# Predicting quantum channels over general product distributions

**Sitan Chen** SITAN@SEAS.HARVARD.EDU
*Harvard University*

**Jaume de Dios Pont** JAUME.DEDIOSPONT@MATH.ETHZ.CH
*ETH Zurich*

**Jun-Ting Hsieh** JUNTINGH@CS.CMU.EDU
*Carnegie Mellon University*

**Hsin-Yuan Huang** HSINYUAN@CALTECH.EDU
*Google Quantum AI, Caltech*

**Jane Lange** JLANGE@MIT.EDU
*MIT*

**Jerry Li** JERRYZLI@U.WASHINGTON.EDU
*University of Washington*

**Editors:** Nika Haghtalab and Ankur Moitra

## Abstract

We investigate the problem of predicting the output behavior of unknown quantum channels. Given query access to an $n$-qubit channel $\mathcal{E}$ and an observable $\mathcal{O}$, we aim to learn the mapping

$$\rho \mapsto \mathrm{Tr}(\mathcal{O}\mathcal{E}[\rho])$$

to within a small error for most $\rho$ sampled from a distribution $\mathcal{D}$. Previously, Huang et al. (2023) proved a surprising result that even if $\mathcal{E}$ is arbitrary, this task can be solved in time roughly $n^{O(\log(1/\varepsilon))}$, where $\varepsilon$ is the target prediction error. However, their guarantee applied only to input distributions $\mathcal{D}$ invariant under all single-qubit Clifford gates, and their algorithm fails for important cases such as general product distributions over product states $\rho$.

In this work, we propose a new approach that achieves accurate prediction over essentially any product distribution $\mathcal{D}$, provided it is not "classical" in which case there is a trivial exponential lower bound. Our method employs a "biased Pauli analysis," analogous to classical biased Fourier analysis. Implementing this approach requires overcoming several challenges unique to the quantum setting, including the lack of a basis with appropriate orthogonality properties. The techniques we develop to address these issues may have broader applications in quantum information.

**Keywords:** Quantum learning, supervised learning, low-degree approximation, Fourier analysis

## 1. Introduction

When is it possible to learn to predict the outputs of a quantum channel $\mathcal{E}$? Such questions arise naturally in a variety of settings, such as the experimental study of complex quantum dynamics (Huang et al., 2022, 2021), and in fast-forwarding simulations of Hamiltonian evolutions (Cirstoiu et al., 2020; Gibbs et al., 2024). However, in the worst case this problem is intractable, as it generalizes the classical problem of learning an arbitrary Boolean function over the uniform distribution from black-box access. To circumvent this, our goal is to understand families of natural restrictions on the problem under which efficient estimation is possible.

One way to avoid this exponential scaling would be to posit further structure on the channel, e.g. by assuming it is given by a shallow quantum circuit (Nadimpalli et al., 2023; Huang et al., 2024) or a structured Pauli channel (Harper et al., 2020; Flammia and O'Donnell, 2021; Harper et al., 2021; Van Den Berg et al., 2023; Arunachalam et al., 2024). However, there are settings where the evolutions may be quite complicated — e.g. the channel might correspond to the time evolution of an evaporating black hole (Hayden and Preskill, 2007; Hayden and Penington, 2019; Penington et al., 2022; Yang and Engelhardt, 2023) — and where it is advantageous to avoid such strong structural assumptions on the underlying channel.

Recently, Huang et al. (2023) considered an alternative workaround in which one only attempts to learn a complicated $n$-qubit channel in an *average-case* sense. Given query access to $\mathcal{E}$, and given an observable $\mathcal{O}$, the goal is to learn the mapping

$$\rho \mapsto \mathrm{Tr}(\mathcal{O}\,\mathcal{E}[\rho])$$

accurately on average over input states $\rho$ drawn from some $n$-qubit distribution $\mathcal{D}$, rather than over worst-case input states. The authors of Huang et al. (2023) came to the surprising conclusion that this average-case task is tractable even for *arbitrary* channels $\mathcal{E}$, provided $\mathcal{D}$ comes from a certain class of "locally flat" distributions. Their key observation was that the Heisenberg-evolved observable $\mathcal{E}^\dagger[\mathcal{O}]$ admits a *low-degree approximation* in the Pauli basis, where the quality of approximation is defined in an average-case sense over input state $\rho$.

Another interesting feature of this result is that their learning algorithm only needs to query $\mathcal{E}$ on random product states, regardless of the choice of locally flat distribution $\mathcal{D}$. This is both an advantage and a shortcoming. On one hand, if one is certain that the states $\rho$ one wants to predict on are samples from a locally flat distribution, no further information about $\mathcal{D}$ is needed to implement the learning protocol in Huang et al. (2023). On the other hand, locally flat distributions are quite specialized: they are constrained to be invariant under any single-qubit Clifford gate. In particular, almost all product distributions over product states fall outside this class. Worse yet, the general approach of low-degree approximation in the Pauli basis can be shown to fail when local flatness does not hold (see Section C.2). We therefore ask:

*Are there more general families of distributions $\mathcal{D}$ under which one can learn to predict arbitrary quantum dynamics?*

Identifying rich settings where it is possible to characterize the average-case behavior of such dynamics, while making minimal assumptions on the dynamics, is of intense practical interest. Unfortunately, our understanding of this remains limited: even for general product distributions, known techniques break down. In this work we take an important first step towards this goal by completely characterizing the complexity of learning to predict arbitrary quantum dynamics in the product setting. Informally stated, our main result is that learning is possible *so long as the distribution is not classical*. That is, for this problem there is a "blessing of quantum-ness": as long as the distribution displays any quantitative level of quantum behavior, there is an efficient algorithm for predicting arbitrary quantum dynamics under this distribution.

More formally, note that if $D$ is the uniform distribution over the computational basis states $|0\rangle$ and $|1\rangle$, then the task of predicting $\mathrm{Tr}(\mathcal{O}\,\mathcal{E}[\rho])$ on average over $\rho \sim \mathcal{D} \triangleq D^{\otimes n}$ for an arbitrary channel $\mathcal{E}$ is equivalent to the task of learning an arbitrary Boolean function from random labeled examples, which trivially requires exponentially many samples. This logic naturally extends to any "two-point" distribution in which $D$ is supported on two diametrically opposite points on the Bloch sphere. Note that any such distribution, up to a rotation, is an embedding of a classical distribution onto the Bloch sphere.

A natural way of quantifying closeness to such distributions is in terms of the second moment matrix $\mathcal{S} \in \mathbb{R}^{3\times 3}$ of the distribution $D$, when $D$ is viewed as a distribution over the Bloch sphere (see Section 2.1 for formal definitions). We refer to this matrix as the *Pauli second moment matrix* of $D$. For the purposes of this discussion, the key property of this matrix is that $\|\mathcal{S}\|_{\mathsf{op}} \leq 1$ for all $D$, and moreover, $\|\mathcal{S}\|_{\mathsf{op}} = 1$ if and only if $D$ is one of the aforementioned two-point distributions. With this, we can now state our main result:

**Theorem 1 (Learning an unknown quantum channel)** *Let $\varepsilon, \delta, \eta \in (0,1)$. Let $D$ be an unknown distribution over the Bloch sphere with Pauli second moment matrix $\mathcal{S}$ such that $\|\mathcal{S}\|_{\mathsf{op}} \leq 1-\eta$. Let $\mathcal{E}$ be an unknown $n$-qubit quantum channel, and let $\mathcal{O}$ be a known $n$-qubit observable. There exists an algorithm with time and sample complexity $\min(2^{O(n)}/\varepsilon^2, n^{O(\log(1/\varepsilon)/\log(1/(1-\eta)))}) \cdot \log(1/\delta)$ that outputs an efficiently computable map $f'$ such that*

$$\mathbb{E}_{\rho\sim D^{\otimes n}}[(\mathrm{Tr}(\mathcal{O}\mathcal{E}[\rho]) - \mathrm{Tr}(f'(\rho)))^2] \leq \varepsilon$$

*with probability at least $1-\delta$.*

Note that the only condition on $D$ we require is a quantitative bound on the spectral norm of its Pauli second moment matrix. In other words, so long as the distribution $D$ is far from any two-point distribution, i.e., it is far from any classical distribution, we demonstrate that there is an efficient algorithm for learning to predict general quantum dynamics under this distribution. Previously it was only known how to achieve the above guarantee in the special case where $D$ has mean zero. Indeed, as soon as one deviates from the mean zero case, the Pauli decomposition approach of Huang et al. (2023) breaks down. In contrast, our guarantee works for any product distribution whose marginal second moment matrices have operator norm bounded away from 1.

**Remark 2** *We note that our techniques generalize to the case where the distribution is the product of different distributions over qubits, so long as each distribution has second moment with operator norm bounded by $1-\eta$. However, for readability we will primarily focus on the case where all of the distributions are the same. See Section D.2 for a discussion of how to easily generalize our techniques to this setting.*

**Beyond low-degree concentration in an orthonormal basis.** Here we briefly highlight the key conceptual novelties of our analysis, which may be of independent interest. We begin by recalling the analysis in Huang et al. (2023) in greater detail. They considered the decomposition of $\mathcal{O}' \triangleq \mathcal{E}^\dagger[\mathcal{O}]$ into the basis of $n$-qubit Pauli operators, i.e. $\mathcal{O}' = \sum_{P\in\{I,X,Y,Z\}^n} \alpha_P \cdot P$ and argued that this is well-approximated by the *low-degree truncation* $\mathcal{O}'_{\mathrm{low}} = \sum_{|P|<t} \alpha_P \cdot P$. This can be readily seen from the following calculation. By rotating the distribution $D$, we may assume the covariance is diagonal, with entries bounded by $1-\eta$. Then the error achieved by the low-degree truncation is given by

$$\mathbb{E}[\mathrm{Tr}((\mathcal{O}' - \mathcal{O}'_{\mathrm{low}})\rho)^2] = \mathbb{E}\Big[\Big(\sum_{|P|\geq t} \alpha_P \cdot \mathrm{Tr}(P\rho)\Big)^2\Big]$$
$$\leq \sum_{|P|\geq t} (1-\eta)^{|P|} \cdot \alpha_P^2 \leq (1-\eta)^t \cdot \frac{1}{2^n}\|\mathcal{O}'\|_F^2 \leq (1-\eta)^t\,,$$

where the second step follows from the fact that $\mathbb{E}[\mathrm{Tr}(P\rho)\,\mathrm{Tr}(Q\rho)] = 0$ if $P \neq Q$ and is at most $(1-\eta)^{|P|}$ if $P = Q$ (since $D$ is mean zero and its covariance is diagonal with entries bounded by $1-\eta$), and the last step follows by the assumption that $\|\mathcal{O}'\|_{\mathsf{op}} \le 1$.

