mathbb{C}$ to rewrite the Gaussian expectation value on the right hand side of Equation (59) as

$$\mathop{\mathbf{E}}_{\mathbf{g},\mathbf{h}\sim\mathcal{N}(0,\mathbb{I})} \left[f(\mathbf{g} + \mathrm{i}\mathbf{h})\right] = \mathop{\mathbf{E}}_{\mathbf{g},\mathbf{h}} \left[f\left(r(\mathbf{g}, \mathbf{h}) |\widehat{\mathbf{g} + \mathrm{i}\mathbf{h}}\rangle\right)\right] = \mathop{\mathbf{E}}_{\mathbf{g},\mathbf{h}} \left[r(\mathbf{g}, \mathbf{h})^{2k} f\left(|\widehat{\mathbf{g} + \mathrm{i}\mathbf{h}}\rangle\right)\right], \tag{62}$$

where we have succinctly accumulated all Gaussian random variables into two vectors $\mathbf{g}$ and $\mathbf{h}$. Now, something interesting happens. The particular form of the standard Gaussian probability density ensures that radius ($r(\mathbf{g}, \mathbf{h})$) and direction ($|\widehat{\mathbf{g} + \mathrm{i}\mathbf{h}}\rangle$) can be decomposed into two statistically independent random variables / vectors. The square r^2 of the former follows a χ^2_{2D}-distribution with $2D$ degrees of freedom while the latter must be a unit vector that is sampled uniformly from the complex unit sphere $\mathbb{S}^{D-1}$. The latter is a consequence of the fact that the distribution of standard Gaussian vectors $\mathbf{g} + \mathrm{i}\mathbf{h}$ is invariant under unitary transformations. Statistical independence, on the other hand, can be deduced from transforming the probability density function (pdf) of a Gaussian random vector into generalized spherical coordinates. Under such a transformation, the pdf decomposes into a product of a radial part (the radius) and an angular part. We can now use these insights to decompose the original expectation value into a product of two expectation values:

$$\begin{aligned} \mathop{\mathbf{E}}_{\mathbf{g},\mathbf{h}\overset{iid}{\sim}\mathcal{N}(0,\mathbb{I})} \left[r(\mathbf{g}, \mathbf{h})^{2k} f\left(|\widehat{\mathbf{g} + \mathrm{i}\mathbf{h}}\rangle\right)\right] &= \mathop{\mathbf{E}}_{r^2\sim\chi^2_{2D},\, \widehat{\mathbf{g}+\mathrm{i}\mathbf{h}}\sim\mu_S} \left[r^{2k} f\left(|\widehat{\mathbf{g} + \mathrm{i}\mathbf{h}}\rangle\right)\right] \\ &= \mathop{\mathbf{E}}_{r^2\sim\chi^2_{2D}} \left[r^{2k}\right] \mathop{\mathbf{E}}_{\widehat{\mathbf{g}+\mathrm{i}\mathbf{h}}\sim\mu_S} \left[f\left(\widehat{\mathbf{g} + \mathrm{i}\mathbf{h}}\rangle\right)\right]. \end{aligned} \tag{63}$$

The second expression describes a uniform integral of f over the complex unit sphere in D dimensions. This is exactly the Haar expectation value on the left hand side of Equation (59). The other expression is the k-th moment of a χ^2 random variable with $2D$ degrees of freedom. These are well-known and amount to

$$\mathop{\mathbf{E}}_{r^2\sim\chi^2_{2D}} \left[r^{2k}\right] = \frac{\Gamma(2D + k/2)}{\Gamma(k/2)} = (2D)(2D+2)\cdots(2D+2(k-1)) = k! 2^k \binom{D+k-1}{k}.$$

Putting everything together, allows us to conclude

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{g},\mathbf{h}\overset{iid}{\sim}\mathcal{N}(0,\mathbb{I})}\left[f\left(\mathbf{g}+\mathrm{i}\mathbf{h}\right)\right] =& \mathop{\mathbf{E}}_{g_i,h_i\overset{iid}{\sim}\mathcal{N}(0,1)}\left[f\left(g_1+\mathrm{i}h_1,\ldots,g_D+\mathrm{i}h_D\right)\right]\\ =& \mathop{\mathbf{E}}_{r^2\sim\chi^2_{2D}}\left[r^{2k}\right]\mathop{\mathbf{E}}_{|\psi\rangle\sim\mu_S}\left[f(|\psi\rangle)\right]\\ =& k!2^k\binom{D+k-1}{k}\mathop{\mathbf{E}}_{|\psi\rangle\sim\mu_S}\left[f(|\psi\rangle)\right].\end{aligned} \tag{64}$$

The claim in Equation (59) is an immediate reformulation of this equality. □

We now have the essential tool at hand to compute the Haar expectation values that matter for this work.

Theorem 18: (Haar average of bounded function expectation values) *Let* $\phi : \{1,\ldots,D\}\to[-1,1]$ *be a function and define* $P_U[\phi]=\sum_{d=1}^{D}\phi(d)\left|\langle d|U|\psi_0\rangle\right|^2$. *Then, the expectation value of* $P_U[\phi]$ *over Haar random unitaries becomes*

$$\mathop{\mathbf{E}}_{U\sim\mu_U}\left[P_U[\phi]\right]=\mathop{\mathbf{E}}_{|\psi\rangle\sim\mu_S}\left[\sum_{d=1}^{D}\phi(d)\left|\langle d|\psi\rangle\right|^2\right]=\sum_{d=1}^{D}\frac{1}{D}\phi(d)=\mathcal{U}\left[\phi\right]. \tag{65}$$

This is a standard result that readily follows from Haar integration via Weingarten calculus and the representation theory of the unitary group. Let us now show how to achieve the same result with Gaussian integration (Theorem 17).

Proof. Let us start by using linearity to rewrite the desired expectation value as

$$\mathop{\mathbf{E}}_{U\sim\mu_U}\left[P_U\left[\phi\right]\right]=\sum_{d=1}^{D}\phi(d)\mathop{\mathbf{E}}_{|\psi\rangle\sim\mu_S}\left[\left|\langle d|\psi\rangle\right|^2\right]. \tag{66}$$

The claim then follows from the following equality:

$$\mathop{\mathbf{E}}_{|\psi\rangle\sim\mu_S}\left[\left|\langle d|\psi\rangle\right|^2\right]=\frac{1}{D}\quad\text{for all } d=1,\ldots,D. \tag{67}$$

which we now prove for any basis vector d. Once d is fixed, we can interpret this as the expectation value of the function $f_d(|\psi\rangle)=\left|\langle d|\psi\rangle\right|^2$ which selects the d'th vector entry (amplitude) and outputs its magnitude squared. Every such function is homogeneous with even degree 2 ($k=1$) and we can use Theorem 17 to rewrite the expectation value as

$$\begin{aligned}\mathop{\mathbf{E}}_{|\psi\rangle\sim\mu_S}\left[\left|\langle d|\psi\rangle\right|^2\right] =& \mathop{\mathbf{E}}_{|\psi\rangle\sim\mu_S}\left[f_d(|\psi\rangle)\right]\\ =& \frac{1}{1!2^1}\binom{D}{1}^{-1}\mathop{\mathbf{E}}_{g_i,h_i\overset{iid}{\sim}\mathcal{N}(0,1)}\left[f_d(g_1+\mathrm{i}h_1,\ldots,g_D+\mathrm{i}h_D)\right]\\ =& \frac{1}{2D}\mathop{\mathbf{E}}_{g_i,h_i\overset{iid}{\sim}\mathcal{N}(0,1)}\left[\left|g_d+\mathrm{i}h_d\right|^2\right]=\frac{1}{2D}\left(\mathop{\mathbf{E}}_{g_d\sim\mathcal{N}(0,1)}\left[g_d^2\right]+\mathop{\mathbf{E}}_{h_d\sim\mathcal{N}(0,1)}\left[h_d^2\right]\right)\end{aligned}$$

$$= \frac{1}{2D}(1+1) = \frac{1}{D}.$$

Here, we have used statistical independence (we can forget all Gaussian random variables which do not feature in the expression), as well as the fact that Gaussian standard variables obey $\mathbf{E}[g_d^2] = \mathbf{E}[h_d^2] = 1$ (unit variance). □

The next expectation value is more intricate by comparison. There, Weingarten calculus would not work, because the function involved is not a well-behaved polynomial. Gaussian integration, however, can handle such non-polynomial expectation values and yields the following elegant display.

Theorem 19: (Haar expectation of the TV distance) *Let* $P_U(d) = |\langle d|U|\psi_0\rangle|^2$ *be the output distribution of a* D *level quantum system in the computational basis and* $\mathcal{U}$ *the* D *dimensional uniform distribution. The expectation value of the total variation distance* $\mathrm{d_{TV}}(P_U, \mathcal{U})$ *over Haar random unitaries* U *obeys*

$$\frac{1}{\mathrm{e}} - \frac{1}{2\sqrt{D}} \leq \mathop{\mathbf{E}}_{U\sim\mu_U}\left[\mathrm{d_{TV}}\left(P_U, \mathcal{U}\right)\right] \leq \frac{1}{\mathrm{e}} + \frac{1}{2\sqrt{D}}.$$

The approximation error is controlled by $1/(2\sqrt{D})$ *which diminishes exponentially for* n*-qubit systems* ($D = 2^n$).

Proof. Let us once more start by using linearity of the expectation value to rewrite the desired expression as

$$\begin{aligned}\mathop{\mathbf{E}}_{U\sim\mu_U}\left[\mathrm{d_{TV}}\left(P_U, \mathcal{U}\right)\right] &= \frac{1}{2}\sum_{d=1}^{D} \mathop{\mathbf{E}}_{|\psi\rangle\sim\mu_S}\left[\frac{1}{2}\sum_{d=1}^{D}\left||\langle d|\psi\rangle|^2 - \frac{1}{D}\right|\right] \\ &= \frac{1}{2D}\sum_{d=1}^{D} \mathop{\mathbf{E}}_{|\psi\rangle\sim\mu_S}\left[\left|D\,|\langle d|\psi\rangle|^2 - 1\right|\right]. \end{aligned} \tag{68}$$

Next, note that unitary invariance of $|\psi\rangle \sim \mu_S$ ensures that each of the summands on the right hand side must yield the same expectation value. This allows us to simplify further and obtain

$$\mathop{\mathbf{E}}_{U\sim\mu_U}\left[\mathrm{d_{TV}}\left(P_U, \mathcal{U}\right)\right] = \frac{1}{2D}\sum_{d=1}^{D} \mathop{\mathbf{E}}_{|\psi\rangle\sim\mu_S}\left[\left|D\,|\langle d|\psi\rangle|^2 - 1\right|\right] = \frac{1}{2} \mathop{\mathbf{E}}_{|\psi\rangle\sim\mu_S}\left[\left|D\,|\langle 1|\psi\rangle|^2 - 1\right|\right].$$

Note that this is not a polynomial function, but we can still use Gaussian integration to accurately bound this expression. To this end, we first use $\langle\psi|\psi\rangle = 1$ to rewrite the expression of interest as the uniform average over a homogeneous function with even degree 2 ($k = 1$):

$$f\left(|\psi\rangle\right) = \frac{1}{2}\left|D\,|\langle 1|\psi\rangle|^2 - \langle\psi|\psi\rangle\right|^2.$$

We can now use Theorem 17 to conclude

$$\begin{aligned}
\mathop{\mathbf{E}}_{U\sim\mu_U}\left[\mathrm{d_{TV}}\left(P_U,\mathcal{U}\right)\right] =& \mathop{\mathbf{E}}_{|\psi\rangle\sim\mu_S}\left[\frac{1}{2}\left|D\left|\langle 1|\psi\rangle\right|^2-\langle\psi|\psi\rangle\right|\right] = \mathop{\mathbf{E}}_{|\psi\rangle\sim\mu_S}\left[f(|\psi\rangle)\right] \\
=& \frac{1}{1!2^1}\binom{D}{1}^{-1} \mathop{\mathbf{E}}_{g_i,h_i\overset{iid}{\sim}\mathcal{N}(0,1)}\left[f\left(g_1+\mathrm{i}h_1,\ldots,g_d+\mathrm{i}h_d\right)\right] \\
=& \frac{1}{2D} \mathop{\mathbf{E}}_{g_j,h_j\overset{iid}{\sim}\mathcal{N}(0,1)}\left[\frac{1}{2}\left||D\langle 1|\mathbf{g}+\mathrm{i}\mathbf{h}\rangle|^2-\langle\mathbf{g}+\mathrm{i}\mathbf{h}|\mathbf{g}+\mathrm{i}\mathbf{h}\rangle\right|\right] \\
=& \frac{1}{4D} \mathop{\mathbf{E}}_{g_j,h_j\overset{iid}{\sim}\mathcal{N}(0,1)}\left[\left|D\left(g_1^2+h_1^2\right)-\sum_{j=1}^{D}\left(g_j^2+h_j^2\right)\right|\right] \\
=& \frac{1}{2} \mathop{\mathbf{E}}_{g_j,h_j\overset{iid}{\sim}\mathcal{N}(0,1)}\left[\left|\frac{1}{2}\left(g_1^2+h_1^2\right)-1+1-\frac{1}{2D}\sum_{j=1}^{D}\left(g_j^2+h_j^2\right)\right|\right].
\end{aligned} \tag{69}$$

