Learning the closest product state

Ainesh Bakshi* John Bostanci† William Kretschmer♭ Zeph Landau♭

Jerry Li♮ Allen Liu* Ryan O'Donnell♯ Ewin Tang♭

*MIT, †Columbia, ♭UC Berkeley, ♮University of Washington, ♯Carnegie Mellon University

Abstract

We study the problem of finding a (pure) product state with optimal fidelity to an unknown n-qubit quantum state ρ, given copies of ρ. This is a basic instance of a fundamental question in quantum learning: is it possible to efficiently learn a simple approximation to an arbitrary state? We give an algorithm which finds a product state with fidelity ε-close to optimal, using $N = n^{\mathrm{poly}(1/\varepsilon)}$ copies of ρ and $\mathrm{poly}(N)$ classical overhead. We further show that estimating the optimal fidelity is NP-hard for error $\varepsilon = 1/\mathrm{poly}(n)$, showing that the error dependence cannot be significantly improved.

For our algorithm, we build a carefully-defined cover over candidate product states, qubit by qubit, and then demonstrate that extending the cover can be reduced to approximate constrained polynomial optimization. For our proof of hardness, we give a formal reduction from polynomial optimization to finding the closest product state. Together, these results demonstrate a fundamental connection between these two seemingly unrelated questions. Building on our general approach, we also develop more efficient algorithms in three simpler settings: when the optimal fidelity exceeds $5/6$; when we restrict ourselves to a discrete class of product states; and when we are allowed to output a matrix product state.

Contents

1 Introduction

When can we obtain a classical description of a complex quantum system? This problem, at the heart of quantum information theory, is one commonly faced by experimentalists: when we have a large, intricate quantum device, how can we tell what it is doing? Due to the exponentiality inherent to quantum mechanics, a generic system of n particles is described by a number of parameters scaling exponentially with n, so in general, an efficient description simply does not exist. However, real-world systems are not generic: the physics governing the device will suggest a corresponding model of the system, giving us a hint for how the system can be efficiently described.

Simultaneously, in real-world applications, the state which one is learning may not—and typically will not—exactly fall within a given model class, due to noise or other forms of imprecision in how our model represents the real world. In light of this, the natural question is to seek the best approximation to the underlying state within the prescribed model. Such an approximation can serve as a far more tractable proxy for the true state when it is complex to describe exactly. In this work, we consider the problem for the class of product states, arguably the most fundamental class of states to consider. Stated plainly, the question we ask is the following:

Can we efficiently learn the best product state approximation to any given state?

We formalize this problem as follows:

Problem 1 (Learning the closest product state)**.** Consider the set of n-qubit product states $\mathcal{P} = \{|\pi_1\rangle \otimes \cdots \otimes |\pi_n\rangle\}$, and let $\varepsilon, \delta > 0$ be error parameters. Given N copies of an arbitrary n-qubit state with density matrix ρ, output a classical description of a state $|\pi\rangle \in \mathcal{P}$ such that, with probability $\geqslant 1 - \delta$,

$$\langle\pi|\,\rho\,|\pi\rangle \geqslant \mathsf{OPT} - \varepsilon\,, \qquad \text{where } \mathsf{OPT} = \max_{|\pi\rangle\in\mathcal{P}} \langle\pi|\,\rho\,|\pi\rangle\ .$$

Product states are natural to study in this context for a number of reasons. Because of the locality inherent in physical systems, we commonly model physical systems with states exhibiting low entanglement. Chief among them are *mean-field theories*, which model systems as states which exhibit *zero entanglement*, e.g., product states [BH16]. The mean-field approximation plays a central role in domains relevant to quantum computing: in particular, in quantum chemistry, mean-field theories like Hartree-Fock theory and density functional theory are the standard algorithmic workhorses for understanding chemical processes [Cha24]. In light of this, we can rephrase Problem 1 as asking for the best (pure) mean-field approximation to an arbitrary quantum state, and for the quality of that approximation. From this perspective, we believe that obtaining an efficient algorithm for Problem 1 will have important implications both for validating the effectiveness of these theories and for understanding their properties in real-world settings.