Note that when $D$ is not mean zero, this step breaks, and we do not have this nice exponential decay in $t$. In fact, in Section C.2 we construct examples of operators which are not well-approximated by their low-degree truncations in the Pauli basis when $D$ has mean bounded away from zero.

A natural attempt at a workaround would be to change the basis under which we truncate. At least classically, biased product distributions over the Boolean hypercube still admit suitable orthonormal bases of functions, namely the *biased Fourier characters*. As we show in Appendix B, this idea can be used to give the following learning guarantee in the classical case where $D$ is only supported along the $Z$ direction in the Bloch sphere:

**Theorem 3 (PAC learning over a concentrated product distribution)** *Let* $\varepsilon, \delta, \eta \in (0,1)$. *Let* $D$ *be an unknown distribution over the interval* $[-(1-\eta), 1-\eta]$. *Let* $f : [-1,1]^n \to [-1,1]$ *be an unknown bounded, multilinear function. There exists an algorithm with time and sample complexity* $n^{O(\log(1/\varepsilon)/\log(1/(1-\eta)))} \cdot \log(1/\delta)$ *that outputs a hypothesis* $f'$ *such that*

$$\mathbb{E}_{x\sim D^{\otimes n}}[(f(x) - f'(x))^2] \le \varepsilon$$

*with probability at least* $1-\delta$.

Unfortunately, when we move beyond the classical setting, the picture becomes trickier. In particular, it is not immediately clear what the suitable analogue of the biased Fourier basis should be in the quantum setting. We could certainly try to consider single-qubit operators of the form $\widetilde{P} = P - \mu_P \cdot I$ for $P \in \{X, Y, Z\}$, where $\mu_P$ denotes the $P$-th coordinate of the mean of $D$ regarded as a distribution over the Bloch sphere. We could then extend naturally to give a basis over $n$ qubits, and the functions $\rho \mapsto \mathrm{Tr}(\widetilde{P}\rho)$ would by design be orthogonal to each other with respect to the distribution $D^{\otimes n}$. Writing $\mathcal{O}' = \sum_P \widetilde{\alpha}_P \cdot \widetilde{P}$ and defining $\widetilde{\mathcal{O}}_{\mathrm{low}} \triangleq \sum_{|P|<t} \widetilde{\alpha}_P \cdot \widetilde{P}$, we can mimic the calculation above and obtain

$$\mathbb{E}[\mathrm{Tr}((\mathcal{O}' - \widetilde{\mathcal{O}}_{\mathrm{low}})\rho)^2] = \mathbb{E}\Big[\Big(\sum_{|P|\ge t} \widetilde{\alpha}_P \cdot \mathrm{Tr}(\widetilde{P}\rho)\Big)^2\Big] \le (1-\eta)^{|P|} \sum_{|P|\ge t} \widetilde{\alpha}_P^2 \,.$$

Unfortunately, at this juncture the above naive approach hits a snag. In the mean zero case, we could easily relate $\sum_P \alpha_P^2$ to $\frac{1}{2^n}\|\mathcal{O}'\|_F^2$ because of a fortuitous peculiarity of the mean-zero setting. In that setting, we implicitly exploited both that the Pauli operators $P$ are orthogonal to each other with respect to the trace inner product, and also that that the functions $\rho \mapsto \mathrm{Tr}(P\rho)$ are orthogonal to each other with respect to $D^{\otimes n}$. In the above approach for nonzero mean, we achieved the latter condition by shifting the Pauli operators to define $\widetilde{P}$, but these shifted operators are no longer orthogonal to each other with respect to the trace inner product.

Crucially, this means that *standard Fourier-analytic tools cannot control the truncation error* in this general setting. Circumventing this issue is the technical heart of our proof. As we will see in Section 3, several technical innovations are needed.

First, instead of assuming that the covariance of $D$ is diagonalized, we will fix a rotation that simplifies the *mean*, $\mathbb{E}[\rho]$. Second, instead of shifting the basis operators $\{X, Y, Z\}$ so that the resulting functions $\rho \mapsto \mathrm{Tr}(\widetilde{P}\rho)$ are orthogonal with respect to $D^{\otimes n}$, we shift them so that $\mathbb{E}[\rho]$

is *orthogonal* to them, and define $\mathcal{O}'_{\text{low}}$ by truncating in this new basis instead. Finally, instead of directly bounding the truncation error $\mathbb{E}[\operatorname{Tr}((\mathcal{O}' - \mathcal{O}'_{\text{low}})\rho)^2]$ using the above sequence of steps, we crucially relate it to the quantity

$$\operatorname{Tr}((\mathcal{O}')^2\,\mathbb{E}[\rho])$$

in order to establish exponential decay. Note that $\operatorname{Tr}((\mathcal{O}')^2\,\mathbb{E}[\rho]) \le \|\mathcal{O}'\|_{\mathsf{op}}^2 \cdot \operatorname{Tr}(\mathbb{E}[\rho]) \le 1$. To our knowledge, all three of these components are new to our analysis. We leave it as an intriguing open question to find other applications of these ingredients to domains where "biased Pauli analysis" arises.

At a conceptual level, we believe our techniques are potentially of independent interest. The breakdown of "standard" Fourier analytic tools suggests that this general setting may be substantially more complex than the mean-zero setting. For instance, it is unclear if Bohnenblust-Hille style inequalities like those proven in Huang et al. (2023) hold for general product distributions; certainly the well-known proof strategies do not carry over to this setting.

**Remark 4 (Predicting multiple observables)** Just as in Huang et al. (2023), we can also easily extend our guarantee to the setting where we wish to learn the joint mapping

$$(\rho, \mathcal{O}) \mapsto \operatorname{Tr}(\mathcal{O}\,\mathcal{E}[\rho])\,. \tag{1}$$

This is the natural channel learning analogue of the question of classical shadows for state learning (Huang et al., 2020) – recall that in the latter setting, one would like to perform measurements on copies of $\rho$ and obliviously produce a classical description of the state that can then be used to compute some collection of observable values. We sketch the argument for extending to learning the joint mapping in Eq. (1) in Section D.1.

**Impossibility for general concentrated distributions.** It is natural to wonder to what extent our results can be generalized, especially to states that are entangled. Could it be that all one needs is some kind of global covariance bound? Unfortunately, we show in Appendix C that even in the classical setting, this is not the case. Since classical distributions can be encoded by distributions over qubits, this implies hardness for learning in the quantum setting as well.

**Theorem 5 (Hardness of learning over general concentrated distributions)** *There exists a distribution $\mathcal{D}$ over $[-(1-\eta), 1-\eta]^n$ and a concept class $\mathcal{C}$ such that no algorithm PAC-learns the class $\mathcal{C}$ over $\mathcal{D}$ in subexponential time.*

There is a wide spectrum of distributional assumptions that interpolates between fully product distributions and general concentrated distributions — for instance, products of $k$-dimensional qudit distributions, output states of small quantum circuits, or distributions over negatively associated variables. As discussed earlier, our understanding of when it is possible to predict the average-case behavior of arbitrary quantum dynamics is still nascent, and understanding learnability with respect to these more expressive distributional assumptions remains an important open question.

**Organization.** In Section 2, we state some preliminaries. In Section 3, we prove Theorem 1, our main result. Finally, in Appendix C, we show impossibility results including Theorem 5 and the failure of low-degree truncation in the standard Pauli basis.

In Appendix B, we prove Theorem 3, which is the classical setting and can be viewed as a warm-up to the quantum setting. In Appendix D, we remark on simple extensions of our results.

### 1.1. Related work

Our work is part of a growing literature bridging classical computational learning theory and its quantum counterpart. Its motivation can be thought of as coming from the general area of quantum process tomography (Mohseni et al., 2008), but as this is an incredibly extensive research direction, here we only focus our attention on surveying directly relevant works.

**Learning quantum channels.** In the appendix, we describe representative works in the literature on learning *full descriptions* of channels using Pauli analysis, the quantum analogue of Fourier analysis. In contrast, our focus is on learning certain *properties* of the channel, and only in an average-case sense over input states. As mentioned previously, this specific question was first studied in Huang et al. (2023). There have been two direct follow-up works to this paper which are somewhat orthogonal to the thrust of our contributions. The first (Volberg and Zhang, 2023) establishes refined versions of the so-called *non-commutative Bohnenblust-Hille inequality* which was developed and leveraged by Huang et al. (2023) to obtain *logarithmic* sample complexity bounds. As mentioned earlier, it is unclear whether Bohnenblust-Hille-style inequalities hold for this general setting, since standard proof techniques fail. We leave it as a very interesting open question as to whether or not one can still achieve logarithmic sample complexities in this setting, with or without Bohnenblust-Hille. The second follow-up Klein et al. (2023) to Huang et al. (2023) studies the natural qudit generalization of the original question where the distribution over qudits is similarly closed under a certain family of single-site transformations.

Finally, we note the recent work of Arunachalam et al. (2024) which studied the learnability of quantum channels with only low-degree Pauli coefficients. Their focus is incomparable to ours as they target a stronger metric, namely $\ell_2$-distance for channels, but need to make a strong assumption on the complexity of the channel being learned. In contrast, we target a weaker metric, namely average-case error for predicting observables, but our guarantee applies for arbitrary channels.

## 2. Preliminaries

### 2.1. Bloch sphere and Pauli covariance matrices

The Pauli matrices $I, X, Y, Z$ provide the basis for $2 \times 2$ Hermitian matrices. This is captured by the following standard fact on expanding a single-qubit state using Pauli matrices.

**Fact 6 (Pauli expansion of states)** *Any single-qubit mixed state $\rho$ can be written as $\rho = \frac{1}{2}(I + \alpha_x X + \alpha_y Y + \alpha_z Z)$ where $\boldsymbol{\alpha} \in \mathbb{R}^3$, $\|\boldsymbol{\alpha}\|_2 \leq 1$, and $X, Y, Z$ are the standard Pauli matrices. The set of all such $\boldsymbol{\alpha}$ of unit norm is the* Bloch sphere.