It seems possible to compute this expectation value directly by rewriting each expectation value as an integral weighted by the Gaussian probability density function $\exp\left(-g_j^2/2\right)$, but doing so would incur a total of $2D$ nested integrations. Here, we instead simplify the derivation considerably by providing reasonably tight upper and lower bounds. We have already suggestively decomposed the expression within the absolute value into two terms that are easier to control individually:

$$\begin{aligned}
M =& \frac{1}{2} \mathop{\mathbf{E}}_{g_1,h_1\overset{iid}{\sim}\mathcal{N}(0,1)}\left[\left|\frac{1}{2}\left(g_1^2+h_1^2\right)-1\right|\right] && \text{(asymptotic mean value)}, \\
\Delta =& \frac{1}{2} \mathop{\mathbf{E}}_{g_j,h_j\overset{iid}{\sim}\mathcal{N}(0,1)}\left[\left|1-\frac{1}{2D}\sum_{j=1}^{D}\left(g_j^2+h_j^2\right)\right|\right] && \text{(approximation error)}.
\end{aligned}$$

Applying the triangle inequality to Equation (69) readily allows us to infer

$$M-\Delta \leq \mathop{\mathbf{E}}_{U\sim\mu_U}\left[\mathrm{d_{TV}}\left(P_U,\mathcal{U}\right)\right] \leq M+\Delta \tag{70}$$

and the two remaining parameters can now be computed independently. We defer the actual calculations to the end of this subsection and only state the results here:

$$\begin{aligned}
M =& 1/\mathrm{e} && \text{(see Theorem 21 below)}, \\
\Delta \leq& 1/\left(2\sqrt{D}\right) && \text{(see Theorem 20 below)}.
\end{aligned}$$

Inserting these numerical values into Equation (70) yields the claim. □

Let us now supply the technical calculations that are essential for completing the proof of Theorem 19.

Lemma 20: *Let $g_1,\ldots,g_d,h_1,\ldots,h_d$ be $2D$ independent standard Gaussian ran-*

dom variables. Then,

$$\Delta = \frac{1}{2} \mathop{\mathbf{E}}_{g_i,h_i \overset{iid}{\sim} \mathcal{N}(0,1)} \left[\left| 1 - \frac{1}{2D} \sum_{j=1}^{D} \left(g_j^2 + h_j^2 \right) \right| \right] \leq \frac{1}{2\sqrt{D}}.$$

Proof. There is no conceptual difference between the g_j and h_j random variables. So, we may replace them with $2D$ independent standard Gaussian random variables $\tilde{g}_1, \ldots, \tilde{g}_{2D} \overset{iid}{\sim} \mathcal{N}(0,1)$. This reformulation yields

$$\Delta = \frac{1}{4D} \mathop{\mathbf{E}}_{\tilde{g}_i \overset{iid}{\sim} \mathcal{N}(0,1)} \left[\left| 2D - \sum_{j=1}^{2D} \tilde{g}_j^2 \right| \right] \leq \frac{1}{4D} \mathop{\mathbf{E}}_{\tilde{g}_i \overset{iid}{\sim} \mathcal{N}(0,1)} \left[\left(2D - \sum_{j=1}^{2D} \tilde{g}_j^2 \right)^2 \right]^{1/2}, \tag{71}$$

where the last inequality is Jensen's. The remaining expectation value is the variance of a χ^2 random variable with $2D$ degrees of freedom. Standard textbooks tell us that this variance is equal to $2(2D) = 4D$. We do, however, believe that it is instructive to compute this variance directly, because it showcases an important subroutine when computing Haar integrals via Gaussian integration. Note that the random variables involved obey

$$\mathop{\mathbf{E}}_{\tilde{g}_j, \tilde{g}_k \overset{iid}{\sim} \mathcal{N}(0,1)} \left[\tilde{g}_j^2 \tilde{g}_k^2 \right] = \begin{cases} \mathbf{E}_{\tilde{g}_j \sim \mathcal{N}(0,1)} \left[\tilde{g}_j^2 \right] \mathbf{E}_{\tilde{g}_k \sim \mathcal{N}(0,1)} \left[\tilde{g}_k^2 \right] = 1 \,, & \text{whenever } j \neq k \text{ and} \\ \mathbf{E} \left[\tilde{g}_j^4 \right] = 3 \,, & \text{else if } j = k. \end{cases}$$

Combining them yields $\mathbf{E}_{\tilde{g}_j, \tilde{g}_k \overset{iid}{\sim} \mathcal{N}(0,1)} \left[\tilde{g}_j^2 \tilde{g}_k^2 \right] = 1 + 2\delta_{j,k}$.

If we also recall $\mathbf{E}_{\tilde{g}_j \sim \mathcal{N}(0,1)} \left[\tilde{g}_j^2 \right] = 1$, we can readily conclude

$$\begin{aligned} \mathop{\mathbf{E}}_{\tilde{g}_i \overset{iid}{\sim} \mathcal{N}(0,1)} \left[\left(2D - \sum_{j=1}^{2D} \tilde{g}_j^2 \right)^2 \right] &= 4D^2 - 4D \sum_{j=1}^{2D} \mathop{\mathbf{E}}_{\tilde{g}_j \sim \mathcal{N}(0,1)} \left[\tilde{g}_j^2 \right] + \sum_{j=1}^{2D} \sum_{k=1}^{2D} \mathop{\mathbf{E}}_{\tilde{g}_j, \tilde{g}_k \overset{iid}{\sim} \mathcal{N}(0,1)} \left[\tilde{g}_j^2 \tilde{g}_k^2 \right] \\ &= 4D^2 - 4D \sum_{j=1}^{2D} 1 + \sum_{j=1}^{2D} \sum_{k=1}^{2D} (1 + 2\delta_{j,k}) \\ &= 4D^2 - 8D^2 + 4D^2 + 4D = 4D. \end{aligned}$$

Inserting this variance expression into Equation (71) yields the claim. □

Lemma 21: *Let g, h be two independent standard Gaussian random variables. Then*

$$M = \frac{1}{2} \mathop{\mathbf{E}}_{g,h \overset{iid}{\sim} \mathcal{N}(0,1)} \left[\left| \frac{1}{2} \left(g^2 + h^2 \right) - 1 \right| \right] = \frac{1}{\mathrm{e}}.$$

Proof. This is the one location in our derivation, where we really utilize the power of Gaussian integration. We start by rewriting the expectation value as an integral over two independent standard Gaussian random variables with (standard Gaussian)

probability density functions $\exp\left(-g^2/2\right)/\sqrt{2\pi}$ and $\exp\left(-h^2/2\right)/\sqrt{2\pi}$, respectively:

$$M = \frac{1}{2} \mathop{\mathbf{E}}_{g,h \overset{iid}{\sim} \mathcal{N}(0,1)} \left[\left| \frac{1}{2}\left(g^2+h^2\right) - 1 \right| \right] = \frac{1}{4} \iint_{-\infty}^{\infty} \left| g^2+h^2-2 \right| \frac{\exp\left(-(g^2+h^2)/2\right)}{2\pi} \mathrm{d}g\mathrm{d}h.$$

Next, we view $(g,h) \in \mathbb{R}^2$ as a 2-dimensional vector and switch into polar coordinates: $(r\cos(\phi), r\sin(\phi))$. Note that $g^2+h^2 = r^2$ and there is no angle dependence in the integral. Accordingly the volume element changes from $\mathrm{d}g\mathrm{d}h$ to $r\mathrm{d}r\mathrm{d}\phi$ and we obtain

$$M = \frac{1}{8\pi} \int_0^{2\pi} \int_0^{\infty} \left| r^2-2 \right| \mathrm{e}^{-r^2/2} r \mathrm{d}r \mathrm{d}\phi = \frac{1}{4} \int_0^{\infty} \left| r^2-2 \right| r\mathrm{e}^{-r^2/2} r \mathrm{d}r,$$

where we have carried out the integral over the angle ϕ which cancels the $1/(2\pi)$-term in front of the expression. Next, we note that the sign of the absolute value changes as we change the integration range. For $r \in \left[0, \sqrt{2}\right]$, we have $|r^2-2| = (2-r^2)$ while $|r^2-2| = (r^2-2)$ for $r \in \left[\sqrt{2}, \infty\right)$. This implies

$$\begin{aligned} M =& \frac{1}{4} \int_0^{\infty} \left| r^2-2 \right| r\mathrm{e}^{-r^2/2} r \mathrm{d}r \\ =& \frac{1}{4} \left(\int_0^{\sqrt{2}} \left(2-r^2\right) r\mathrm{e}^{-r^2/2} \mathrm{d}r + \int_{\sqrt{2}}^{\infty} \left(r^2-2\right) r\mathrm{e}^{-r^2/2} \mathrm{d}r \right) \\ =& \frac{1}{4} \left(2\int_0^{\sqrt{2}} r\mathrm{e}^{-r^2/2} \mathrm{d}r - \int_0^{\sqrt{2}} r^3\mathrm{e}^{-r^2/2} \mathrm{d}r + \int_{\sqrt{2}}^{\infty} r^3\mathrm{e}^{-r^2/2} - 2\int_{\sqrt{2}}^{\infty} r\mathrm{e}^{-r^2/2} \mathrm{d}r \right). \end{aligned} \tag{72}$$

These four remaining integrals can be determined from the following well-known Gaussian integration formulas for $a \leq b$:

$$\int_a^b r\mathrm{e}^{r^2/2} = \mathrm{e}^{-a^2/2} - \mathrm{e}^{-b^2/2} \quad \text{and} \quad \int_a^b r^3\mathrm{e}^{-r^2/2} \mathrm{d}r = (a^2+2)\mathrm{e}^{-a^2/2} - (b^2+2)\mathrm{e}^{-b^2/2}.$$

The limit $b \to \infty$ produces a vanishing contributions, because $\lim_{b\to\infty} \mathrm{e}^{-b^2/2} = 0$ and $\lim_{b\to\infty}(b^2+2)\mathrm{e}^{-b^2/2} = 0$ Inserting these values into Equation (72) yields

$$\begin{aligned} M =& \frac{1}{4} \left(2\left(\mathrm{e}^{-0} - \mathrm{e}^{-1}\right) - \left((2+0)\mathrm{e}^{-0} - (2+2)\mathrm{e}^{-1}\right) + \left((2+2)\mathrm{e}^{-1} - 0\right) - 2\left(\mathrm{e}^{-1} - 0\right) \right) \\ =& \frac{1}{4} \left(2 - 2/\mathrm{e} - 2 + 4/\mathrm{e} + 4/\mathrm{e} - 2/\mathrm{e}\right) = \frac{1}{\mathrm{e}}. \end{aligned}$$

□

A.2 Lipschitz constants for function evaluations and TV distances

In this appendix section, we derive Lipschitz constants for the functions whose expectation value value we computed in the previous subsection. We will see that these Lipschitz constants are small ($L = 2$ and $L = 1$, respectively) which is the only requirement we need to invoke Levy's lemma to show exponential concentration around these expectation values.

Lemma 22: *To start fix* $\phi : \{1, \dots, D\} \to [-1, 1]$ *and reinterpret* $P_U[\phi] = \sum_{d=1}^{D} \phi(d) |\langle d|U|\psi_0\rangle|^2$ *as a function in the pure state* $|\psi\rangle = U|\psi_0\rangle$*, namely* $P_\phi(|\psi\rangle) = \sum_{d=1}^{D} \phi(d) |\langle d|\psi\rangle|^2$*. This function has Lipschitz constant* $L = 2$*, namely*

$$|P_\phi(|\psi\rangle) - P_\phi(|\chi\rangle)| \leq 2 \, \| |\psi\rangle - |\chi\rangle \|_{\ell_2} \quad \textit{for all pure states } |\psi\rangle, |\chi\rangle \in \mathbb{C}^D.$$

Proof. Let us start by rewriting $P_\phi(|\psi\rangle)$ as a linear function in the (pure) density matrix $|\psi\rangle\langle\psi|$:

$$P_\phi(|\psi\rangle) = \sum_{d=1}^{D} \phi(d) \, |\langle d|\psi\rangle|^2 = \langle\psi| \left(\sum_{d=1}^{D} \phi(d) |d\rangle\langle d| \right) |\psi\rangle = \operatorname{tr}\left(\Phi \, |\psi\rangle\langle\psi|\right), \tag{73}$$

where we have introduced the diagonal $D \times D$ matrix $\Phi = \sum_{d=1}^{D} \phi(d) |d\rangle\langle d|$. Note that this matrix has operator norm $\|\Phi\|_\infty = \max_{1 \leq d \leq D} |\phi(d)| \leq 1$, because the function values $\phi(d)$ are confined to $[-1, 1]$ by assumption. The matrix Hoelder inequality then implies

$$\begin{aligned} |P_\phi(|\psi\rangle) - P_\phi(|\chi\rangle)| = |\operatorname{tr}\left(\Phi \, |\psi\rangle\langle\psi|\right) - \operatorname{tr}\left(\Phi \, |\chi\rangle\langle\chi|\right)| &= |\operatorname{tr}\left(\Phi \;\; (|\psi\rangle\langle\psi| - |\chi\rangle\langle\chi|)\right)| \\ &\leq \|\Phi\|_\infty \, \| |\psi\rangle\langle\psi| - |\chi\rangle\langle\chi| \|_1, \end{aligned}$$

where $\| \cdot \|_1$ denotes the trace norm. Since $\|\Phi\|_\infty \leq 1$, the claim – Lipschitz constant $L = 2$ – then follows from the following relation between trace distance of pure states and Euclidean distance of the state vectors involved:

$$\frac{1}{2} \, \| |\psi\rangle\langle\psi| - |\chi\rangle\langle\chi| \|_1 \leq \| |\psi\rangle - |\chi\rangle \|_2 \quad \text{for pure states } |\psi\rangle, |\chi\rangle \in \mathbb{C}^D. \tag{74}$$

Let us now derive this useful relation. One way is to use the Fuchs-van de Graaf inequalities (which are tight for pure states) to relate the trace distance to a pure state fidelity:

$$\frac{1}{2} \, \| |\psi\rangle\langle\psi| - |\chi\rangle\langle\chi| \|_1 = \sqrt{1 - F\left(|\psi\rangle, |\chi\rangle\right)} = \sqrt{1 - |\langle\psi|\chi\rangle|^2}. \tag{75}$$

Finally, we can use $|\langle\psi|\chi\rangle| \leq 1$, as well as $\langle\psi|\psi\rangle = \langle\chi|\chi\rangle = 1$ and $2\,|\langle\psi|\chi\rangle| \geq 2\mathrm{Re}\left(\langle\psi|\chi\rangle\right) = \langle\psi|\chi\rangle + \langle\chi|\psi\rangle$ to obtain

$$\begin{aligned} \sqrt{1 - |\langle\psi|\chi\rangle|^2} &= \sqrt{1 + |\langle\psi|\chi\rangle|} \sqrt{1 - |\langle\psi|\chi\rangle|} \leq \sqrt{1+1} \sqrt{1 - \mathrm{Re}\left(\langle\psi|\chi\rangle\right)} \\ &= \sqrt{1 - \langle\psi|\chi\rangle - \langle\chi|\psi\rangle + 1} = \sqrt{\langle\psi|\psi\rangle - \langle\psi|\chi\rangle - \langle\chi|\psi\rangle + \langle\chi|\chi\rangle} \\ &= \sqrt{\left(\langle\psi| - \langle\chi|\right)\left(|\psi\rangle - |\chi\rangle\right)} = \| |\psi\rangle - |\chi\rangle \|_2. \end{aligned} \tag{76}$$

□

Lemma 23: *Let* $P_U(d) = |\langle d|U|\psi_0\rangle|^2$ *be the output distribution of a* D *level quantum system in the computational basis and let* $\mathcal{U}$ *be the* D *dimensional uniform distribution. The total variation distance between both distributions defines a function in the pure state* $|\psi\rangle = U|\psi_0\rangle$*, namely* $f_{\mathrm{TV}}(|\psi\rangle) = \frac{1}{2} \sum_d \left| |\langle d|\psi\rangle|^2 - 1/D \right|$*. This*

function has Lipschitz constant $L = 1$*, i.e.,*

$$|f_{\mathrm{TV}}(|\psi\rangle) - f_{\mathrm{TV}}(|\chi\rangle)| \leq \||\psi\rangle - |\chi\rangle\|_2 \quad \textit{for all pure states } |\psi\rangle, |\chi\rangle \in \mathbb{C}^D.$$

Proof. Let us start by rewriting the absolute value of the difference of different function values as

$$\begin{aligned}
&|f_{\mathrm{TV}}(|\psi\rangle) - f_{\mathrm{TV}}(|\chi\rangle)| = \\
&= \frac{1}{2}\left|\sum_{d=1}^{D}\left(\left||\langle d|\psi\rangle|^2 - 1/D\right| - \left||\langle d|\chi\rangle|^2 - 1/D\right|\right)\right| \\
&= \frac{1}{2}\left|\sum_{d=1}^{D}\left(\left|\left(|\langle d|\chi\rangle|^2 - 1/D\right) + \left(|\langle d|\psi\rangle|^2 - |\langle d|\chi\rangle|^2\right)\right| - \left||\langle d|\chi\rangle|^2 - 1/D\right|^2\right)\right| \\
&\leq \frac{1}{2}\sum_{d=1}^{D}\left||\langle d|\psi\rangle|^2 - |\langle d|\chi\rangle|^2\right|, \qquad (77)
\end{aligned}$$

where the last inequality follows from applying the triangle inequality to each summand in order to break up the two contributions in the first absolute value of the second line. The first contribution then cancels with the final term and we obtain the advertised display. We can now rewrite this new expression as

$$\frac{1}{2}\sum_{d=1}^{D}\left||\langle d|\psi\rangle|^2 - |\langle d|\chi\rangle|^2\right| = \frac{1}{2}\sum_{d=1}^{D}\left|\langle d|\left(|\psi\rangle\langle\psi| - |\chi\rangle\langle\chi|\right)|d\rangle\right|, \qquad (78)$$

which accumulates the sum of the absolute values of the diagonal entries of the (pure) state difference $(|\psi\rangle\langle\psi| - |\chi\rangle\langle\chi|)$. This sum of absolute diagonal entries is always smaller than the trace norm of the matrix in question[2]. This relation implies

$$|f_{\mathrm{TV}}(|\psi\rangle) - f_{\mathrm{TV}}(|\chi\rangle)| \leq \frac{1}{2}\sum_{d=1}^{D}\left|\langle d|\left(|\psi\rangle\langle\psi| - |\phi\rangle\langle\phi|\right)|d\rangle\right| \leq \frac{1}{2}\left\||\psi\rangle\langle\psi| - |\phi\rangle\langle\phi|\right\|_1, \qquad (79)$$

and the claim – Lipschitz constant $L = 1$ – now follows from reusing Equation (74) to convert this trace norm distance into a Euclidean distance of the state vectors involved. □

B Unitary designs

In this appendix we provide context for approximate unitary designs; a key tool for the results in this work. Moreover, we discuss recent bounds on the generation of designs by random quantum circuits. Recall the definition of the moment operator:

$$\Phi^{(t)}(\nu)(A) := \int U^{\otimes t} A (U^\dagger)^{\otimes t} \mathrm{d}\nu(U). \qquad (80)$$

[2]This relation is well known in matrix analysis and follows, for instance, from Helstrom's theorem.

Definition 6: (ε-Approximate Design) A probability distribution ν over $\mathrm{U}(D)$ is an ε-approximate unitary design if

$$\left\|\Phi^{(t)}(\nu) - \Phi^{(t)}(\mu_U)\right\|_\diamond \leq \frac{\varepsilon}{D^t}, \tag{81}$$

where $\|\bullet\|_\diamond$ denotes the diamond norm, or channel distinguishability, defined as the stabilized $1 \to 1$ norm.

In this work we will only be concerned with averages over states, that is the case $A = (|\psi\rangle\langle\psi|)^{\otimes t}$. In this case we have the standard formula (see e.g. [Har13]).

$$\Phi^{(t)}(\mu_U)(A) = \int U^{\otimes t} \left(|\psi\rangle\langle\psi|\right)^{\otimes t} (U^\dagger)^{\otimes t} \mathrm{d}\mu_U(U) = \frac{P_{\mathrm{sym},t}}{\binom{D+t-1}{t}}, \tag{82}$$

where $P_{\mathrm{sym},t}$ we denote the projector onto the symmetric subspace $S^t(\mathbb{C}^D)$. With the above definition of an approximate unitary design, we obtain that for ν an ε-approximate unitary design, we have

$$\left\|\Phi^{(t)}(\nu)\left(\left(|\psi\rangle\langle\psi|\right)^{\otimes t}\right) - \frac{P_{\mathrm{sym},t}}{\binom{D+t-1}{t}}\right\|_1 \leq \frac{\varepsilon}{D^t}, \tag{83}$$

where $\|\bullet\|_1$ denotes the Schatten 1-norm, or trace norm.

The key result for the following is that random quantum circuits are in fact approximate unitary t-designs in polynomial depth [BHH16; HL09; Haf22]. These bounds come with large explicit constants. For small values of $t = 2, 4$, we even have good explicit constants. We present the bound from Haferkamp [Haf22]:

Theorem 24: *For $n \geq \left\lceil 2\log_2(4t) + 1.5\sqrt{\log_2(4t)}\right\rceil$, random quantum circuits in a brickwork architecture are ε-approximate unitary t-designs in depth*

$$T \geq C \ln^5(t) t^{4+3\frac{1}{\sqrt{\log_2(t)}}} (2nt + \log_2(1/\varepsilon)), \tag{84}$$

where C can be taken to be 10^{13}.

Note that the large constants are likely an artefact of the proof technique based on the martingale technique [Nac96] in [BHH16; Haf22], which focus on the scaling in t.

Using instead finite-size criteria [Kna88] combined with numerics, one can greatly improve these constants for $t \leq 5$. Compare [HH21, Table I]. It is likely that we could obtain comparable constants for $t = 8$ as well. Unfortunately, this seems to require numerics for daunting system sizes.

C Moment calculations

C.1 Haar moments

To begin we give explicit formulas for Haar random moments. We will make use of the following standard formula repeatedly:

$$\begin{aligned}\mathop{\mathbf{E}}_{\psi\sim\mu_S}\left[|\langle\psi|\phi\rangle|^{2t}\right] &= \mathop{\mathbf{E}}_{\psi\sim\mu_S}\left[\mathrm{Tr}\left[\left(|\psi\rangle\langle\psi|\right)^{\otimes t}\left(|\phi\rangle\langle\phi|\right)^{\otimes t}\right]\right] \\ &= \mathrm{Tr}\left[\mathop{\mathbf{E}}_{\psi\sim\mu_S}\left[|\psi\rangle\langle\psi|\right]^{\otimes t}\left(|\phi\rangle\langle\phi|\right)^{\otimes t}\right] \\ &= \mathrm{Tr}\left[\frac{P_{\mathrm{sym},t,D}}{\binom{D+t-1}{t}}\left(|\phi\rangle\langle\phi|\right)^{\otimes t}\right] \\ &= \binom{D+t-1}{t}^{-1}.\end{aligned}$$

In fact, we will need a more general formula for the proof of Theorem 9, which we state as the following lemma.