*Any single-qubit distribution $D$ can be viewed as a distribution over the Bloch sphere. We use $\mathbb{E}_D[\rho] \in \mathbb{C}^{2\times 2}$ and $\boldsymbol{\mu} \in \mathbb{R}^3$ to refer to the expected state and the expected Bloch vector respectively.*

By taking tensor products of Pauli matrices, we obtain the collection of $4^n$ Pauli observables $\{I, X, Y, Z\}^{\otimes n}$, which form a basis for the space of $2^n \times 2^n$ Hermitian matrices:

**Fact 7 (Pauli expansion of observables)** *Let $\mathcal{O}$ be an $n$-qubit observable. Then $\mathcal{O}$ can be written in the form $\mathcal{O} = \sum_{P \in \{I,X,Y,Z\}^{\otimes n}} \widehat{\mathcal{O}}(P) \cdot P$ where $\widehat{\mathcal{O}}(P) \triangleq \mathrm{Tr}(\mathcal{O}P)/2^n$.*

Next, we define the Pauli covariance and second moment matrices associated to any distribution $D$ over the single-qubit Bloch sphere.

**Definition 8 (Pauli covariance and second moment matrices)** *Let $D$ be a distribution over the Bloch sphere. We will associate with $D$ the second moment matrix $\mathcal{S} \in \mathbb{R}^{3\times 3}$ and covariance matrix $\Sigma \in \mathbb{R}^{3\times 3}$, indexed by the non-identity Pauli components $X, Y$, and $Z$. We define $\mathcal{S}$ such that $\mathcal{S}_{P,Q} = \mathbb{E}_{\rho\sim D}[\mathrm{Tr}(P\rho)\mathrm{Tr}(Q\rho)]$, and $\Sigma$ such that $\Sigma_{P,Q} = \mathbb{E}_{\rho\sim D}[\mathrm{Tr}(P\rho)\mathrm{Tr}(Q\rho) - \mathrm{Tr}(P\,\mathbb{E}[\rho])\mathrm{Tr}(Q\,\mathbb{E}[\rho])]$.*

**Fact 9** *For any distribution over the Bloch sphere, $\mathrm{Tr}(\mathcal{S}) = 1$.*

### 2.2. Access model

In this paper we consider the following, standard access model, see e.g. Huang et al. (2023). We assume that we can interact with the unknown channel $\mathcal{E}$ by preparing any input state, passing it through $\mathcal{E}$, and performing a measurement on the output. Additionally, we are given access to training examples from some distribution $\mathcal{D} = D^{\otimes n}$ over product states, and their corresponding *classical descriptions*. Because product states over $n$ qubits can be efficiently represent efficiently using $O(n)$ bits, the training set can be stored efficiently on classical computers. The standard approach to represent a product state on a classical computer is as follows. For each state $\rho = \otimes_{i=1}^{n} |\psi_i\rangle$ sampled from $\mathcal{D}$, the classical description can be given by their 1-qubit Pauli expectation values: $\mathrm{Tr}(P|\psi_i\rangle\langle\psi_i|)$ for all $i \in [n]$ ranging over each qubit and $P \in \{X, Y, Z\}$.

Given these classical samples from $\mathcal{D}$ and the ability to query $\mathcal{E}$, the learning goal is to produce a hypothesis $f'$ which takes as input the classical description of a product state $\rho$ and outputs an estimate for $\mathrm{Tr}(\mathcal{O}\,\mathcal{E}[\rho])$. Formally, we want this hypothesis to have small test loss in the sense that $\mathbb{E}_{\rho\sim\mathcal{D}}[(\mathrm{Tr}(\mathcal{O}\,\mathcal{E}[\rho]) - \mathrm{Tr}(f'(\rho)))^2] \leq \varepsilon$ with probability at least $1-\delta$ over the randomness of the learning algorithm and the training examples from $\mathcal{D}$.

## 3. Learning an unknown quantum channel

In this section we will prove Theorem 1. First we will show that under any product distribution with second moment $\mathcal{S}$ such that $\|\mathcal{S}\|_{\mathsf{op}} \leq 1-\eta$ for some $\eta \in (0,1)$, every observable has a low-degree approximation. A distribution has $\|\mathcal{S}\|_{\mathsf{op}} = 1$ only if it is effectively a classical distribution; i.e. it is supported on two antipodal points in the Bloch sphere. So we are showing that any product distribution which is "spread out" within the Bloch sphere behaves well with low-degree approximation.

To do this, we cannot use exactly the same argument as in Appendix B, because there is not necessarily an orthonormal basis for our product distribution $D^{\otimes n}$ that is a "stretched" basis for some other distribution over the Bloch sphere. Instead, we compare the variance of the observable under $D^{\otimes n}$ to the quantity $\mathbb{E}_{\rho\sim D^{\otimes n}}[\mathrm{Tr}(\mathcal{O}^2\rho)]$. This allows us to use the boundedness of $\mathcal{O}$ to derive bounds for the contribution of the degree-$d$ part to the variance of $\mathcal{O}$ under $D^{\otimes n}$. The learning algorithm will find the low-degree approximation by linear regression over the degree-$\log(1/\varepsilon)$ Pauli coefficients. The notion of low-degreeness will be with respect to a basis adapted to $D^{\otimes n}$.

**Definition 10** *Let $D$ be a distribution over the Bloch sphere. Let $U^\dagger A U$ be the eigendecomposition of $\overline{\rho} = \mathbb{E}_D[\rho]$. Let $\tilde{X}, \tilde{Y}, \tilde{Z} := U^\dagger X U, U^\dagger Y U, \frac{U^\dagger Z U - \mathrm{Tr}(\overline{\rho} U^\dagger Z U) I}{\sqrt{1-\mathrm{Tr}(\overline{\rho} U^\dagger Z U)^2}}$.*

*Let $B = \{I, \tilde{X}, \tilde{Y}, \tilde{Z}\}^{\otimes n}$. The degree of $P \in B$ is the number of non-identity elements in the product. The degree of a linear combination $\sum_{P\in B} \alpha_P P$ of elements in $B$ is the largest degree of $P \in B$ such that $\alpha_P \neq 0$.*

The fact that the degree is defined for any observable (which is equivalent to $\tilde{X}, \tilde{Y}, \tilde{Z}$ forming a basis of the space of operators) is the content of Lemma 12. The existence of low-degree approximation is guaranteed by the following lemma.

**Lemma 11 (Low-degree approximation)** *Let $\mathcal{O}$ be a bounded $n$-qubit observable and let $D$ be a distribution over the Bloch sphere with mean $\boldsymbol{\mu}$ and Pauli second moment matrix $\mathcal{S}$ such that $\|\mathcal{S}\| \leq 1 - \eta$ for some $\eta \in (0, 1)$. Then there exists a degree-$d$ observable $\mathcal{O}^{\leq d}$ and a constant $\eta' \in (0, 1)$ such that*

$$\mathbb{E}_{\rho\sim D^{\otimes n}}[(\mathrm{Tr}(\mathcal{O}\rho) - \mathrm{Tr}(\mathcal{O}^{\leq d}\rho))^2] \leq (1-\eta')^d,$$

*where $\eta'$ is a function of $\eta$.*

Once Lemma 11 is established, Theorem 1 follows readily from an application of Corollary 17.

**Proof of Theorem 1, using Lemma 11** Let $1 - \eta$ be a known upper bound on $\|\mathcal{S}\|_{\mathsf{op}}$. We assume the access model of Section 2.2, where we get a set $S$ of examples $[\mathrm{Tr}(P\rho)]$ for 1-local $P$. Our algorithm is as follows:

1. Compute $\eta'$ as in the last line of Lemma 15. Let $d \coloneqq O(\log(1/\varepsilon)/\log(1/(1-\eta')))$.
2. Draw a set $S$ of size $n^d \cdot \log(1/\delta)$, and initialize $S'$ to be empty.
3. For each $x \in S$, prepare a set $T$ of $\frac{1}{\varepsilon^2}\log(|S|/\delta)$ copies of the state $\rho$ that matches the 1-local expectations of $x$. Let $\mathrm{est}(\mathrm{Tr}(\mathcal{O}\mathcal{E}[\rho]))$ be the estimate of $\mathrm{Tr}(\mathcal{O}\mathcal{E}[\rho])$, where for each $\rho \in T$, we measure with respect to $\{\mathcal{E}^\dagger[\mathcal{O}], I - \mathcal{E}^\dagger[\mathcal{O}]\}$, and $\mathrm{est}(\mathrm{Tr}(\mathcal{O}\mathcal{E}[\rho]))$ is the empirical probability of measuring the first outcome. Add
$$(x^{\otimes \log(1/\varepsilon)/\log(1/(1-\eta'))}, \mathrm{est}(\mathrm{Tr}(\mathcal{O}\mathcal{E}[\rho])))$$
to the set $S'$.
4. Run linear regression on $S'$ and output the returned hypothesis $h$.

By Hoeffding's inequality, each estimate of $\mathrm{Tr}(\mathcal{O}\mathcal{E}[\rho])$ is within $\varepsilon$ of its expectation with probability $1-\exp(-\Omega(|T|\varepsilon^2))$. By union bound over the $|S|$ estimates, with probability $\geq 1-\delta$, all estimates are within $\varepsilon$ of $\mathrm{Tr}(\mathcal{O}\mathcal{E}[\rho])$.

The time, sample, and error bounds follow from Lemma 11, which guarantees that $\mathcal{E}^\dagger[\mathcal{O}]$ is $\varepsilon$-close to some degree-$d$ polynomial. This implies the labels of our sample set are $2\varepsilon$-close to such a polynomial. The dimension of the linear regression problem is $\leq \binom{n}{d} \cdot 3^d \leq n^{O(d)}$, as there are 3 choices for each non-identity component. Then by Corollary 17, linear regression has time and sample complexity

$$n^{O(\log(1/\varepsilon)/\log(1/(1-\eta')))} \cdot \log(1/\delta)$$

and outputs a hypothesis $h$ such that $\mathbb{E}_{\rho\sim D^{\otimes n}}[(\mathrm{Tr}(\mathcal{O}\mathcal{E}[\rho]) - \mathrm{Tr}(h\rho))^2] \leq O(\varepsilon)$. ■

The proof of Lemma 11 follows from three technical Lemmas. The first gives, for any product distribution over $n$-qubit states, a (non-orthogonal) decomposition of any observable into operators which are centered and bounded in variance with respect to that distribution.