Lemma 25: *Let $|i_1\rangle,\ldots,|i_k\rangle$ with $i_1,\ldots,D\in\{1,\ldots,D\}$ be mutually orthogonal state vectors and $\lambda=(\lambda_1,\ldots,\lambda_k)$ a partition of t for $t\leq D$. Then, we find the formula*

$$\mathop{\mathbf{E}}_{|\psi\rangle\sim\mu_S}\left[\prod_{l=1}^{k}|\langle\psi|i_l\rangle|^{2\lambda_l}\right]=\frac{\prod_{l=1}^{k}\lambda_l!}{D\cdots(D+t-1)}. \tag{85}$$

Proof. The proof follows directly from the following calculation:

$$\begin{aligned}\mathop{\mathbf{E}}_{|\psi\rangle\sim\mu_S}\prod_{l=1}^{k}|\langle\psi|i_l\rangle|^{2\lambda_l} &= \mathop{\mathbf{E}}_{\psi\sim\mu_S}\left[\mathrm{Tr}\left[\left(|\psi\rangle\langle\psi|\right)^{\otimes t}\bigotimes_{l=1}^{k}\left(|i_l\rangle\langle i_l|\right)^{\otimes\lambda_l}\right]\right] \\ &= \binom{D+t-1}{t}^{-1}\mathrm{Tr}\left[P_{\mathrm{sym},t,D}\bigotimes_{l=1}^{k}\left(|i_l\rangle\langle i_l|\right)^{\otimes\lambda_l}\right] \\ &= \frac{1}{D\cdots(D+t-1)}\sum_{\pi\in S_t}\mathrm{Tr}\left[r(\pi)\bigotimes_{l=1}^{k}\left(|i_l\rangle\langle i_l|\right)^{\otimes\lambda_l}\right],\end{aligned} \tag{86}$$

where we used the notation r for the representation of S_t that, for each permutation $\pi\in S_t$, permutes the t tensor factors according to π:

$$r(\pi)|i_1\rangle\otimes\cdots\otimes|i_t\rangle:=|i_{\pi^{-1}(1)}\rangle\otimes\cdots\otimes|i_{\pi^{-1}(t)}\rangle. \tag{87}$$

Moreover, we used the formula $P_{\mathrm{sym},t,D}=\frac{1}{t!}\sum_\pi r(\pi)$. Notice that

$$\mathrm{Tr}\left[r(\pi)\bigotimes_{l=1}^{k}\left(|i_l\rangle\langle i_l|\right)^{\otimes\lambda_l}\right]=\begin{cases}1 & \text{if } \pi\in S_{\lambda_1}\times\ldots\times S_{\lambda_k}\\ 0 & \text{else.}\end{cases} \tag{88}$$

Hence, we find

$$\begin{aligned}\mathop{\mathbf{E}}_{|\psi\rangle\sim\mu_S}\prod_{l=1}^{k}|\langle\psi|i_l\rangle|^{2\lambda_l} &= \frac{1}{D\cdots(D+t-1)}|S_{\lambda_1}|\cdots|S_{\lambda_k}| \\ &= \frac{\prod_{l=1}^{k}\lambda_l!}{D\cdots(D+t-1)}.\end{aligned} \tag{89}$$

□

In the special case $D = 2^n$ we thus obtain the explicit formulas for the first and second moment:

$$\mathop{\mathbf{E}}_{U\sim\mu_U}[P_U(x)] = \frac{1}{2^n}, \tag{90}$$

$$\mathop{\mathbf{E}}_{U\sim\mu_U}[P_U(x)P_U(y)] = \frac{1}{2^n(2^n+1)}[1+\delta_{x,y}]\,. \tag{91}$$

C.2 Restricted depth moments

Next, we state bounds on the first two moments over brickwork random quantum circuits of depth d:

Lemma 26: (Moments over circuits of restricted depth – adapted from [BCG21]) *For μ_C the measure over brickwork random quantum circuits on n qubits of depth d, it holds*

$$\mathop{\mathbf{E}}_{U\sim\mu_C}[P_U(x)] = \frac{1}{2^n}\,, \tag{92}$$

$$\mathop{\mathbf{E}}_{U\sim\mu_C}[P_U(x)P_U(y)] \leq \frac{1}{2^{2n}}(1+\delta_{x,y})\left[1+n\left(\frac{4}{5}\right)^d\right]. \tag{93}$$

where the bound in Equation (93) holds for $d \geq \frac{\log n}{\log 5/4}$.

Proof. We note that μ_C is an exact 1-design at any depth d[3]. Hence, the first moment is the same as in Equation (90) i.e.

$$\mathop{\mathbf{E}}_{U\sim\mu_C}[P_U(x)] = \mathop{\mathbf{E}}_{U\sim\mu_U}[P_U(x)] = \frac{1}{2^n}. \tag{94}$$

To obtain the second moment given in Equation (93), we adapt and modify a calculation presented in Section 6.3 of [BCG21]. Specifically, using a mapping to a statistical mechanics model, the second moment with respect to the random circuit,

[3]In fact, already a single layer of randomly drawn unitary gates forms an exact 1-design. This is because this layer contains as a subgroup the Pauli group which is known to form an exact 1-design. It follows from the invariance of the Haar measure under left multiplication that random unitary circuits form an exact 1-design also for $d \geq 1$.

$\mathbf{E}_{\mu_C}[P_U(x)P_U(y)]$, can be expressed as a partition function. The value of this partition function can then be bounded by counting domain walls. In Section 6.3 of Ref. [BCG21], this technique was already used to obtain an upper bound on $\mathbf{E}_{\mu_C}[P_U(x)^2]$, for random circuits of depth $d \geq \frac{\log n}{\log 5/4}$. More specifically, Ref. [BCG21] has obtained the upper bound

$$\mathop{\mathbf{E}}_{U\sim\mu_C}\left[P_U(x)^2\right] \leq \left(1+\left(\frac{4}{5}\right)^d\right)^{n/2} \mathop{\mathbf{E}}_{U\sim\mu_U}\left[P_U(x)^2\right], \tag{95}$$

which is given in terms of the Haar expectation value $\mathbf{E}_{\mu_U}[P_U(x)^2]$, and indeed converges to this Haar value in the infinite circuit depth-limit $d \to \infty$. A similar analysis allows us to obtain the following bound on the expectation value of the cross terms $P_U(x)P_U(y)$,

$$\mathop{\mathbf{E}}_{U\sim\mu_C}[P_U(x)P_U(y)] \leq \left(1+\left(\frac{4}{5}\right)^d\right)^{n/2} \mathop{\mathbf{E}}_{U\sim\mu_U}[P_U(x)P_U(y)]. \tag{96}$$

Note that this upper bound is also given in terms of the corresponding Haar value $\mathbf{E}_{\mu_U}[P_U(x)P_U(y)]$. We use the second moment already calculated in Equation (91). Finally, we bound the prefactor: By Bernoulli's inequality, we have that $(1+x^d)^n \leq e^{nx^d}$. For $d \geq \frac{\log n}{\log 5/4}$ and $x < 1$ we can then use the convexity of the exponential function $e^y \leq (1-y)e^0 + ye^1$ to obtain $e^{nx^d} \leq 1 - nx^d + enx^d \leq 1 + 2nx^d$. This allows us to show that

$$\left(1+\left(\frac{4}{5}\right)^d\right)^{n/2} \leq 1 + n\left(\frac{4}{5}\right)^d. \tag{97}$$

Substituting Equations (91) and (97) into Equation (96) then yields Equation (93). □

D Random Clifford unitaries

The Clifford group forms a 3-design. Therefore, we can carry over the bounds on $\mathfrak{f}$ obtained via Chebychev and hence second moments from the global Haar measure to the uniform measure over global Clifford operations. The same analogy holds for local Haar random unitaries and local Clifford unitaries. Thus, the bounds on $\mathfrak{f}$ from the restricted depth moments from Section C.2 also hold for restricted depth Clifford circuits. The key difference between the case of Clifford and Haar random unitaries lies thus in the far from uniform behavior. This is emphasized in the following lemma.

Lemma 27: *The probability that the output distribution of a uniformly random*

global Clifford circuit on n qubits is the uniform distribution is given by

$$\mathop{\mathbf{Pr}}_{U\sim \mathrm{Cl}(2^n)}\left(P_U = \mathcal{U}\right) = \frac{1}{\prod_{i=1}^{n}\left(1+\frac{1}{2^i}\right)}\,. \tag{98}$$

In particular, it asymptotically approaches

$$\mathop{\mathbf{Pr}}_{U\sim \mathrm{Cl}(2^n)}\left(P_U = \mathcal{U}\right) \overset{n\to\infty}{\to} 0.41942244... \tag{99}$$

from above and for any number of qubits n, the probability is larger than 0.41.

Thus, even though "non-trivial" learning is hard for random Clifford unitary output distributions as characterized by $\mathfrak{f}$, the trivial learning algorithm, which always returns $\mathcal{U}$, will succeed with probability larger than 0.41 over the uniformly drawn $U \sim \mathrm{Cl}(2^n)$.

Proof. The result of drawing a uniformly random Clifford unitary $U \sim \mathrm{Cl}(2^n)$ and applying to $|0\rangle^{\otimes n}$ is a uniformly random stabilizer state.

The number of n-qubit stabilizer states is given by [AG04]

$$|\mathcal{S}_n| = 2^n \prod_{i=i}^{n}\left(2^i+1\right) = 2^{n+n(n+1)/2}\prod_{i=1}^{n}\left(1+\frac{1}{2^i}\right) \tag{100}$$

The number of n-qubit stabilizer states giving rise to the uniform distribution is given by

$$|\mathcal{S}_n^n| = 2^n \cdot 2n(n+1)/2 = 2^{n+n(n+1)/2} \tag{101}$$

This follows from [KG15]. In particular, Corollary 2 in Ref. [KG15] gives a formula for the number of stabilizer states with pre-described inner product with respect to a fixed reference stabilizer state. For our purposes, it suffices to take as reference state the all-zero state $|0^n\rangle$ and find the number of stabilizer states $|\psi\rangle$ such that $|\langle\psi|0^n\rangle| = 2^{-n}$. Such states are precisely the n-qubit stabilizer states giving rise to the uniform distribution.

Hence,

$$\mathop{\mathbf{Pr}}_{U\sim \mathrm{Cl}(2^n)}\left(P_U = \mathcal{U}\right) = \frac{2^{n+n(n+1)/2}}{2^{n+n(n+1)/2}\prod_{i=1}^{n}\left(1+\frac{1}{2^i}\right)} = \frac{1}{\prod_{i=1}^{n}\left(1+\frac{1}{2^i}\right)} \tag{102}$$

The asymptotic behavior of this product for $n\to\infty$ is found in [Ben21]. □

E Deterministic algorithms

The aim of this appendix is to give a detailed proof of Theorem 1 which is restated as Theorem 30. We follow a similar strategy as Feldman in [Fel17] by proving the result for learning via a reduction to a suitably chosen decision problem.

Problem 2: (Decide $\mathcal{D}$ versus Q) Let $\mathcal{D}$ be some distribution class and Q some fixed reference distribution. The task decide $\mathcal{D}$ versus Q is defined as, given access to an unknown $P \in \mathcal{D} \cup \{Q\}$ to decide whether "$P = Q$" or "$P \in \mathcal{D}$".

We connect the query complexity of learning with the query complexity of deciding by the following lemma. .

Lemma 28: (Learning is as hard as deciding) *Let $\mathcal{D}$ be a distribution class and let Q be such that $\mathrm{d_{TV}}(P, Q) > \epsilon + \tau$ for all $P \in \mathcal{D}$. Let $0 < \tau \leq \epsilon \leq 1$. Let $\mathcal{A}$ be a deterministic algorithm that ϵ-learns $\mathcal{D}$ from q many τ accurate statistical queries. Then there exists a deterministic algorithm that decides $\mathcal{D}$ versus Q from $q+1$ many τ accurate statistical queries.*

Proof. We run $\mathcal{A}$ on the unknown distribution $P \in \mathcal{D} \cup \{Q\}$ and obtain either

- a representation of some P' which is ϵ close to P if $P \in \mathcal{D}$, or
- anything if $P = Q$.

In case we do not receive a representation of any distribution return "$P = Q$". Now, assume we receive a representation of some distribution P'. Using this representation compute whether P' is ϵ close to any distribution in $\mathcal{D}$. While this step is computationally costly, it does not require any further queries to $\mathtt{Stat}(P)$. If there does not exists such a distribution in $\mathcal{D}$ which is ϵ-close to P', return "$P = Q$".

Now assume there exists an $H \in \mathcal{D}$ such that $\mathrm{d_{TV}}(P', H) < \epsilon$. To assure, that $\mathcal{A}$ is not biased towards returning distributions close to $\mathcal{D}$ if it fails, compute the set S that maximizes the total variation distance between Q and P', $|P'(S) - Q(S)| = \mathrm{d_{TV}}(P', Q)$. Denote by $\phi = \mathbb{1}_S$ the characteristic function on S and query $v_\phi \leftarrow \mathtt{Stat}_\tau(P)[\phi]$. If $|Q[\phi] - v_\phi| \leq \tau$ return "$P = Q$", else return "$P \in \mathcal{D}$".

We analyze the algorithm for each case separately. Common to both is that the algorithm makes, by definition, at most $q + 1$ statistical queries.

To begin with assume $P \in \mathcal{D}$. By the correctness of $\mathcal{A}$ we receive a representation of some P' that is at most ϵ far from P. By assumption, for any $H \in \mathcal{D}$ it holds $\mathrm{d_{TV}}(H, Q) > \epsilon + \tau$. Then, by the definition of S using the reverse triangle inequality we find

$$|Q[\phi] - v_\phi| \geq ||Q[\phi] - P[\phi]| - |P[\phi] - v_\phi|| > |\epsilon + \tau - \tau| = \epsilon \geq \tau \,. \tag{103}$$

Hence, we correctly decide "$P \in \mathcal{D}$".

For the other case assume $P = Q$. If $\mathcal{A}$ does not return a valid representation or, if $\mathcal{A}$ returns a representation of some P' that is more than ϵ far away from any distribution in $\mathcal{D}$, we know, by the correctness of $\mathcal{A}$, that it must hold $P = Q$. It remains to show the last step. Assume there is an $H \in \mathcal{D}$ which is at most ϵ far from P'. Then, by assumption, for every ϕ it must hold $|Q[\phi] - v_\phi| \leq \tau$. Hence, we correctly decide "$P = Q$". □

Using Theorem 28 it will be sufficient to bound the query complexity of deciding. This is achieved by the next lemma.

Lemma 29: (Hardness of deciding, deterministic version) *Let $\mathcal{A}$ be a deterministic algorithm that decides $\mathcal{D}$ versus Q from q many τ-accurate statistical queries, then for any measure μ over $\mathcal{D}$ it holds*

$$q \geq \left(\max_{\phi} \mathop{\mathbf{Pr}}_{P\sim\mu} \left[|P[\phi] - Q[\phi]| > \tau \right] \right)^{-1} . \tag{104}$$

Proof. Assume we run $\mathcal{A}$ and answer every query ϕ by $Q[\phi]$. Denote by $\phi_1, \ldots, \phi_q$ the queries made. Assume for a contradiction, that for some $P \in \mathcal{D}$ there does not exist any distinguishing query. Then, the responses $Q[\phi_i]$ for $i = 1, \ldots, q$ would be valid responses for some $\mathtt{Stat}_\tau(P)$ contradicting the assumption that $\mathcal{A}$ successfully decides whether $P = Q$. Thus, for any $P \in \mathcal{D}$ there must exist at least one i that distinguishes Q from P. In particular,

$$\begin{aligned} 1 &= \mathop{\mathbf{Pr}}_{P\sim\mu} \left[\exists i,\, |P[\phi_i] - Q[\phi_i]| > \tau \right] && (105)\\ &\leq \sum_{i=1}^{q} \mathop{\mathbf{Pr}}_{P\sim\mu} \left[|P[\phi_i] - Q[\phi_i]| > \tau \right] && (106)\\ &\leq q \max_{\phi} \mathop{\mathbf{Pr}}_{P\sim\mu} \left[|P[\phi_i] - Q[\phi_i]| > \tau \right] , && (107) \end{aligned}$$

which completes the proof. □

We are now set to prove our bound for the deterministic average case query complexity. Note, that Theorem 28 holds for learners that learn all of $\mathcal{D}$. Thus, the core of the remaining proof will be to translate the implications on worst to average case learners.

Theorem 30: (Theorem 1 restated) *Suppose there is a deterministic algorithm $\mathcal{A}$ that ϵ-learns a β fraction of $\mathcal{D}$ with respect to μ from q many τ-accurate statistical queries. Then for any Q it holds*

$$q + 1 \geq \frac{\beta - \mathbf{Pr}_{P\sim\mu} \left[\mathrm{d_{TV}}(P, Q) \leq \epsilon + \tau \right]}{\max_{\phi} \mathbf{Pr}_{P\sim\mu} \left[|P[\phi] - Q[\phi]| > \tau \right]} , \tag{108}$$

where again, the max *is over all functions $\phi : X \to [-1, 1]$.*

Proof. Let $\mathcal{D}' \subseteq \mathcal{D}$ with $\mu(\mathcal{D}') = \beta$ be a set on which $\mathcal{A}$ is successful. Define

$$\mathcal{D}_Q = \{ P \in \mathcal{D}' : \mathrm{d_{TV}}(P, Q) > \epsilon + \tau \}$$

and let μ_Q be the measure μ conditioned on $\mathcal{D}_Q$, such that $\mu_Q(P) = \mu(P \mid P \in \mathcal{D}_Q)$. Then, by the definition of the conditional probability,

$$\beta - \mathop{\mathbf{Pr}}_{P\sim\mu} \left[\mathrm{d_{TV}}(P, Q) \leq \epsilon + \tau \right] \leq \mu(\mathcal{D}_Q) . \tag{109}$$

Therefore, for any ϕ it holds

$$\begin{aligned}\Pr_{P\sim\mu_Q}\left[|P[\phi]-Q[\phi]|>\tau\right] &= \Pr_{P\sim\mu}\left[|P[\phi]-Q[\phi]|>\tau \mid P\in\mathcal{D}_Q\right] \\ &\leq \frac{\mathbf{Pr}_{P\sim\mu}\left[|P[\phi]-Q[\phi]|>\tau\right]}{\beta-\mathbf{Pr}_{P\sim\mu}\left[\mathrm{d_{TV}}(P,Q)\leq\epsilon+\tau\right]}\end{aligned} \tag{110}$$

where we used the definition of the conditional probability.

The claim then follows from the observation that the average learner $\mathcal{A}$ for $\mathcal{D}$ implies a learner for $\mathcal{D}_Q$, the complexity of which can be bounded by the complexity of deciding $\mathcal{D}_Q$ vs Q via the reduction Theorem 32. We obtain a bound for the latter from Theorem 33 applying the measure μ_Q. □

Before we end this section we state a variant of this bound due to Feldman to discuss the connection to Theorem 30. We restate his proof adapted to our notation for the sake of completeness.

Lemma 31: (Variation of Lemma C.2 from [Fel17] for deterministic algorithms) *Suppose there is a deterministic algorithm $\mathcal{A}$ that ϵ-learns a β fraction of $\mathcal{D}$ with respect to μ from q many τ-accurate statistical queries. Then for any Q it holds*

$$q\geq\frac{\beta-\sup_D\mathbf{Pr}_{P\sim\mu}\left[\mathrm{d_{TV}}(P,D)<\epsilon\right]}{\max_\phi\mathbf{Pr}_{P\sim\mu}\left[|P[\phi]-Q[\phi]|>\tau\right]}\,, \tag{111}$$

where the max *is over all functions $\phi: X\to[-1,1]$ and the* sup *is over all distributions D over the domain X.*

Proof. Denote by $\mathcal{D}'\subseteq\mathcal{D}$ the subset of size $\mu(\mathcal{D}')=\beta$ on which $\mathcal{A}$ successfully ϵ-learns from q queries. We run $\mathcal{A}$ and answer every query ϕ by $Q[\phi]$. By assumption $\mathcal{A}$ makes q queries $\phi_1,\ldots,\phi_q$ to Q and, without loss of generality, we assume that the algorithm returns some distribution D. Now, let P be any distribution in $\mathcal{D}'$ at least ϵ-far from D. Exactly as in the proof of Theorem 29, there must exists at least one query function ϕ_i that distinguishes Q from P. In particular, it must hold

$$\begin{aligned}\beta-\Pr_{P\sim\mu}\left[\mathrm{d_{TV}}(P,D)<\epsilon\right] &\leq \Pr_{P\sim\mu}\left[\exists i:\ |P[\phi_i]-Q[\phi_i]|>\tau\right] \\ &\leq \sum_{i=1}^{q}\Pr_{P\sim\mu}\left[|P\phi_i]-Q[\phi_i]>\tau|\right] \\ &\leq q\max_\phi\Pr_{P\sim\mu}\left[|P[\phi]-Q[\phi]|>\tau\right]\,.\end{aligned} \tag{112}$$

Now assume that, after interacting with Q, the algorithm does not return any valid distribution. Then, again by $\mathcal{A}$'s determinism, for *any* $P\in\mathcal{D}'$ there must exist a distinguishing query ϕ_i that distinguishes Q from P. The claim then follows by taking the supremum over $D\in\mathcal{D}_X$ to bound the ϵ-ball around the unknown D. □

Note 2: We want to highlight that Theorem 31 is tight with respect to β: The trivial algorithm, which makes zero queries and always outputs that D which maximizes

the open ϵ-ball will, with probability $\mu(B_\epsilon(D))$ over $P \sim \mu$ be correct.

Note 3: As stated above, Theorem 31 gives the optimal lower bound with respect to β. However, in some cases directly bounding the weight of all ϵ-balls may not be convenient. In Section G we give a general recipe for obtaining bounds for all ϵ-balls just from the two ingredients used in Theorem 30: The maximally distinguishable fraction and the mass of the ball around the reference distribution. While this strategy is straight forward, the bounds obtained are slightly worse than those obtained by directly invoking Theorem 30, which is why we stick to the latter result in this work.

F Quantum and probabilistic algorithms

In this appendix we will detail the connection between statistical query learning via deterministic and random algorithms. Throughout the section we will refer by random algorithm to both classical probabilistic as well as quantum algorithms.

The randomized average case query complexity for ϵ-learning $\mathcal{D}$ with respect to the probability measure μ depends on the two parameters α and β, where

- α denotes the success probability with respect to the internal randomness of the learning algorithm and
- β denotes the fraction of distributions in $\mathcal{D}$ measured with respect to μ on which the learning algorithm must be successful.

The aim of this appendix is to bound the randomized average case query complexity for ϵ-learning $\mathcal{D}$ by (c.f. Theorem 34)

$$q \geq \frac{2 \cdot \left(\alpha - \frac{1}{2}\right) \cdot \left(\beta - \mathbf{Pr}_{P\sim\mu}\left[\mathrm{d_{TV}}(P,Q) \leq \epsilon + \tau\right]\right)}{\max_\phi \mathbf{Pr}_{P\sim\mu}\left[|P[\phi] - Q[\phi]| > \tau\right]}\,. \tag{113}$$

Thus, the randomized average case query complexity of ϵ-learning is bounded by the same bounds as the deterministic average case query complexity up to a prefactor $2(\alpha - 1/2)$, which becomes trivial for $\alpha \leq 1/2$.

We will follow the same strategy as in the deterministic case laid out in Section E. The main difference is that we need a bound on the decision problem Problem 2 for random algorithms. The subtlety, why the arguments from Theorem 29 fail, is that a random algorithm does not need to make a distinguishing query to solve the problem. Rather, it may guess the correct solution using its internal randomness. The main technical ingredient of this Appendix is thus a result by Feldman which, first bounding the probability of guessing correctly, bounds the statistical query complexity for Problem 2 which is stated as Theorem 33.

To begin with, we can follow the same strategy as before to reduce deciding to learning, also in the random setting.

Lemma 32: (Learning is as hard as deciding) *Let $\mathcal{D}$ be a distribution class and let Q be such that $\mathrm{d}_{\mathrm{TV}}(P,Q) > \epsilon + \tau$ for all $P \in \mathcal{D}$. Let $0 < \tau \leq \epsilon \leq 1$. Let $\mathcal{A}$ be an algorithm that ϵ-learns $\mathcal{D}$ from q many τ accurate statistical queries with probability α over its internal randomness. Then there exists an algorithm that, with probability α over its internal randomness, decides $\mathcal{D}$ versus Q from $q+1$ many τ accurate statistical queries.*

Proof. The proof is identical to that of Theorem 28. The only difference is that the learner $\mathcal{A}$ only succeeds with probability α, which leads to the decider only succeeding with probability α. The reduction itself however is deterministic and, as such, does not change the statistics. □

We now state the result by Feldman on the randomized statistical query complexity of Problem 2. For the sake of self consistency we provide the proof adapted to our notation.

Lemma 33: (Hardness of deciding. Taken from Theorem 3.9 from [Fel17]) *Let $\mathcal{A}$ be a random algorithm that decides $\mathcal{D}$ versus Q with probability at least α from q many τ accurate statistical queries. Then, for any measure μ over $\mathcal{D}$ it holds*

$$q \geq \frac{2 \cdot (\alpha - \frac{1}{2})}{\max_{\phi} \mathbf{Pr}_{P\sim\mu} \left[|P[\phi] - Q[\phi]| > \tau\right]} \,. \tag{114}$$

Proof. Let $\mathcal{A}$ be an algorithm that decides $\mathcal{D}$ vs. Q with probability α over its internal randomness. We run $\mathcal{A}$ and, on every query ϕ return $Q[\phi]$. Denote by $\phi_1, \ldots, \phi_q$ the queries made. These queries then can be interpreted as random variables with respect to $\mathcal{A}$'s randomness. Let $P \in \mathcal{D}$ and denote by

$$p(P) = \mathop{\mathbf{Pr}}_{\mathcal{A}} \left[\exists i \,:\, |P[\phi_i] - Q[\phi_i]| > \tau\right] \,. \tag{115}$$

We now show that $p(P) \geq 2(\alpha - 1/2)$: By the correctness of $\mathcal{A}$ the corresponding output will be "$P \in \mathcal{D}$" with probability at most $1-\alpha$. For the sake of contradiction assume $p(P) < 2(\alpha - 1/2)$. Thus, when run on P, for some valid answers $\mathcal{A}$ will still return "$P = Q$" with probability at least $> 1 - p(P) - (1-\alpha) > 1 - 2\alpha + 1 - 1 + \alpha = 1 - \alpha$. Since this probability is bounded by $1-\alpha$ we find a contradiction $\alpha < \alpha$. Thus $p(P) \geq 2(\alpha - 1/2)$.