**Lemma 12** *Let $D$ be a distribution over the Bloch sphere. Let $U^\dagger AU$ be the eigendecomposition of $\overline{\rho} = \mathbb{E}_D[\rho]$, and let $\tilde{X}, \tilde{Y}, \tilde{Z} := U^\dagger XU, U^\dagger YU, \frac{U^\dagger ZU - \mathrm{Tr}(\overline{\rho}U^\dagger ZU)I}{\sqrt{1-\mathrm{Tr}(\overline{\rho}U^\dagger ZU)^2}}$. Then $\{I, \tilde{X}, \tilde{Y}, \tilde{Z}\}$ is a basis for the set of $2 \times 2$ unitary Hermitian matrices, and thus every $n$-qubit observable $\mathcal{O}$ can be written as*

$$\mathcal{O} = \sum_{P \in B} \widehat{\mathcal{O}}(P)P$$

*for $B = \{I, \tilde{X}, \tilde{Y}, \tilde{Z}\}^{\otimes n}$. Furthermore, each non-identity $P \in B$ satisfies $\mathbb{E}_{\rho\sim D^{\otimes n}}[\mathrm{Tr}(P\rho)] = 0$ and $\mathbb{E}_{\rho\sim D^{\otimes n}}[\mathrm{Tr}(P\rho)^2] \leq 1$.*

**Proof** It is clear that $\{I, \tilde{X}, \tilde{Y}, \tilde{Z}\}$ is linearly independent and thus $B$ forms a basis for any $n$-qubit observable. Note that $\mathrm{Tr}(P\rho) = \prod_{i=1}^n \mathrm{Tr}(P_i\rho_i)$ as $\rho$ is a product state, so we can restrict the analysis to single qubit states drawn from $D$.

For $P = \tilde{X}$ or $\tilde{Y}$, it is clear that $\mathbb{E}_{\rho\sim D}[\mathrm{Tr}(P\rho)] = 0$ and $\mathbb{E}_{\rho\sim D}[\mathrm{Tr}(P\rho)^2] = 1$. For $\tilde{Z}$, we have

$$\mathbb{E}_{\rho\sim D}[\mathrm{Tr}(\tilde{Z}\rho)] = \frac{\mathbb{E}[\mathrm{Tr}(\rho U^\dagger ZU)] - \mathrm{Tr}(\overline{\rho}U^\dagger ZU)}{\sqrt{1-\mathrm{Tr}(\overline{\rho}U^\dagger ZU)^2}} = 0\,.$$

Moreover,

$$\mathbb{E}_{\rho\sim D}[\mathrm{Tr}(\tilde{Z}\rho)^2] = \frac{\mathrm{Var}[\mathrm{Tr}(\rho U^\dagger ZU)]}{1-\mathrm{Tr}(\overline{\rho}U^\dagger ZU)^2} \leq 1\,,$$

because $\mathrm{Var}[\mathrm{Tr}(\rho U^\dagger ZU)] + \mathbb{E}[\mathrm{Tr}(\rho U^\dagger ZU)]^2 = \mathbb{E}[\mathrm{Tr}(\rho U^\dagger ZU)^2] \leq 1$. ■

**Remark 13** *Going forward, we will assume w.l.o.g. that the mean of the distribution $\mathbb{E}_D[\rho]$ is diagonal because we can always transform the basis according to $U$. Thus, we will assume that the mean state $\mathbb{E}_D[\rho] = \frac{1}{2}(I + \mu Z)$ and the mean Bloch vector $\boldsymbol{\mu} = (0, 0, \mu)$ for some $\mu \in [-1, 1]$,*

We next prove two important components required in the proof of Lemma 11. The first lemma gives an eigenvalue lower bound on a Hermitian matrix that arises when we expand $\mathrm{Tr}(\mathcal{O}^2\,\mathbb{E}[\rho])$.

**Lemma 14** *Let $\mu \in (-1, 1)$ and*

$$\widetilde{M} = \begin{pmatrix} 1 & i\mu \\ -i\mu & 1 \end{pmatrix}.$$

*Then $\lambda_{\min}(\mathrm{Re}(\widetilde{M}^{\otimes k})) \geq (1-\mu^2)^{k/2}$ for any $k \in \mathbb{N}$.*

**Proof** Note that $\widetilde{M} = I + \mu Y$ and $Y$ is imaginary. Thus,

$$\mathrm{Re}((I + \mu Y)^{\otimes k}) = \sum_{P \in \{I,Y\}^k : |P| \text{ even}} \mu^{|P|} \bigotimes_{i=1}^k P_i\,.$$

Note also that the eigenvalues of the above remain the same if we replace the $Y$'s with $Z$'s. Thus, the eigenvalues of the above can be indexed by $x \in \{\pm 1\}^k$, and can be expressed as follows:

$$\lambda(x) = \sum_{S \subseteq [k] : |S| \text{ even}} \mu^{|S|} x^S\,.$$

Let $f : \mathbb{R}^k \to \mathbb{R}$ be the function $f(x) = \prod_{i=1}^k (1 + x_i) = \sum_{S \subseteq [k]} x^S$. Then, we have $\lambda(x) = \frac{1}{2}(f(\mu x) + f(-\mu x))$. By the AM-GM inequality,

$$\lambda(x) \geq \sqrt{f(\mu x) f(-\mu x)} = \prod_{i=1}^k \sqrt{(1 + \mu x_i)(1 - \mu x_i)} = (1 - \mu^2)^{k/2} .$$

■

The next lemma gives an upper bound on the Pauli covariance matrix $\Sigma$ (recall Definition 8) scaled by a specific diagonal matrix.

**Lemma 15** *Let $D$ be a distribution over the Bloch sphere with mean $\boldsymbol{\mu} = (0, 0, \mu)$, Pauli second moment matrix $\mathcal{S}$ such that $\|\mathcal{S}\|_{\mathsf{op}} \leq 1 - \eta$ for some $\eta \in (0, 1)$, and Pauli covariance matrix $\Sigma$. Let $\Delta = \operatorname{diag}((1 - \mu^2)^{-1/4}, (1 - \mu^2)^{-1/4}, (1 - \mu^2)^{-1/2})$. Then there exists $\eta' \in (0, 1)$ such that $\|\Delta \Sigma \Delta\|_{\mathsf{op}} \leq 1 - \eta'$, where $\eta'$ is a function of $\eta$.*

**Proof** We split into two cases based on whether $\mu^2 \geq \eta/2$. The case where $\mu^2 < \eta/2$ is the simpler case: since $\|\Sigma_{\mathsf{op}}\| \leq \|\mathcal{S}\|_{\mathsf{op}} \leq 1 - \eta$ and $\|\Delta\|_{\mathsf{op}}^2 \leq (1 - \mu^2)^{-1}$, we have

$$\|\Delta \Sigma \Delta\|_{\mathsf{op}} \leq \frac{\|\Sigma\|_{\mathsf{op}}}{1 - \mu^2} \leq \frac{1 - \eta}{1 - \mu^2} \leq \frac{1 - \eta}{1 - \eta/2} \leq 1 - \eta/2 .$$

Now we consider the case where the inequality is not satisfied. We note that $\Sigma = \mathcal{S} - \boldsymbol{\mu}\boldsymbol{\mu}^\top$ and $\boldsymbol{\mu} = (0, 0, \mu)$, thus we have that $\Delta \Sigma \Delta$ has the following block structure:

$$\Delta \Sigma \Delta = \begin{pmatrix} \frac{\mathcal{S}_{2\times 2}}{\sqrt{1-\mu^2}} & b \\ b^\dagger & \frac{\mathcal{S}_{zz} - \mu^2}{1 - \mu^2} \end{pmatrix} ,$$

where $\mathcal{S}_{2\times 2}$ is the top left $2 \times 2$ block of $\mathcal{S}$, $\mathcal{S}_{zz}$ is the bottom right entry, and $b$ is the remaining part (its values will not be directly relevant). Note also that $\Delta \Sigma \Delta \succeq 0$, and it is well-known that any PSD matrix of the form $\begin{bmatrix} A & b \\ b^\dagger & c \end{bmatrix} \succeq 0$ has operator norm at most $\|A\|_{\mathsf{op}} + c$. We thus know that

$$\|\Delta \Sigma \Delta\|_{\mathsf{op}} \leq \frac{\mathcal{S}_{zz} - \mu^2}{1 - \mu^2} + \frac{\|\mathcal{S}_{2\times 2}\|_{\mathsf{op}}}{\sqrt{1 - \mu^2}} \leq \frac{\mathcal{S}_{zz} - \mu^2}{1 - \mu^2} + \frac{\operatorname{Tr}(\mathcal{S}_{2\times 2})}{\sqrt{1 - \mu^2}} \leq \frac{\mathcal{S}_{zz} - \mu^2}{1 - \mu^2} + \frac{1 - \mathcal{S}_{zz}}{\sqrt{1 - \mu^2}} ,$$

where we use the fact that $1 = \operatorname{Tr}(\mathcal{S}) = \operatorname{Tr}(\mathcal{S}_{2\times 2}) + \mathcal{S}_{zz}$. Then we have

$$\begin{aligned} \|\Delta \Sigma \Delta\|_{\mathsf{op}} &\leq \frac{\mathcal{S}_{zz} - \mu^2}{1 - \mu^2} + \frac{1 - \mathcal{S}_{zz}}{\sqrt{1 - \mu^2}} \\ &= \mathcal{S}_{zz}\Big(\frac{1}{1 - \mu^2} - \frac{1}{\sqrt{1 - \mu^2}}\Big) + \frac{1}{\sqrt{1 - \mu^2}} - \frac{\mu^2}{1 - \mu^2} \\ &\leq (1 - \eta)\Big(\frac{1}{1 - \mu^2} - \frac{1}{\sqrt{1 - \mu^2}}\Big) + \frac{1}{\sqrt{1 - \mu^2}} - \frac{\mu^2}{1 - \mu^2} \\ &= 1 - \eta\Big(\frac{1}{1 - \mu^2} - \frac{1}{\sqrt{1 - \mu^2}}\Big) \\ &\leq 1 - \eta \frac{1 - \sqrt{1 - \eta/2}}{1 - \eta/2} . \end{aligned}$$

The third line is by the fact that $\mathcal{S}_{zz} \le \|\mathcal{S}\|_{\mathsf{op}} \le 1-\eta$, and the last inequality is because the function $\frac{1}{1-\mu^2} - \frac{1}{\sqrt{1-\mu^2}}$ is increasing with $\mu^2$ and that we are in the $\mu^2 \ge \eta/2$ case.