The remainder now follows from the union bound as in Theorem 29:

$$\begin{aligned} 2(\alpha - 1/2) \leq p(P) &= \mathop{\mathbf{Pr}}_{\mathcal{A}, P\sim\mu} \left[\exists i \,:\, |P[\phi_i] - Q[\phi_i]| > \tau\right] \\ &\leq \sum_{i=1}^{q} \mathop{\mathbf{Pr}}_{\mathcal{A}, P\sim\mu} \left[|P[\phi_i] - Q[\phi_i]| > \tau\right] \\ &\leq q \max_{\phi} \mathop{\mathbf{Pr}}_{P\sim\mu} \left[|P[\phi] - Q[\phi]| > \tau\right] \,. \end{aligned} \tag{116}$$

□

Thus, following Section E, we can state the main theorem of this appendix.

Theorem 34: (Randomized average case query complexity of learning) *Let $\mathcal{A}$ be a random algorithm for average case ϵ-learning $\mathcal{D}$ with respect to μ with parameters α and β from q many τ accurate statistical queries. Then for any Q it holds*

$$q + 1 \geq 2 \cdot \frac{(\alpha - \frac{1}{2}) \cdot (\beta - \mathbf{Pr}_{P\sim\mu}\left[\mathrm{d_{TV}}(P, Q) \leq \epsilon + \tau\right])}{\max_\phi \mathbf{Pr}_{P\sim\mu}\left[|P[\phi] - Q[\phi]| > \tau\right]}\,. \tag{117}$$

Proof. The proof is identically to that of Theorem 30 using Theorem 32 and Theorem 33 instead of Theorem 28 and Theorem 29. □

For the sake of context relating to the discussion in the end of Section E and [Fel17, Lemma C.2], we finish this appendix with two additional insights.

Note 4: If we restrict the result to probabilistic algorithms we can follow the argument from [Fel17, Lemma C.2]: One can make the randomness explicit by writing the random algorithm $\mathcal{A}$ as an ensemble of deterministic algorithms $\{\mathcal{A}_x\}$ with $x \sim \mathcal{A}$ the internal randomness. Then the randomized average case query complexity can be bounded by the deterministic average case query complexity replacing β by $\alpha \cdot \beta$. This yields

$$q \geq \frac{\alpha \cdot \beta - \sup_D \mathbf{Pr}_{P\sim\mu}\left[\mathrm{d_{TV}}(P, D) < \epsilon\right]}{\max_\phi \mathbf{Pr}_{P\sim\mu}\left[|P[\phi] - Q[\phi]| > \tau\right]}\,, \tag{118}$$

where, for the sake of transparency, we used Feldmans bound stated as Theorem 31 for the deterministic reference bound.

Note that Equation (118) is tight with respect to $\alpha \cdot \beta$ and the joint measure $\mu \times \mathcal{A}$. However, it has two disadvantages for our usecase. First, it only holds for classical probabilistic algorithms, but not for other random algorithms such as quantum algorithms. Second, we are interested in the average case hardness as in Definition 2. This means, we would like a statement that is tight with respect to β with respect to μ only. Thus, to obtain the corresponding tight bound for quantum algorithms, we add the following lemma.

Lemma 35: *Let $\mathcal{A}$ be a random algorithm for average case ϵ-learning $\mathcal{D}$ with respect to μ with parameters α and β from q many τ-accurate statistical queries. Then for any Q it holds*

$$q \geq 2 \cdot \frac{(\alpha - \frac{1}{2}) \cdot (\beta - \sup_D \mathbf{Pr}_{P\sim\mu}\left[\mathrm{d_{TV}}(P, D) < \epsilon\right])}{\max_\phi \mathbf{Pr}_{P\sim\mu}\left[|P[\phi] - Q[\phi]| > \tau\right]}\,, \tag{119}$$

where the sup *is over all distributions with the same domain X and the* max *is over all functions $\phi : X \to [-1, 1]$.*

Proof. Assume $\mathcal{A}$ is a random algorithm that ϵ-learns $\mathcal{D}$ with respect to μ, α and β

from q many τ-accurate statistical queries. We run $\mathcal{A}$ and answer each query ϕ by $Q[\phi]$. Denote by $\phi_1, \ldots, \phi_q$ the queries made and, without loss of generality, assume $\mathcal{A}$ returns the representation of some distribution D. Denote by $\mathcal{D}' \subseteq \mathcal{D}$ the set on which, with probability at least α, the algorithm is successful. Further, let $p(P)$ as in the proof of Theorem 33 and let

$$\mathcal{D}_D = \{P \in \mathcal{D}' : \mathrm{d_{TV}}(P, D) \geq \epsilon\} \ .$$

Since $p(P) \geq 2(\alpha - 1/2)$ for any $P \in \mathcal{D}_D \subseteq \mathcal{D}'$ we find

$$\begin{aligned} 2(\alpha - 1/2) &\leq \mathop{\mathbf{Pr}}_{P\sim\mu,\mathcal{A}}\left[\exists i \ , \ |P[\phi_i] - Q[\phi_i]| > \tau \mid P \in \mathcal{D}_D\right] \\ &= \frac{\mathbf{Pr}_{P\sim\mu,\mathcal{A}}\left[\exists i \ , \ |P[\phi_i] - Q[\phi_i]| > \tau\right]}{\mu(\mathcal{D}_D)} \leq \frac{\mathbf{Pr}_{P\sim\mu,\mathcal{A}}\left[\exists i \ , \ |P[\phi_i] - Q[\phi_i]| > \tau\right]}{\beta - \mathbf{Pr}_{P\sim\mu}\left[\mathrm{d_{TV}}(P, D) < \epsilon\right]} \qquad (120) \\ &\leq \sum_{i=1}^{q} \frac{\mathbf{Pr}_{P\sim\mu,\mathcal{A}}\left[|P[\phi_i] - Q[\phi_i]| > \tau\right]}{\beta - \mathbf{Pr}_{P\sim\mu}\left[\mathrm{d_{TV}}(P, D) < \epsilon\right]} \leq q \frac{\max_\phi \mathbf{Pr}_{P\sim\mu}\left[|P[\phi] - Q[\phi]| > \tau\right]}{\beta - \mathbf{Pr}_{P\sim\mu}\left[\mathrm{d_{TV}}(P, D) < \epsilon\right]} \ , \end{aligned}$$

where we used the definition of the conditional probability, the bound on $\mu(\mathcal{D}_D)$ similar to that on $\mu(\mathcal{D}_Q)$ from the proof of Theorem 30 and the union bound. The claim then follows from taking the maximum over all distributions in order to estimate the unknown D. □

It is easy to see that Theorem 35 is tight with respect to β: The trivial algorithm that always returns D, where D is the distribution with the ϵ-ball of highest weight, will succeed with probability $\mathbf{Pr}_{P\sim\mu}\left[\mathrm{d_{TV}}(P, D) < \epsilon\right]$. We conclude this appendix with a note similar to Note 3.

Note 5: In general Theorem 35 gives the optimal lower bounds with respect to β. However, in some cases directly bounding the weight of all ϵ-balls may not be convenient. The following appendix Section G gives a general recipe for obtaining such a bound just from the two ingredients used in Theorem 34: The maximally distinguishable fraction and the mass of the ball around the reference distribution. While this strategy is straight forward, the bounds obtained are slightly worse than those obtained by directly invoking Theorem 34, which is why we stick to the latter result in this work.

G Far from any fixed distribution

In the main text, we obtained multiple "far-from-uniform"-results for the output distributions of random circuits for different depth regimes. In this section, we show that random quantum circuits actually exhibit a more general property. Namely, their output distributions are with overwhelming probability far from any fixed distribution. This result was stated in the main text as Informal Theorem 2. Here, we restate it formally and then go on to prove it.

Theorem 36: (Formal version of Informal Theorem 2) *Let μ_C be the measure on $\mathrm{U}(2^n)$ induced by local random quantum circuits of depth d. Then, there is a $d' = O(n)$ such that at any depth $d \geq d'$, for any $\epsilon \leq 1/225$, and for any distribution D over $\{0,1\}^n$ it holds*

$$\Pr_{U\sim\mu_C}\left[\mathrm{d_{TV}}(P_U, D) \geq \epsilon\right] \geq 1 - c2^{-n}\,, \tag{121}$$

where c is a constant that can be bounded by $c < 7 \times 10^6 < 2^{20}$.

In the following, let D denote the arbitrary but fixed distribution as in the above theorem. We note that to prove Theorem 36, we can distinguish two cases: Either D is itself close to uniform, then a far-from-uniform bound implies a far-from-D bound. In the other case, D is at least some distance away from the uniform distribution. As made explicit by the following lemma, the "far-from-any-fixed-distribution"-result for such D is implied by a bound on the maximal distinguishable fraction with respect to the uniform distribution, $\mathfrak{f} = \mathrm{frac}(\mu, \mathcal{U}, \tau)$. In fact, the lemma holds not only for the uniform distribution but any choice of reference distribution Q.

Thus, a "far-from-any-fixed-distribution"-result follows from two ingredients: A "far-from-Q" result and a bound on the maximally distinguishable fraction against Q, for any reference distribution Q. We happen to have calculated these bounds for the particular choice of $Q = \mathcal{U}$ already in the main text.

Lemma 37: *Let $\epsilon, \tau > 0$, X be some domain and let $Q \in \mathcal{D}_X$ be the reference distribution. Moreover, let $D \in \mathcal{D}_X$ be such that*

$$\mathrm{d_{TV}}\left(Q, D\right) > \epsilon + \tau\,. \tag{122}$$

Then for any measure μ over $\mathcal{D}_X$ it holds

$$\Pr_{P\sim\mu}\left[\mathrm{d_{TV}}(P, D) < \epsilon\right] \leq \mathrm{frac}(\mu, Q, \tau)\,. \tag{123}$$

Proof. Recall that by the variational characterization of the total variation distance it holds that $\mathrm{d_{TV}}(Q, D) = \max_{T\subset X} |Q(T) - D(T)|$. Let $S \subseteq X$ be such a set maximizing the total variation distance and denote by $\phi = \mathbb{1}_S$ the characteristic function on S. This is, $\mathrm{d_{TV}}(D, Q) = |D[\phi] - Q[\phi]|$.

By the reverse triangle inequality for any $P' \in B_\epsilon(D)$ it then holds

$$\begin{aligned} |P'[\phi] - Q[\phi]| &\geq \big||P'[\phi] - D[\phi]| - |D[\phi] - Q[\phi]|\big| \\ &\geq |\mathrm{d_{TV}}(Q, D) - \mathrm{d_{TV}}(P', D)| \\ &> |\epsilon + \tau - \epsilon| = \tau\,, \end{aligned} \tag{124}$$

where we have used that $|P'[\phi] - D[\phi]| \leq \mathrm{d_{TV}}(P', D) < \epsilon$ together with $\mathrm{d_{TV}}(Q, D) > \epsilon + \tau > \epsilon > \mathrm{d_{TV}}(P', D)$.

Hence, by Equation (124) it holds

$$\Pr_{P\sim\mu}\left[\mathrm{d_{TV}}(P,D)<\epsilon\right]\leq\Pr_{P\sim\mu}\left[|P[\phi]-Q[\phi]|>\tau\right]\leq\mathrm{frac}(\mu,Q,\tau)\,, \tag{125}$$

where the last inequality is due to the maximum over all functions ϕ in the definition of $\mathrm{frac}(\mu,Q,\tau)$ (Definition 3). □

Applying Theorem 37 to the choice of $Q=\mathcal{U}$ and using our bound for the maximally distinguishable fraction $\mathfrak{f}$ against uniform from Theorem 7 in Section 4.1, we find the following:

Corollary 38: *Let $\epsilon,\tau>0$ and let D be any probability distribution over $\{0,1\}^n$ satisfying*

$$\mathrm{d_{TV}}\left(D,\mathcal{U}\right)>\epsilon+\tau \tag{126}$$

where $\mathcal{U}$ is the uniform distribution. Let μ_C denote the measure over brickwork random quantum circuits as in Definition 5. Then, there is a $d'=O(n)$ such that at any depth $d\geq d'$ it holds

$$\Pr_{U\sim\mu_C}\left[\mathrm{d_{TV}}(D,P_U)<\epsilon\right]\leq\frac{3}{2^n\tau^2}. \tag{127}$$

We now complete the proof of Theorem 36 following the two-cases argument laid out above.

Proof of Theorem 36. Let D be an arbitrary distribution, let d large enough and let $\epsilon=1/450$ such that $3\epsilon=\epsilon'=1/150$. We distinguish two cases:

1. $\mathrm{d_{TV}}(D,\mathcal{U})\leq 2\epsilon$,
2. $\mathrm{d_{TV}}(D,\mathcal{U})>2\epsilon$.