Thus, combining the two cases, there always exists $\eta'$ such that $\|\Delta\Sigma\Delta\|_{\mathsf{op}} \le 1-\eta'$. Specifically, we have $\eta' \ge \min\left\{\eta\frac{1-\sqrt{1-\eta/2}}{1-\eta/2},\ \eta/2\right\} > 0$. ■

Now we prove Lemma 11 which shows the existence of a low-degree approximator.

**Proof of Lemma 11** Assume w.l.o.g., as in Remark 13, that $\mathbb{E}_D[\rho] = \frac{1}{2}(I+\mu Z)$ and $\boldsymbol{\mu} = (0,0,\mu)$. Let $\mathcal{O}$ be expressed in the basis $B = \{I, X, Y, \frac{Z-\mu I}{\sqrt{1-\mu^2}}\}^{\otimes n}$ as in Lemma 12. For a subset $S \subseteq [n]$, we will denote by $\{P \in B : P \sim S\}$ the set of basis elements with non-identity components at indices in $S$ and identity components at indices in $\overline{S}$. We will also denote by $|P|$ the number of non-identity components in $P$.

Let $\mathcal{O}^{>d} := \sum_{|S|>d} \sum_{P\sim S} \widehat{\mathcal{O}}(P)P$. We will show that $\mathcal{O}^{>d}$ satisfies $\mathbb{E}_{\rho\sim D^{\otimes n}}[\mathrm{Tr}(\mathcal{O}^{>d}\rho)^2] \le (1-\eta')^d$ where $\eta' \in (0,1)$ depends on $\eta$.

We first note a bound on the related quantity $\mathbb{E}_{\rho\sim D^{\otimes n}}[\mathrm{Tr}(\mathcal{O}^2\rho)]$. Because $\|\mathcal{O}\|_{\mathsf{op}} \le 1$ and $\mathrm{Tr}(\mathbb{E}[\rho]) \le 1$, $\mathrm{Tr}(\mathcal{O}^2\,\mathbb{E}[\rho]) \le 1$ as well. We will expand this quantity as a quadratic form.

$$\begin{aligned}
\mathbb{E}_{\rho\sim D^{\otimes n}}[\mathrm{Tr}(\mathcal{O}^2\rho)] &= \sum_{P,Q\in B} \widehat{\mathcal{O}}(P)\widehat{\mathcal{O}}(Q)\mathrm{Tr}(PQ\,\mathbb{E}[\rho]) \\
&= \sum_{P,Q\in B} \widehat{\mathcal{O}}(P)\widehat{\mathcal{O}}(Q)\mathrm{Tr}(\otimes_{i=1}^n P_iQ_i\,\mathbb{E}[\rho_i]) \\
&= \sum_{P,Q\in B} \widehat{\mathcal{O}}(P)\widehat{\mathcal{O}}(Q)\prod_{i=1}^n \mathrm{Tr}(P_iQ_i\,\mathbb{E}[\rho_i])\,.
\end{aligned}$$

Note that the product is 0 whenever exactly one of $P_i$, $Q_i$ is $I$ for any $i \in [n]$ (as $\mathrm{Tr}((Z-\mu I)\cdot(I+\mu Z)) = 0$). Therefore, we can partition the terms into groups that share a subset of identity variables:

$$\begin{aligned}
\mathbb{E}_{\rho\sim D^{\otimes n}}[\mathrm{Tr}(\mathcal{O}^2\rho)] &= \sum_{S\subseteq[n]} \sum_{P,Q\sim S} \widehat{\mathcal{O}}(P)\widehat{\mathcal{O}}(Q)\prod_{i=1}^n \mathrm{Tr}(P_iQ_i\,\mathbb{E}[\rho_i]) \\
&= \sum_{S\subseteq[n]} \widehat{\mathcal{O}}_S^\dagger M^{\otimes|S|}\widehat{\mathcal{O}}_S\,,
\end{aligned}$$

where $\widehat{\mathcal{O}}_S$ is the vector of coefficients for the set $\{P : P \sim S\}$, and $M$ is the $3\times 3$ matrix such that $M_{P,Q} = \mathrm{Tr}(PQ\,\mathbb{E}_D[\rho])$:

$$M = \begin{pmatrix} 1 & i\mu & 0 \\ -i\mu & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Here, the entry $i\mu$ arises because $XY = iZ$ and $\mathrm{Tr}(XY\cdot\frac{1}{2}(I+\mu Z)) = i\mu$. Since $M$ is positive semidefinite, we have $\widehat{\mathcal{O}}_S^\dagger M^{\otimes|S|}\widehat{\mathcal{O}}_S \ge 0$ for all $S$, and therefore we have

$$1 \ge \mathbb{E}_{\rho\sim D^{\otimes n}}[\mathrm{Tr}(\mathcal{O}^2\rho)] = \sum_{S\subseteq[n]} \widehat{\mathcal{O}}_S^\dagger M^{\otimes|S|}\widehat{\mathcal{O}}_S \ge \sum_{S\subseteq[n]:|S|>d} \widehat{\mathcal{O}}_S^\dagger M^{\otimes|S|}\widehat{\mathcal{O}}_S\,. \tag{2}$$

Now we will write the desired quantity $\mathbb{E}_{\rho\sim D^{\otimes n}}[\mathrm{Tr}(\mathcal{O}^{>d}\rho)^2]$ as a quadratic form as well:

$$\mathbb{E}_{\rho\sim D^{\otimes n}}[\mathrm{Tr}(\mathcal{O}^{>d}\rho)^2] = \sum_{P,Q\subseteq B\colon |P|,|Q|>d} \widehat{\mathcal{O}}(P)\widehat{\mathcal{O}}(Q) \prod_{i\in[n]} \mathbb{E}[\mathrm{Tr}(P_i\rho)\mathrm{Tr}(Q_i\rho)]\,.$$

Similarly, the product is 0 whenever exactly one of $P_i$ and $Q_i$ is identity (by the guarantee of Lemma 12 that $\mathbb{E}[\mathrm{Tr}(P_i\rho)] = 0$ for $P_i \neq I$), so we can make the same partition:

$$\mathbb{E}_{\rho\sim D^{\otimes n}}[\mathrm{Tr}(\mathcal{O}^{>d}\rho)^2] = \sum_{|S|>d} \widehat{\mathcal{O}}_S^\dagger M'^{\otimes|S|}\widehat{\mathcal{O}}\,. \tag{3}$$

Here $M'$ is $3\times 3$ matrix such that $M'_{PQ} = \mathbb{E}_{\rho\sim D}[\mathrm{Tr}(P\rho)\mathrm{Tr}(Q\rho)]$ for $P,Q \in \{X, Y, \frac{Z-\mu I}{\sqrt{1-\mu^2}}\}$; in other words, it is the second moment matrix of the non-identity elements of our biased Pauli basis. Below, we show the entries of $M'$ in terms of the Pauli covariance matrix $\Sigma$:

$$M' = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} & \frac{\Sigma_{xz}}{\sqrt{1-\mu^2}} \\ \Sigma_{xy} & \Sigma_{yy} & \frac{\Sigma_{yz}}{\sqrt{1-\mu^2}} \\ \frac{\Sigma_{xz}}{\sqrt{1-\mu^2}} & \frac{\Sigma_{yz}}{\sqrt{1-\mu^2}} & \frac{\Sigma_{zz}}{1-\mu^2} \end{pmatrix}.$$

Our aim is to bound $M'$ in terms of $M$ in order to show the existence of some $\eta' \in (0,1)$ such that

$$\mathbb{E}[\mathrm{Tr}(\mathcal{O}^{>d}\rho)^2] \le (1-\eta')^d \cdot \mathrm{Tr}((\mathcal{O}^{>d})^2\,\mathbb{E}[\rho]) \le (1-\eta')^d.$$

Comparing Eq. (2) and (3), it suffices to show that

$$(1-\eta')^{|S|}\widehat{\mathcal{O}}_S^\dagger M^{\otimes|S|}\widehat{\mathcal{O}}_S \ge \widehat{\mathcal{O}}_S^\dagger M'^{\otimes|S|}\widehat{\mathcal{O}}_S$$

for all vectors $\mathcal{O}_S$. Crucially, since the Pauli coefficients $\mathcal{O}_S$ must be *real*, it suffices to prove that

$$M'^{\otimes|S|} \preceq (1-\eta')^{|S|}\cdot \mathrm{Re}(M^{\otimes|S|})\,. \tag{4}$$

Let $M_{2\times 2}$ be the top left $2\times 2$ block in $M$, which is exactly the matrix in Lemma 14. From Lemma 14, we have $\lambda_{\min}(\mathrm{Re}(M_{2\times 2}^{\otimes|S|})) \ge (1-\mu^2)^{|S|/2}$. By the block structure of $M$, it follows that $\mathrm{Re}(M^{\otimes|S|}) \succeq \mathrm{diag}(\sqrt{1-\mu^2}, \sqrt{1-\mu^2}, 1)^{\otimes|S|}$. Thus, we can establish Eq. (4) by proving that

$$M' \preceq (1-\eta')\cdot \mathrm{diag}\Big(\sqrt{1-\mu^2}, \sqrt{1-\mu^2}, 1\Big)\,. \tag{5}$$