In case 1, we have that

$$\Pr_{U\sim\mu_C}\left[\mathrm{d_{TV}}(P_U,D)<\epsilon\right]\leq\Pr_{U\sim\mu_C}\left[\mathrm{d_{TV}}\left(P_U,\mathcal{U}\right)<\epsilon'\right]<3200\cdot 2^{-n}\leq O\left(2^{-n}\right) \tag{128}$$

by our far-from-uniform result from the main text, namely Theorem 11.

In case 2, we have that $\mathrm{d_{TV}}(D,\mathcal{U})>2\epsilon$. Setting $\tau=\epsilon$ we can apply Theorem 38 to obtain

$$\Pr_{U\sim\mu_C}\left[\mathrm{d_{TV}}(P_U,D)<\epsilon\right]<\frac{3}{2^n\epsilon^2}=607500\times 2^{-n}\,. \tag{129}$$

Hence, for any distribution D, any $\epsilon\leq 1/450$ we find

$$\Pr_{U\sim\mu_C}\left[\mathrm{d_{TV}}(P_U,D)<\epsilon\right]<607500\times 2^{-n}=O\left(2^{-n}\right)\,. \tag{130}$$

□

Finally, we summarize the connection between the ϵ-ball with the largest weight and the maximally distinguishable fraction as advertised at the end of Section E as follows.

Lemma 39: *Let $\epsilon, \tau > 0$, X be some domain, μ be a measure over $\mathcal{D}_X$ and $Q \in \mathcal{D}_X$ an arbitrary distribution. Then for any $D \in \mathcal{D}_X$ it holds*

$$\Pr_{P\sim\mu}\left[\mathrm{d}_{\mathrm{TV}}(P,D) < \epsilon\right] \leq \max\left\{\mathrm{frac}(\mu, Q, \tau)\,,\ \Pr_{P\sim\mu}\left[\mathrm{d}_{\mathrm{TV}}(P,Q) \leq 2\epsilon + \tau\right]\right\}\,. \tag{131}$$

Proof. We consider two cases. In case $\mathrm{d}_{\mathrm{TV}}(D,Q) > \epsilon + \tau$ we obtain the contribution from $\mathrm{frac}(\mu, Q, \tau)$ via Theorem 37. So consider the case $\mathrm{d}_{\mathrm{TV}}(D,Q) \leq \epsilon + \tau$. In this case we can bound

$$\Pr_{P\sim\mu}\left[\mathrm{d}_{\mathrm{TV}}(P,D) < \epsilon\right] \leq \Pr_{P\sim\mu}\left[\mathrm{d}_{\mathrm{TV}}(P,Q) \leq 2\epsilon + \tau\right]\,, \tag{132}$$

which yields the claim. □

References

[AA24] Anurag Anshu and Srinivasan Arunachalam. "A Survey on the Complexity of Learning Quantum States". In: *Nature Reviews Physics* 6.1 (Jan. 2024), pp. 59–69. DOI: 10.1038/s42254-023-00662-4 (page 3).

[AAK21] Naman Agarwal, Pranjal Awasthi, and Satyen Kale. "A Deep Conditioning Treatment of Neural Networks". In: *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*. Vol. 132. Proceedings of Machine Learning Research. PMLR, 2021, pp. 249–305. DOI: 10.48550/arXiv.2002.01523. URL: https://proceedings.mlr.press/v132/agarwal21b.html (pages 8, 9).

[Aar07] Scott Aaronson. "The learnability of quantum states". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 463.2088 (2007), pp. 3089–3114. DOI: 10.1098/rspa.2007.0113. arXiv: quant-ph/0608142 (page 9).

[ABG+14] Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. "Provable Bounds for Learning Some Deep Representations". In: *Proceedings of the 31st International Conference on Machine Learning*. Vol. 32. Proceedings of Machine Learning Research 1. Bejing, China: PMLR, 2014, pp. 584–592. DOI: 10.48550/arXiv.1310.6343. URL: https://proceedings.mlr.press/v32/arora14.html (page 9).

[AC17] Scott Aaronson and Lijie Chen. "Complexity-theoretic foundations of quantum supremacy experiments". In: *32nd Computational Complexity Conference (CCC 2017)*. Vol. 79. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017, 22:1–22:67. ISBN: 978-3-95977-040-8. DOI: 10.4230/LIPIcs.CCC.2017.22 (pages 4, 7, 9, 12, 24, 31).

[Ad17] Srinivasan Arunachalam and Ronald de Wolf. "Guest Column: A Survey of Quantum Learning Theory". In: *SIGACT News* 48.2 (June 2017), pp. 41–67. DOI: 10.1145/3106700.3106710 (page 3).

[AG04] Scott Aaronson and Daniel Gottesman. "Improved simulation of stabilizer circuits". In: *Phys. Rev. A* 70 (5 2004). DOI: 10.1103/PhysRevA.70.052328 (pages 8, 45).

[AGS21] Srinivasan Arunachalam, Alex Bredariol Grilo, and Aarthi Sundaram. "Quantum Hardness of Learning Shallow Classical Circuits". In: *SIAM Journal on Computing* 50.3 (2021). DOI: 10.1137/20M1344202, pp. 972–1013 (page 8).

[AGY20] Srinivasan Arunachalam, Alex B. Grilo, and Henry Yuen. *Quantum statistical query learning*. 2020. DOI: 10.48550/arXiv.2002.08240 (page 8).

[AHS23] Srinivasan Arunachalam, Vojtěch Havlíček, and Louis Schatzki. "On the role of entanglement and statistics in learning". In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NIPS '23. New Orleans, LA, USA: Curran Associates Inc., 2023. DOI: 10.48550/arXiv.2306.03161 (page 8).

[AK21] Eddie Aamari and Alexander Knop. "Statistical query complexity of manifold estimation". In: *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. Virtual Italy: ACM, 2021, pp. 116–122. ISBN: 978-1-4503-8053-9. DOI: 10.1145/3406325.3451135 (page 8).

[AK22] Eric R. Anschuetz and Bobak T. Kiani. "Quantum variational algorithms are swamped with traps". In: *Nature Communications* 13.1 (Dec. 2022). ISSN: 2041-1723. DOI: 10.1038/s41467-022-35364-5. URL: http://dx.doi.org/10.1038/s41467-022-35364-5 (page 8).

[AS23] Emmanuel Abbe and Colin Sandon. "Polynomial-time universality and limitations of deep learning". In: *Communications on Pure and Applied Mathematics* 76.11 (2023). DOI: 10.1002/cpa.22121, pp. 3493–3549. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.22121 (page 9).

[Bar17] Boaz Barak. "The Complexity of Public-Key Cryptography". In: *Tutorials on the Foundations of Cryptography: Dedicated to Oded Goldreich*. Ed. by Yehuda Lindell. Cham: Springer International Publishing, 2017, pp. 45–77. ISBN: 978-3-319-57048-8. DOI: 10.1007/978-3-319-57048-8_2 (page 10).

[BCG21] Boaz Barak, Chi-Ning Chou, and Xun Gao. "Spoofing linear cross-entropy benchmarking in shallow quantum circuits". In: *12th Innovations in Theoretical Computer Science Conference (ITCS 2021)*. Vol. 185. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2021, 30:1–30:20. ISBN: 978-3-95977-177-1. DOI: 10.4230/LIPIcs.ITCS.2021.30. URL: https://drops.dagstuhl.de/opus/volltexte/2021/13569 (pages 11, 30, 43, 44).

[BCH+21] Fernando G. S. L. Brandão, Wissam Chemissany, Nicholas Hunter-Jones, Richard Kueng, and John Preskill. "Models of quantum complexity growth". In: *PRX Quantum* 2 (3 2021). DOI: 10.1103/PRXQuan-

tum.2.030316. URL: https://link.aps.org/doi/10.1103/PRXQuantum.2.030316 (pages 3, 19).

[BDM+05] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. "Practical privacy: The SuLQ framework". In: *Proceedings of the Twenty-Fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems - PODS '05*. Baltimore, Maryland: ACM Press, 2005, p. 128. ISBN: 978-1-59593-062-0. DOI: 10.1145/1065167.1065184 (page 8).

[Ben21] Ben. *Solve the limit of the sucession* $\left(1+\frac{1}{2}\right)\left(1+\frac{1}{4}\right)\times\cdots\times\left(1+\frac{1}{2^n}\right)$. Mathematics Stack Exchange. 2021. URL: https://math.stackexchange.com/q/3386319 (page 45).

[Ber97] Bonnie Berger. "The Fourth Moment Method". In: *SIAM Journal on Computing* 26.4 (1997). DOI: 10.1137/S0097539792240005 (page 24).

[BFJ+94] Avrim Blum, Merrick Furst, Jeffrey Jackson, Michael Kearns, Yishay Mansour, and Steven Rudich. "Weakly learning DNF and characterizing statistical query learning using Fourier analysis". In: *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing*. STOC '94. Montreal, Quebec, Canada: Association for Computing Machinery, 1994. DOI: 10.1145/195058.195147 (page 8).

[BFK+94] Avrim Blum, Merrick Furst, Michael Kearns, and Richard J. Lipton. "Cryptographic Primitives Based on Hard Learning Problems". In: *Advances in Cryptology — CRYPTO' 93*. Ed. by Douglas R. Stinson. Berlin, Heidelberg: Springer, 1994, pp. 278–291. ISBN: 978-3-540-48329-8. DOI: 10.1007/3-540-48329-2_24 (page 10).

[BHH16] Fernando G. S. L. Brandao, Aram W. Harrow, and Michał Horodecki. "Local random quantum circuits are approximate polynomial-designs". In: *Communications in Mathematical Physics* 346.2 (2016), pp. 397–434. DOI: 10.1007/s00220-016-2706-8 (pages 12, 13, 41).

[BJS11] Michael J. Bremner, Richard Jozsa, and Dan J. Shepherd. "Classical simulation of commuting quantum computations implies collapse of the polynomial hierarchy". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 467.2126 (2011), pp. 459–472. DOI: 10.1098/rspa.2010.0301. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rspa.2010.0301. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rspa.2010.0301 (page 8).

[BMR+20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901. DOI: 10.48550/arXiv.2005.14165 (page 9).

[BMS17] Michael J. Bremner, Ashley Montanaro, and Dan J. Shepherd. "Achieving quantum supremacy with sparse and noisy commuting quantum computations". In: *Quantum* 1 (2017), p. 8. DOI: 10.22331/q-2017-04-25-8. arXiv: 1610.01808 [quant-ph] (page 8).

[CHM+15] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gerard Ben Arous, and Yann LeCun. "The Loss Surfaces of Multilayer Networks".

In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*. Vol. 38. Proceedings of Machine Learning Research. San Diego, California, USA: PMLR, 2015, pp. 192–204. DOI: 10.48550/arXiv.1412.0233. URL: https://proceedings.mlr.press/v38/choromanska15.html (page 9).

[CLL22] Sitan Chen, Jerry Li, and Yuanzhi Li. "Learning (very) simple generative models is hard". In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS '22. New Orleans, LA, USA: Curran Associates Inc., 2022. ISBN: 9781713871088. DOI: 10.48550/arXiv.2205.16003 (pages 8, 9).

[CMD+20] Brian Coyle, Daniel Mills, Vincent Danos, and Elham Kashefi. "The Born supremacy: Quantum advantage and training of an Ising born machine". In: *npj Quantum Information* 6.1 (2020), pp. 1–11. DOI: 10.1038/s41534-020-00288-9 (page 4).

[Dan05] Christoph Dankert. "Efficient simulation of random quantum states and operators". In: *arXiv* (2005). DOI: 10.48550/arXiv.quant-ph/0512217 (page 17).

[Dan17] Amit Daniely. "SGD Learns the conjugate kernel class of the network". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 2419–2427. ISBN: 978-1-5108-6096-4. DOI: 10.48550/arXiv.1702.08503 (page 9).

[DKS17] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. "Statistical query lower bounds for robust estimation of high-dimensional Gaussians and Gaussian mixtures". In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. 2017, pp. 73–84. DOI: 10.1109/FOCS.2017.16 (page 8).

[DMN+06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. "Calibrating noise to sensitivity in private data analysis". In: *Theory of Cryptography*. Ed. by Shai Halevi and Tal Rabin. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2006, pp. 265–284. ISBN: 978-3-540-32732-5. DOI: 10.1007/11681878_14 (page 8).

[DV20] Amit Daniely and Gal Vardi. "Hardness of learning neural networks with natural weights". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020, pp. 930–940. ISBN: 978-1-71382-954-6. DOI: 10.48550/arXiv.2006.03177 (page 9).