Now, note that $M'$ can be written as $\tilde{\Delta}\Sigma\tilde{\Delta}$, where $\Sigma$ is the covariance matrix of $D$ and $\tilde{\Delta} = \mathrm{diag}(1, 1, (1-\mu^2)^{-1/2})$. Letting $\Delta = \mathrm{diag}((1-\mu^2)^{-1/4}, (1-\mu^2)^{-1/4}, (1-\mu^2)^{-1/2})$ (which is the diagonal matrix defined in Lemma 15), we can see that Eq. (5) is equivalent to

$$\|\Delta M'\Delta\|_{\mathsf{op}} \le 1-\eta',$$

By Lemma 15, this is true by our assumption that $\|\mathcal{S}\|_{\mathsf{op}} \leq 1-\eta$. This establishes Eq. (5) and thus Eq. (4), and from Eq. (2) and (3), we have

$$\begin{aligned}1 \geq \mathbb{E}_{\rho\sim D^{\otimes n}}[\mathrm{Tr}(\mathcal{O}^2\rho)] &\geq \sum_{|S|>d} \widehat{\mathcal{O}}_S^\dagger M^{\otimes|S|}\widehat{\mathcal{O}}_S \\ &\geq (1-\eta')^{-d}\sum_{|S|>d} \widehat{\mathcal{O}}_S^\dagger M'^{\otimes|S|}\widehat{\mathcal{O}}_S \\ &= (1-\eta')^{-d}\,\mathbb{E}_{\rho\sim D^{\otimes n}}[\mathrm{Tr}(\mathcal{O}^{>d}\rho)^2]\,.\end{aligned}$$

Therefore, $\mathbb{E}_{\rho\sim D^{\otimes n}}[(\mathrm{Tr}(\mathcal{O}\rho)-\mathrm{Tr}(\mathcal{O}^{\leq d}\rho))^2] \leq (1-\eta')^d$ as claimed. ■

## Acknowledgments

SC and HH would like to thank Ryan O’Donnell for a helpful discussion at an early stage of this project. Much of this work was completed while JD and JL were interns at Microsoft Research.

## References

Srinivasan Arunachalam, Arkopal Dutt, Francisco Escudero Gutiérrez, and Carlos Palazuelos. Learning low-degree quantum objects. *arXiv preprint arXiv:2405.10933*, 2024.

Zongbo Bao and Penghui Yao. Nearly optimal algorithms for testing and learning quantum junta channels. *arXiv preprint arXiv:2305.12097*, 2023.

Thomas Chen, Shivam Nadimpalli, and Henry Yuen. Testing and learning quantum juntas nearly optimally. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1163–1185. SIAM, 2023.

Cristina Cirstoiu, Zoe Holmes, Joseph Iosue, Lukasz Cincio, Patrick J Coles, and Andrew Sornborger. Variational fast forwarding for quantum simulation beyond the coherence time. *npj Quantum Information*, 6(1):82, 2020.

Alexandros Eskenazis and Paata Ivanisvili. Learning low-degree functions from a logarithmic number of random queries. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 203–207, 2022.

Steven T Flammia and Ryan O’Donnell. Pauli error estimation via population recovery. *Quantum*, 5:549, 2021.

Merrick L Furst, Jeffrey C Jackson, and Sean W Smith. Improved learning of ac 0 functions. In *COLT*, volume 91, pages 317–325, 1991.

Joe Gibbs, Zoe Holmes, Matthias C Caro, Nicholas Ezzell, Hsin-Yuan Huang, Lukasz Cincio, Andrew T Sornborger, and Patrick J Coles. Dynamical simulation via quantum machine learning with provable generalization. *Physical Review Research*, 6(1):013241, 2024.

Robin Harper, Steven T Flammia, and Joel J Wallman. Efficient learning of quantum noise. *Nature Physics*, 16(12):1184–1188, 2020.

Robin Harper, Wenjun Yu, and Steven T Flammia. Fast estimation of sparse quantum noise. *PRX Quantum*, 2(1):010322, 2021.

Patrick Hayden and Geoffrey Penington. Learning the alpha-bits of black holes. *Journal of High Energy Physics*, 2019(12):1–55, 2019.

Patrick Hayden and John Preskill. Black holes as mirrors: quantum information in random subsystems. *Journal of high energy physics*, 2007(09):120, 2007.

Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics*, 16(10):1050–1057, 2020.

Hsin-Yuan Huang, Richard Kueng, and John Preskill. Information-theoretic bounds on quantum advantage in machine learning. *Physical Review Letters*, 126(19):190505, 2021.

Hsin-Yuan Huang, Michael Broughton, Jordan Cotler, Sitan Chen, Jerry Li, Masoud Mohseni, Hartmut Neven, Ryan Babbush, Richard Kueng, John Preskill, et al. Quantum advantage in learning from experiments. *Science*, 376(6598):1182–1186, 2022.

Hsin-Yuan Huang, Sitan Chen, and John Preskill. Learning to predict arbitrary quantum processes. *PRX Quantum*, 4(4):040337, 2023.

Hsin-Yuan Huang, Yunchao Liu, Michael Broughton, Isaac Kim, Anurag Anshu, Zeph Landau, and Jarrod R McClean. Learning shallow quantum circuits. *arXiv preprint arXiv:2401.10095*, 2024.

Adam Tauman Kalai, Adam R Klivans, Yishay Mansour, and Rocco A Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

Ohad Klein, Joseph Slote, Alexander Volberg, and Haonan Zhang. Quantum and classical low-degree learning via a dimension-free remez inequality. *arXiv preprint arXiv:2301.01438*, 2023.

Adam R Klivans and Rocco Servedio. Learning DNF in time $2^{\tilde{O}(n^{1/3})}$. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 258–265, 2001.

Eyal Kushilevitz and Yishay Mansour. Learning decision trees using the fourier spectrum. In *Proceedings of the twenty-third annual ACM symposium on Theory of computing*, pages 455–464, 1991.

Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, fourier transform, and learnability. *Journal of the ACM (JACM)*, 40(3):607–620, 1993.

Grigorii Aleksandrovich Margulis. Probabilistic characteristics of graphs with large connectivity. *Problemy peredachi informatsii*, 10(2):101–108, 1974.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. 2018.

Masoud Mohseni, Ali T Rezakhani, and Daniel A Lidar. Quantum-process tomography: Resource analysis of different strategies. *Physical Review A*, 77(3):032322, 2008.

Ashley Montanaro and Tobias J Osborne. Quantum boolean functions. *arXiv preprint arXiv:0810.2435*, 2008.

Shivam Nadimpalli, Natalie Parham, Francisca Vasconcelos, and Henry Yuen. On the Pauli Spectrum of QAC0. *arXiv preprint arXiv:2311.09631*, 2023.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

Geoff Penington, Stephen H Shenker, Douglas Stanford, and Zhenbin Yang. Replica wormholes and the black hole interior. *Journal of High Energy Physics*, 2022(3):1–87, 2022.

Theodore J Rivlin. *Chebyshev polynomials*. Courier Dover Publications, 1990.

Cambyse Rouzé, Melchior Wirth, and Haonan Zhang. Quantum talagrand, kkl and friedgut's theorems and the learnability of quantum boolean functions. *arXiv preprint arXiv:2209.07279*, 2022.

Lucio Russo. On the critical percolation probabilities. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 56(2):229–237, 1981.

Lucio Russo. An approximate zero-one law. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 61(1):129–139, 1982.

Ewout Van Den Berg, Zlatko K Minev, Abhinav Kandala, and Kristan Temme. Probabilistic error cancellation with sparse pauli–lindblad models on noisy quantum processors. *Nature Physics*, 19 (8):1116–1121, 2023.

Alexander Volberg and Haonan Zhang. Noncommutative Bohnenblust–Hille inequalities. *Mathematische Annalen*, pages 1–20, 2023.

Lisa Yang and Netta Engelhardt. The complexity of learning (pseudo) random dynamics of black holes and other chaotic systems. *arXiv preprint arXiv:2302.11013*, 2023.

## Appendix A. Further related work

**Other works on quantum Boolean analysis.** In (Montanaro and Osborne, 2008), it was proposed to study Pauli decompositions of *Hermitian* unitaries as the natural quantum analogue of Boolean functions. One notable follow-up work (Rouzé et al., 2022) proved various quantum versions of classical Fourier analytic results like Talagrand's variance inequality and the KKL theorem in this setting, and also obtained corollaries about learning Hermitian unitaries in Frobenius norm given oracle access (see also Chen et al. (2023); Bao and Yao (2023) for the non-Hermitian case). Recently, Nadimpalli et al. (2023) considered the Pauli spectrum of the *Choi representation* of quantum channels and proved low-degree concentration for channels implemented by $\mathsf{QAC}^0$. One technical difference with our work is that these notions of Pauli decomposition are specific to the channel, whereas the object whose Pauli decomposition we consider is specific to the Heisenberg-evolved operator $\mathcal{E}^\dagger[\mathcal{O}]$. Additionally, we note that all of the above mentioned works focus on questions more akin to learning a full description of the channel and thus are inherently tied to channels with specific structure.

**Classical low-degree learning.** The general technique of low-degree approximation in classical learning theory is too prevalent to do full justice to in this section. This idea of learning Boolean functions by approximating their low-degree Fourier truncation was first introduced in the seminal work of Linial et al. (1993). Fourier-analytic techniques have been used to obtain new classical learning results for various concept classes like decision trees (Kushilevitz and Mansour, 1991), linear threshold functions (Kalai et al., 2008), Boolean formulas (Klivans and Servedio, 2001), low-degree polynomials (Eskenazis and Ivanisvili, 2022), and more.

While Fourier analysis over biased distributions dates back to early work of Margulis and Russo (Margulis, 1974; Russo, 1981, 1982), it was first applied in a learning-theoretic context in Furst et al. (1991), extending the aforementioned result of Linial et al. (1993).

### A.1. Generalization bounds for learning

For our learning protocol, we will use the following elementary results about linear and polynomial regression:

**Fact 16 (Rademacher complexity generalization bound Mohri et al. (2018))** *Let $\mathcal{F}$ be the class of bounded linear functions $[-1,1]^d \to [-B,B]$, and let $\ell$ be a loss function with Lipschitz constant $L$ and a uniform upper bound of $c$. With probability $1-\delta$ over the choice of a training set $S$ of size $m$ drawn i.i.d. from distribution $\mathcal{D}$,*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(x),y)] \le \mathbb{E}_{(x,y)\sim S}[\ell(f(x),y)] + 4LB\sqrt{\frac{d}{m}} + 2c\sqrt{\frac{\log 1/\delta}{2m}}\,.$$

**Corollary 17** *Let $f$ be a function that is $\varepsilon$-close to a degree-$\le d$ polynomial $f^\star$:*

$$\mathbb{E}_{x\sim\mathcal{D}}[(f(x)-f^\star(x))^2] \le \varepsilon.$$

*Then linear regression over the set of degree-$d$ polynomials with coefficients in $[-1,1]$ has time and sample complexity $\mathrm{poly}(n^d, \log 1/\varepsilon)\cdot\log 1/\delta$ and finds $h$ such that*

$$\mathbb{E}_{x\sim\mathcal{D}}[(f(x)-h(x))^2] \le O(\varepsilon)$$

*with probability $1-\delta$.*

## Appendix B. Warm-up: the classical case

In this section we will prove Theorem 3, which is a special case of Theorem 1 where the distribution is classical, i.e. supported only on the $Z$ component.

The key ingredient in the proof is to show that for any function which is $L^2$-integrable with respect to a product distribution over $[-(1-\eta), 1-\eta]^n$ and whose extension to the hypercube is bounded, the function admits a "low-degree" approximation under an appropriate orthonormal basis. Roughly, the intuition is that the space of linear functions over a distribution $D$ on $[-(1-\eta), 1-\eta]$ has an orthonormal basis which is a $(1-\eta)$-scaling of a basis for a distribution on $\{-1, 1\}$. Therefore, the space of multilinear functions over the corresponding product distribution $D^{\otimes n}$ has a basis whose degree-$d$ components are scaled by $(1-\eta)^d$. This, combined with the assumption that the function is bounded on the hypercube, allows us to conclude that the contribution of the degree-$d$ component to the variance of $f$ over $D^{\otimes n}$ is at most $(1-\eta)^{2d}$.

We prove this structural result in Section B.1 and conclude the proof of Theorem 3 in Section B.2.

### B.1. Existence of low-degree approximation

We first review some basic facts about classical biased Fourier analysis. For a more extensive overview of this topic, we refer the reader to (O'Donnell, 2014, Chapter 8).

Given a measure $\mu$, we let $L^2(\mu)$ denote the space of $L^2$-integrable functions with respect to $\mu$.

**Fact 18 (Biased Fourier basis)** *Let $D$ be a distribution over $\{-1, 1\}$ with mean $\mu \in (-1, 1)$. Given $f \in L_2(D^{\otimes n})$, the $\mu$-biased Fourier expansion of $f : \{-1, 1\}^n \to \mathbb{R}$ is*

$$f(x) = \sum_{S \subseteq [n]} \widehat{f}(S) \phi_S(x) \, ,$$

*where $\phi_S(x) = \prod_{i \in S} \frac{x_i - \mu}{\sqrt{1-\mu^2}}$ and $\widehat{f}(S) = \mathbb{E}_{x \sim D^{\otimes n}}[\phi_S(x) f(x)]$.*

The functions $\phi_S$ provide an orthonormal basis for the space of functions $L^2(D^{\otimes n})$, where $D$ is the distribution over $\{1, -1\}$ with mean $\mu$. We can naturally extend this to arbitrary product distributions over $\mathbb{R}^d$ as follows:

**Fact 19 (Basis for an arbitrary product distribution)** *Let $D$ be a distribution over $\mathbb{R}$ with mean $\mu$ and variance $\sigma^2 > 0$. Then $\{1, \frac{x-\mu}{\sigma}\}$ is an orthonormal basis for $L^2(D)$, and thus $\{1, \frac{x-\mu}{\sigma}\}^{\otimes n}$ is an orthonormal basis for $L^2(D^{\otimes n})$.*

The orthonormality of the basis immediately implies the following simple fact:

**Fact 20 (Parseval's Theorem)** *For any function $f$ expressed as $f(x) = \sum_{S \subseteq [n]} \widehat{f}(S) \phi_S(x)$, we have $\mathbb{E}_{x \sim D^{\otimes n}}[f(x)^2] = \sum_{S \subseteq [n]} \widehat{f}(S)^2$.*

The following is the crucial structural result in the classical setting that gives rise to Theorem 3. Roughly speaking, it ensures that for the "concentrated product distributions" $D$ considered therein, any bounded multilinear function has decaying coefficients when expanded in the orthonormal basis for $L^2(D^{\otimes n})$.

**Lemma 21** *Let $f : [-1,1]^n \to [-1,1]$ be a multilinear function and let $D$ be a distribution over $\mathbb{R}$ with mean $\mu$ and variance $\sigma^2 < 1 - \mu^2$. Then there exists a function $f^{\leq d}$ such that*

$$\mathbb{E}_{x \sim D^{\otimes n}}[(f(x) - f^{\leq d}(x))^2] \leq \left(\frac{\sigma^2}{1-\mu^2}\right)^d .$$

**Proof** Let $f$ be expressed in the basis $B_{\text{hypercube}} = \{1, \frac{x-\mu}{\sqrt{1-\mu^2}}\}^{\otimes n}$; i.e. as

$$f(x) = \sum_{S \subseteq [n]} \widehat{f}(S) \psi_S(x)$$

where $\psi_S = \prod_{i \in S} \frac{x_i - \mu}{\sqrt{1-\mu^2}}$. Note that $\{\psi_S\}_{S \subseteq [n]}$ is orthonormal with respect to the distribution $\widetilde{D}^{\otimes n}$ where $\widetilde{D}$ is supported on $\{\pm 1\}$ with mean $\mu$. Since $|f(x)| \leq 1$ for $x \in [-1,1]^n$, it follows that

$$1 \geq \mathbb{E}_{x \sim \widetilde{D}^{\otimes n}}[f(x)^2] = \sum_{S \subseteq [n]} \widehat{f}(S)^2$$

via Fact 20.

Now, consider the basis $\phi_S := \prod_{i \in S} \frac{x_i - \mu}{\sigma} = (\frac{\sqrt{1-\mu^2}}{\sigma})^{|S|} \cdot \psi_S$. By Fact 19, we know that $\{\phi_S\}_{S \subseteq [n]}$ is orthonormal with respect to $D$. Let $f^{>d} := \sum_{|S|>d} \widehat{f}(S) \psi_S$. We have

$$\begin{aligned}
\mathbb{E}_{x \sim D^{\otimes n}}[f^{>d}(x)^2] &= \mathbb{E}_{x \sim D^{\otimes n}}\left[\Big(\sum_{|S|>d} \widehat{f}(S)\psi_S\Big)^2\right] \\
&= \mathbb{E}_{x \sim D^{\otimes n}}\left[\Big(\sum_{|S|>d} \widehat{f}(S)\Big(\frac{\sigma}{\sqrt{1-\mu^2}}\Big)^{|S|}\phi_S\Big)^2\right] \\
&= \sum_{|S|>d} \widehat{f}(S)^2 \Big(\frac{\sigma^2}{1-\mu^2}\Big)^{|S|} \\
&\leq \Big(\frac{\sigma^2}{1-\mu^2}\Big)^d \sum_{|S|>d} \widehat{f}(S)^2 ,
\end{aligned}$$

again using Fact 20. Since $\sum_S \widehat{f}(S)^2 \leq 1$, we have

$$\mathbb{E}_{x \sim D^{\otimes n}}[(f(x) - f^{\leq d}(x))^2] \leq \Big(\frac{\sigma^2}{1-\mu^2}\Big)^d$$

as desired. ■

### B.2. Sample complexity and error analysis

In light of Lemma 21, the proof of Theorem 3 is straightforward given the following elementary fact:

**Fact 22 (Bernoulli maximizes variance)** *Let $D$ be a distribution over the interval $[-(1-\eta), 1-\eta]$ with mean $\mu$. Then $\mathrm{Var}_{x\sim D}(x) \le (1-\eta)^2(1-\mu^2)$.*

We can now conclude the proof of Theorem 3:

**Proof of Theorem 3** By Fact 22, we have $\mathrm{Var}_D(x) \le (1-\eta)^2(1-\mu^2)$. Then Lemma 21 gives us a degree-$d$ approximation $f^{\le d}$ to $f$ such that

$$\mathbb{E}_{x\sim D^{\otimes n}}[(f(x) - f^{\le d}(x))^2] \le (1-\eta)^{2d}.$$

Taking $d \coloneqq \log(1/\varepsilon)/\log(1/(1-\eta))$ gives an approximation with error $\le \varepsilon$. Then by Corollary 17, linear regression on the space of polynomials of degree $\log(1/\varepsilon)/\log(1/(1-\eta))$ finds an $O(\varepsilon)$-error hypothesis in $n^{O(\log(1/\varepsilon)/\log(1/(1-\eta)))} \cdot \log(1/\delta)$ time and samples. ■

**Remark 23** *In the classical setting linear regression is not required, as we can estimate the mean, and the coefficients satisfy*

$$\widehat{f}(S) = \mathbb{E}_{x\sim D^{\otimes n}}[f(x)\phi_S(x)]\,,$$

*so they can be estimated directly. We give the guarantee in terms of linear regression because the approach of directly estimating the coefficients does not generalize to the quantum setting.*

## Appendix C. Lower bounds

In this section, we show lower bounds in the classical case, which automatically imply hardness in the quantum case. In Section C.1, we prove Theorem 5, which shows hardness of learning without the product distribution assumption in Theorem 3. In Section C.2, we show that truncating in the unbiased basis fails when the distribution is not mean zero.

### C.1. Lower bounds for learning non-product distributions

In this section, we prove Theorem 5. We show that if $\mathcal{D}$ is an arbitrary distribution, then even if $\mathcal{D}$ is supported on $[-(1-\eta), (1-\eta)]^n$ (i.e., in the interior of the hypercube), there is no learning algorithm.

Let $C$ be a code over $\{0,1\}^n$ of size $2^{\Theta(n)}$ and distance $n/4$. The following fact is standard.

**Fact 24** *For any constant $\varepsilon > 0$, any learning algorithm that can learn an arbitrary function $f : \{\pm 1\}^n \to \{\pm 1\}$ over $C$ to $\varepsilon$ error requires $2^{\Omega(n)}$ queries.*

**Proof of Theorem 5** Let $\eta = 0.1$. We set the distribution $\mathcal{D}$ to be the uniform distribution over $(1-\eta)\cdot C$, which is supported in $[-(1-\eta), (1-\eta)]^n$. Let $f : \{\pm 1\}^n \to \{\pm 1\}$ be an arbitrary function. Since the code $C$ has distance $n/4$, we can without loss of generality assume that $f(x') = f(x)$ whenever $d(x, x') \le n/8$.