[Fel08] Vitaly Feldman. "Evolvability from learning algorithms". In: *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*. Victoria British Columbia Canada: ACM, 2008, pp. 619–628. ISBN: 978-1-60558-047-0. DOI: 10.1145/1374376.1374465 (page 8).

[Fel17] Vitaly Feldman. "A General Characterization of the Statistical Query Complexity". In: *Proceedings of the 2017 Conference on Learning Theory*. Vol. 65. Proceedings of Machine Learning Research. 2017. DOI: 10.48550/arXiv.1608.02198. URL: https://proceedings.mlr.press/v65/feldman17c.html (pages 7, 8, 16, 45, 48, 50, 51).

[FGR+17] Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh S Vempala, and Ying Xiao. “Statistical algorithms and a lower bound for detecting planted cliques”. In: *Journal of the ACM (JACM)* 64.2 (2017), pp. 1–37. DOI: 10.1145/3046674 (pages 4, 7, 8).

[FGV21] Vitaly Feldman, Cristóbal Guzmán, and Santosh Vempala. “Statistical Query Algorithms for Mean Vector Estimation and Stochastic Convex Optimization”. In: *Mathematics of Operations Research* 46.3 (2021), pp. 912–945. DOI: 10.1287/moor.2020.1111 (page 8).

[FPV18] Vitaly Feldman, Will Perkins, and Santosh Vempala. “On the Complexity of Random Satisfiability Problems with Planted Solutions”. In: *SIAM Journal on Computing* 47.4 (2018), pp. 1294–1338. DOI: 10.1137/16M1078471 (page 8).

[GAE07] David Gross, Koenraad Audenaert, and Jens Eisert. “Evenly distributed unitaries: On the structure of unitary designs”. In: *Journal of Mathematical Physics* 48.5 (2007), p. 052104. DOI: 10.1063/1.2716992 (page 17).

[Got98] Daniel Gottesman. *The Heisenberg representation of quantum computers*. 1998. DOI: 10.48550/arXiv.quant-ph/9807006 (page 8).

[GPM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. “Generative adversarial nets”. In: *Advances in neural information processing systems*. 2014, pp. 2672–2680. DOI: 10.48550/arXiv.1406.2661 (page 9).

[Haf22] Jonas Haferkamp. “Random quantum circuits are approximate unitary t-designs in depth $O\left(nt^{5+o(1)}\right)$”. In: *Quantum* 6 (Sept. 2022), p. 795. ISSN: 2521-327X. DOI: 10.22331/q-2022-09-08-795 (pages 12, 13, 41).

[Har13] Aram W. Harrow. “The church of the symmetric subspace”. In: (2013). DOI: 10.48550/arXiv.1308.6595 (page 41).

[HBM+22] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. “Training compute-optimal large language models”. In: *arXiv* (2022). DOI: 10.48550/arXiv.2203.15556 (page 9).

[HE22] Dominik Hangleiter and Jens Eisert. *Computational advantage of quantum random sampling*. 2022. DOI: 10.48550/arXiv.2206.04079. URL: https://arxiv.org/abs/2206.04079 (page 3).

[HH21] Jonas Haferkamp and Nicholas Hunter-Jones. “Improved spectral gaps for random quantum circuits: large local dimensions and all-to-all interactions”. In: *Physical Review A* 104.2 (2021), p. 022417. DOI: 10.1103/PhysRevA.104.022417 (pages 13, 23, 25, 41).

[HIN+23] M. Hinsche, M. Ioannou, A. Nietner, J. Haferkamp, Y. Quek, D. Hangleiter, J.-P. Seifert, J. Eisert, and R. Sweke. “One T Gate Makes Distribution Learning Hard”. In: *Phys. Rev. Lett.* 130 (24 2023), p. 240602. DOI: 10.1103/PhysRevLett.130.240602. URL: https://link.aps.org/doi/10.1103/PhysRevLett.130.240602 (pages 4, 15).

[HL09] Aram W. Harrow and Richard A. Low. “Random quantum circuits are approximate 2-designs”. In: *Communications in Mathematical Physics* 291.1 (2009), pp. 257–302. DOI: 10.1007/s00220-009-0873-6 (page 41).

[HLB+24] Hsin-Yuan Huang, Yunchao Liu, Michael Broughton, Isaac Kim, Anurag Anshu, Zeph Landau, and Jarrod R. McClean. “Learning Shallow Quantum Circuits”. In: *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*. STOC 2024. Association for Computing Machinery, 2024. DOI: 10.1145/3618260.3649722. URL: https://doi.org/10.1145/3618260.3649722 (page 9).

[Hun19] Nicholas Hunter-Jones. *Unitary designs from statistical mechanics in random quantum circuits*. 2019. DOI: 10.48550/ARXIV.1905.12053. URL: https://arxiv.org/abs/1905.12053 (pages 11, 13, 30).

[IL90] Russell Impagliazzo and Leonid A. Levin. “No better ways to generate hard NP instances than picking uniformly at random”. In: *Proceedings [1990] 31st Annual Symposium on Foundations of Computer Science*. 1990, 812–821 vol.2. DOI: 10.1109/FSCS.1990.89604 (page 10).

[JEP+21] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. “Highly accurate protein structure prediction with AlphaFold”. In: *Nature* 596.7873 (2021), pp. 583–589. URL: https://doi.org/10.1038/s41586-021-03819-2 (page 9).

[JSA16] Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. *Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods*. 2016. DOI: 10.48550/arXiv.1506.08473 (page 9).

[Kea93] Michael Kearns. “Efficient Noise-Tolerant Learning from Statistical Queries”. In: *Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing*. STOC ’93. San Diego, California, USA: Association for Computing Machinery, 1993, pp. 392–401. ISBN: 0897915917. DOI: 10.1145/167088.167200. URL: https://doi.org/10.1145/167088.167200 (pages 3, 7).

[KG15] Richard Kueng and David Gross. “Qubit stabilizer states are complex projective 3-designs”. In: (2015). DOI: 10.48550/arXiv.1510.02767 (page 45).

[Kha93] Michael Kharitonov. “Cryptographic hardness of distribution-specific learning”. In: *Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing*. STOC ’93. San Diego, California, USA: Association for Computing Machinery, 1993, pp. 372–381. ISBN: 0897915917. DOI: 10.1145/167088.167197. URL: https://doi.org/10.1145/167088.167197 (page 8).

[KLN+11] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. “What can we learn privately?” In: *SIAM Journal on Computing* 40.3 (2011). DOI: 10.1137/090756090 (page 8).

[KMR+94] Michael J. Kearns, Yishay Mansour, D. Ron, Ronitt Rubinfeld, Robert E. Schapire, and Linda Sellie. “On the learnability of discrete distributions”. In: *Proceedings of the Twenty-sixth Annual ACM Symposium on*

Theory of Computing. STOC '94. Montreal, Quebec, Canada: ACM, 1994, pp. 273–282. ISBN: 0-89791-663-8. DOI: 10.1145/195058.195155. URL: http://doi.acm.org/10.1145/195058.195155 (pages 7, 8).

[Kna88] Stefan Knabe. "Energy gaps and elementary excitations for certain VBS-quantum antiferromagnets". In: *Journal of Statistical Physics* 52.3 (1988), pp. 627–638. DOI: 10.1007/BF01019721 (page 41).

[KV94] Michael J. Kearns and Umesh Vazirani. *An Introduction to Computational Learning Theory*. The MIT Press, Aug. 1994. ISBN: 9780262276863. DOI: 10.7551/mitpress/3897.001.0001. URL: https://doi.org/10.7551/mitpress/3897.001.0001 (page 9).

[LL25] Zeph Landau and Yunchao Liu. "Learning Quantum States Prepared by Shallow Circuits in Polynomial Time". In: *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*. STOC '25. Prague, Czechia: Association for Computing Machinery, 2025, pp. 1828–1838. DOI: 10.1145/3717823.3718311. URL: https://doi.org/10.1145/3717823.3718311 (page 9).

[LMN93] Nathan Linial, Yishay Mansour, and Noam Nisan. "Constant depth circuits, Fourier transform, and learnability". In: *J. ACM* 40.3 (1993), pp. 607–620. ISSN: 0004-5411. DOI: 10.1145/174130.174138. URL: https://doi.org/10.1145/174130.174138 (page 8).

[LSS14] Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. "On the Computational Efficiency of Training Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger. Vol. 27. Curran Associates, Inc., 2014. DOI: 10.48550/arXiv.1410.1141. URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/9abb5a595720451e64d761e3a7827814-Paper.pdf (page 9).

[MNK+18] Kosuke Mitarai, Makoto Negoro, Masahiro Kitagawa, and Keisuke Fujii. "Quantum circuit Learning". en. In: *Physical Review A* 98.3 (2018). arXiv:1803.00745, p. 032309. DOI: 10.1103/PhysRevA.98.032309. URL: http://arxiv.org/abs/1803.00745 (page 8).

[Mon17] Ashley Montanaro. "Learning stabilizer states by Bell sampling". In: (2017). DOI: 10.48550/arXiv.1707.04012 (page 9).

[Nac96] Bruno Nachtergaele. "The spectral gap for some spin chains with discrete symmetry breaking". In: *Communications in Mathematical Physics* 175.3 (1996), pp. 565–606. DOI: 10.1007/BF02099509 (page 41).

[Nan21] Mikito Nanashima. "A theory of heuristic learnability". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Conference on Learning Theory. optissn: 2640-3498. PMLR, 2021, pp. 3483–3525. URL: https://proceedings.mlr.press/v134/nanashima21a.html (page 10).

[Nie23] Alexander Nietner. *Unifying (Quantum) Statistical and Parametrized (Quantum) Algorithms*. 2023. DOI: 10.48550/arXiv.2310.17716 (page 8).

[SBG+19] Maria Schuld, Ville Bergholm, Christian Gogolin, Josh Izaac, and Nathan Killoran. "Evaluating analytic gradients on quantum hardware". en. In: *Physical Review A* 99.3 (2019). arXiv:1811.11184, p. 032331. DOI:

10.1103/PhysRevA.99.032331. URL: http://arxiv.org/abs/1811.11184 (page 8).

[Sha18] Ohad Shamir. "Distribution-Specific Hardness of Learning Neural Networks". In: *Journal of Machine Learning Research* 19.32 (2018), pp. 1–29. DOI: 10.48550/arXiv.1609.01037. URL: http://jmlr.org/papers/v19/17-537.html (page 9).

[Spa98] James C. Spall. "An overview of the simultaneous perturbation method for efficient optimization". en. In: *Johns Hopkins APL Technical Digest* 19.4 (1998), p. 11 (page 8).

[SVW16] Jacob Steinhardt, Gregory Valiant, and Stefan Wager. "Memory, Communication, and Statistical Queries". In: *29th Annual Conference on Learning Theory*. Ed. by Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir. Vol. 49. Proceedings of Machine Learning Research. Columbia University, New York, New York, USA: PMLR, 2016, pp. 1490–1516. URL: https://proceedings.mlr.press/v49/steinhardt16.html (page 8).

[TD02] Barbara M. Terhal and David P. DiVincenzo. "Classical simulation of noninteracting-fermion quantum circuits". In: *Physical Review A* 65.3 (2002), p. 032325. DOI: 10.1103/PhysRevA.65.032325. arXiv: quant-ph/0108010 (page 8).

[Val09] Leslie G. Valiant. "Evolvability". In: *Journal of the ACM* 56.1 (2009), pp. 1–21. DOI: 10.1145/1462153.1462156 (page 8).

[Val12] Leslie G. Valiant. "Quantum circuits that can be simulated classically in polynomial time". In: *SIAM Journal on Computing* (2012). DOI: 10.1137/S0097539700377025 (page 8).

[Val84] Leslie G. Valiant. "A theory of the learnable". In: *Proceedings of the sixteenth annual ACM symposium on Theory of computing*. STOC '84. New York, NY, USA: Association for Computing Machinery, 1984, pp. 436–445. ISBN: 978-0-89791-133-7. DOI: 10.1145/800057.808710. URL: https://doi.org/10.1145/800057.808710 (page 7).

[WD25] Chirag Wadhwa and Mina Doosti. "Learning Quantum Processes with Quantum Statistical Queries". In: *Quantum* 9 (May 2025), p. 1739. ISSN: 2521-327X. DOI: 10.22331/q-2025-05-12-1739. URL: https://doi.org/10.22331/q-2025-05-12-1739 (page 8).

[ZLK+24] Haimeng Zhao, Laura Lewis, Ishaan Kannan, Yihui Quek, Hsin-Yuan Huang, and Matthias C. Caro. "Learning Quantum States and Unitaries of Bounded Gate Complexity". In: *PRX Quantum* 5 (4 2024), p. 040306. DOI: 10.1103/PRXQuantum.5.040306. URL: https://link.aps.org/doi/10.1103/PRXQuantum.5.040306 (pages 7, 9).