Suppose for contradiction that there is an algorithm that, with only $2^{o(n)}$ queries to $f$, outputs a function $g : [-1,1]^n \to \mathbb{R}$ such that $\mathbb{E}_{x\sim\mathcal{D}}[(g(x) - f(x))^2] \le \varepsilon$. Let $p \coloneqq 1 - \eta/2$, and let $\mathrm{Ber}(p)$ be the distribution where we have $+1$ with probability $p$ and $-1$ otherwise. Then, for any $x \in \{\pm 1\}^n$, we have $f((1-\eta)x) = \mathbb{E}_{z\sim\mathrm{Ber}(p)^{\otimes n}}[f(x \circ z)]$, where $\circ$ denotes entry-wise product.

We claim that for any $x \in C$, $|\mathbb{E}_{z\sim\mathrm{Ber}(p)^{\otimes n}}[f(x \circ z)] - f(x)| \leq o_n(1)$. The number of $-1$ coordinates in $z$ is distributed as a $\mathrm{Bin}(n, \eta/2)$. By the Chernoff bound, $\Pr[\mathrm{Bin}(n, \eta/2) \geq (1+\delta)\frac{1}{2}\eta n] \leq e^{-O(\delta^2\eta n)} = o_n(1)$. In particular, for $\eta = 0.1$, we have that $\Pr[\mathrm{Bin}(n, \eta/2) \geq n/8] \leq o_n(1)$. Thus, with probability $1 - o_n(1)$, $d(x \circ z, x) < n/8$, which means that $f(x \circ z) = f(x)$. This proves that $|f((1-\eta)x) - f(x)| \leq o_n(1)$ for all $x \in C$.

Then, let $h(x) = g((1-\eta)x)$.

$$\begin{aligned}\mathbb{E}_{x\in C}\left[(h(x) - f(x))^2\right] &= \mathbb{E}_{x\in C}\left[(g((1-\eta)x) - f(x))^2\right] \\ &\leq \mathbb{E}_{x\in C}\left[(g((1-\eta)x) - f((1-\eta)x))^2\right] + o_n(1) \\ &= \mathbb{E}_{x\sim\mathcal{D}}\left[(g(x) - f(x))^2\right] + o_n(1) \\ &\leq \varepsilon + o_n(1)\,.\end{aligned}$$

This means that $h$ is an $\varepsilon$-approximation of $f$ over $C$. This contradicts Fact 24 thus completing the proof. ■

### C.2. Lower bounds for unbiased degree truncation

In Theorem 3, if the product distribution $\mathcal{D}$ over $[-(1-\eta), 1-\eta]^n$ has mean zero, then directly truncating $f$ with respect to the standard monomial basis at degree $O(\log(1/\varepsilon)/\log(1/(1-\eta)))$ (independent of $n$) suffices. However, in this section, we will show that without the mean zero assumption, even truncating at $\Omega(n)$ degree w.r.t. the monomial basis does not give a small approximation error. Our counter-example implies that truncation in any distribution-oblivious basis will fail on some product distribution. In the quantum setting, this implies that low-degree truncation in the standard Pauli basis fails on some product distribution as well.

The counter-example is quite simple: $f$ is the multilinear extension of the Boolean majority function, and $\mathcal{D}$ is supported on a single nonzero point (which is in fact a product distribution).

**Fact 25 (O'Donnell (2014))** *Let $f$ be the multilinear extension of the Boolean majority function. Then the $\ell_2$ Fourier weight on terms of degree $k$ is $\Theta(k^{-3/2})$.*

The following is a well-known fact in approximation theory.

**Fact 26 (Chebyshev extremal polynomial inequality Rivlin (1990))** *Let $p$ be a degree-$d$ univariate polynomial with leading coefficient 1. Then, $\max_{x\in[-1,1]} |p(x)| \geq 2^{-d+1}$.*

**Lemma 27** *Let $f : \{\pm1\}^n \to \{\pm1\}$ be the majority function extended to the domain $[-1,1]^n$. Let $0 < a < b < 1$ be fixed constants. Then, there exist $\delta := \delta(a,b) \in (0,1)$ and $t^* \in [a,b]$ such that the degree-$\delta n$ truncation $f^{\leq\delta n}$ has $|f^{\leq\delta n}(t^* \cdot \mathbf{1})| \geq \omega_n(1)$.*

**Proof**

Let $g(t) := f^{\leq d}(t \cdot \mathbf{1})$, which is a univariate polynomial of degree $d$, and consider the shifted polynomial $\widetilde{g}(t) = g(\frac{b-a}{2}t + \frac{a+b}{2})$. The leading coefficient of $\widetilde{g}$, denoted $\widetilde{c}_d$, is $c_d(\frac{b-a}{2})^d$, where

$c_d$ is the leading coefficient of $g$. Since $g(t) = f^{\leq d}(t \cdot \mathbf{1})$, we have $c_d = \sum_{S:|S|=d} \widehat{f}(S)$. For the majority function, Fact 25 implies that $\binom{n}{d}\widehat{f}(S)^2 = \Theta(d^{-3/2})$ for all $S$ of size $d$, thus

$$|\widetilde{c}_d| = \left(\frac{b-a}{2}\right)^d \binom{n}{d}^{1/2} \Theta(d^{-3/4})\,.$$

Then, by Fact 26, there must be a $s^* \in [-1, 1]$ such that

$$|\widetilde{g}(s^*)| \geq 2^{-d+1} \cdot \left(\frac{b-a}{2}\right)^d \binom{n}{d}^{1/2} \Theta(d^{-3/4})\,.$$

If $d = \delta n$, then $\binom{n}{d} \geq (\frac{1}{\delta})^d = e^{d\log(1/\delta)}$. Thus, given $0 < a < b < 1$, there exists a $\delta \in (0, 1)$ such that the above is $\exp(\Omega(d))$. Thus, there exists a $t^* \in [a, b]$ such that $|f^{\leq d}(t^* \cdot \mathbf{1})| \geq \omega_n(1)$. ■

# Appendix D. Extensions

## D.1. Learning the joint mapping

As mentioned in Remark 4, our techniques can be extended to learn not just the mapping $\rho \mapsto \mathrm{Tr}(O\mathcal{E}[\rho])$, but also the joint mapping $(O, \rho) \mapsto \mathrm{Tr}(O\mathcal{E}[\rho])$. As this is standard, here we only briefly sketch the main ideas.

The general strategy is to produce a classical description of the channel that we can then use to make predictions about properties of output states. To do this, we draw many input states $\rho_1, \ldots, \rho_N$ from $\mathcal{D}$, query the channel on each them, and for each of the output states $\mathcal{E}[\rho_j]$, we apply a randomized Pauli measurement to each of them and use these to form unbiased estimators for the output state. Concretely, given an output state $\mathcal{E}[\rho_j]$, a randomized Pauli measurement will result in a stabilizer state $|\psi^{(j)}\rangle = \otimes_{i=1}^n |s_i^{(j)}\rangle \in \{|0\rangle, |1\rangle, |+\rangle, |-\rangle, |y+\rangle, |y-\rangle\}^{\otimes n}$, and the expectation of $\otimes_{i=1}^n (3\,|s_i^{(j)}\rangle\langle s_i^{(j)}| - I)$ is $\mathcal{E}[\rho_j]$. The classical description of the channel is given by the $O(\log(1/\varepsilon))$-body reduced density matrices of the input states $\rho_1, \ldots, \rho_N$, together with the classical encodings of $|\psi^{(1)}\rangle, \ldots, |\psi^{(N)}\rangle$.

Given an observable $O$, we can then perform regression to predict the labels $\mathrm{Tr}(O \otimes_{i=1}^n (3\,|s_i^{(j)}\rangle\langle s_i^{(j)}| - I))$ given the features $\{\mathrm{Tr}(P\rho_j)\}_{|P| \leq O(\log(1/\varepsilon))}$. Because the labels are unbiased estimates of $\mathrm{Tr}(O\,\mathcal{E}[\rho_j])$, the resulting estimator will be an accurate approximation to $\rho \mapsto \mathrm{Tr}(O\,\mathcal{E}[\rho])$ for $\rho \sim \mathcal{D}$.

In this work, we do not belabor these details as they are already investigated in depth in Huang et al. (2023). Instead, we focus on the single observable case as this is where the main difficulty lies in extending the results of Huang et al. (2023) to more general input distributions.

## D.2. Non-iid product distributions

In the proof of Lemma 11, we relate the quantity $\mathbb{E}_{\rho \sim D^{\otimes n}}[\mathrm{Tr}(\mathcal{O}^{>d}\rho)^2]$ that we wish to bound to the related quantity $\mathbb{E}_{\rho \sim D^{\otimes n}}[\mathrm{Tr}(\mathcal{O}^2\rho)]$, which is at most 1 since $\|\mathcal{O}\|_{\mathsf{op}} \leq 1$. Due to the choice of our biased Pauli basis $B = \{I, X, Y, \frac{Z-\mu I}{\sqrt{1-\mu^2}}\}^{\otimes n}$, we may write both quantities as sums of $\widehat{\mathcal{O}}_S^\dagger M'^{\otimes |S|} \widehat{\mathcal{O}}_S$ and $\widehat{\mathcal{O}}_S^\dagger M^{\otimes |S|} \widehat{\mathcal{O}}_S$ (see Eq. (2) and (3)).

Suppose $\rho$ is a product of different distributions where each qubit has mean Bloch vector $\boldsymbol{\mu}_i = (0, 0, \mu_i)$ (after rotation; see Remark 13) and second moment bounded by $1 - \eta$. We will instead use the basis $B = \otimes_{i=1}^n \{I, X, Y, \frac{Z - \mu_i I}{\sqrt{1-\mu_i^2}}\}$. One can easily see that the above quantities are essentially the same, with $M^{\otimes |S|}$ replaced by $\otimes_{i \in S} M_i$ and similarly for $M'$. Then, the next steps in the proof (establishing Eq. (4) and (5)) are exactly the same.