As an example application, physicists already run computations to solve Problem 1 in the setting where the input is not a quantum state, but a description of a condensed matter system. Collective entanglement of a multipartite state is often measured by OPT, the best fidelity of the state with a product state, also known as the *geometric measure of entanglement* [WG03]. Since its introduction in 2003, this entanglement measure has been used to understand a variety of condensed matter systems (see related work). An algorithm for Problem 1 can be used to compute the geometric measure of entanglement for states which are efficiently preparable on

a quantum computer, by preparing copies of the state and running the algorithm to estimate OPT, giving an advantage when such states are classically intractable.

Despite the apparent simplicity of the problem, relatively little was known about the computational complexity of Problem 1. From a statistical point of view, one can obtain sample-efficient learners via classical shadow estimation [HKP20] or shadow tomography [Aar20], but these estimators require exponential runtime. On the other hand, efficient algorithms were known only for highly restricted versions of the problem [GIKL24]. This lack of efficient algorithms might be surprising, as when when the unknown state ρ is a product state, i.e. $\mathsf{OPT} = 1$, this task is easy: many algorithms work, including learning every register separately. However, these algorithms are brittle, and fail catastrophically when $\mathsf{OPT} < 0.99$. Even algorithms for the related problem of product state testing, initiated by the important work of Harrow and Montanaro [HM13], do not admit estimates of OPT when OPT is bounded away from 1. In contrast, one would hope to obtain efficient algorithms even when OPT is a small constant (say, 0.1): product states with constant fidelity are still great approximations, considering that almost all product states will have fidelity exponentially small in n.

Beyond specific applications, we hope that understanding this algorithmic task can shed light on a broader program in quantum learning theory. An emerging line of work has been studying "learning the closest state in a hypothesis class", also known as *agnostic tomography*: formally, this problem is Problem 1, except the class of product states $\mathcal{P}$ is replaced with a different hypothesis class $\mathcal{C}$. Product states appear as a special case of several well-studied classes of quantum states, including states described by low-depth quantum circuits, matrix product states, and Gibbs states and ground states of local Hamiltonians. Understanding the computational complexity of agnostic tomography of product states is therefore an important stepping stone to building up to richer approximations. As we demonstrate below, it turns out that learning the closest product state is already a surprisingly deep problem.

1.1 Results

We answer the aforementioned question in the affirmative and provide the first efficient algorithm for agnostic tomography of product states:

Theorem 1.1 (Learning the closest product state, Theorem 5.2)**.** *There is an algorithm which, given as input $\varepsilon > 0$ and $N = n^{\mathrm{poly}(1/\varepsilon)}$ copies of an unknown n-qubit[1] state ρ, runs in time $\mathrm{poly}(N)$ and outputs the classical description of a pure product state $|\phi\rangle$ that, with probability at least 0.99, satisfies*

$$\langle\phi|\rho|\phi\rangle \geqslant \mathsf{OPT} - \varepsilon\,. \tag{1}$$

The algorithm also produces an estimate of OPT to ε error.

We pause to make several comments about this result. First, the regime we are primarily interested in is when ε is a constant (though possibly small). In this regime, our algorithm runs in polynomial time. This resolves an open question posed in [GIKL24].

Secondly, our result holds for all values of OPT, and not just OPT close to 1. The setting where OPT is a small constant (say, 0.1) is particularly challenging: in this regime, there may not be a unique closest product state. In this setting, our algorithm in fact actually outputs a net (albeit in a relatively weak sense) over *all* product states which are close to the unknown state ρ; see Section 5 for more detailed discussion. Moreover, our algorithm does not need to

[1] For simplicity, we only consider when the local systems are qubits. We believe that the results should generalize to qudits without too much struggle: see Remark 3.5.

know the value of OPT, nor does it need even a lower bound on OPT (though if OPT is large the algorithm's complexity improves—see Remark 5.3). Note, however, that the guarantee on the fidelity of $|\phi\rangle$ with ρ is only nontrivial when $\mathsf{OPT} > \varepsilon$.

Finally, prior to this work, the only algorithms for this task were *sample*-efficient, but not time-efficient. For example, a polynomial number of random Clifford measurements suffices to estimate every fidelity with a product state $\langle\pi|\rho|\pi\rangle$ to ε error [HKP20]. However, there are an exponential number of these product states, and computing even one fidelity from these randomized measurements requires exponential time [JV14].

Improved product state testing. Agnostic tomography of product states is closely related to the well-studied problem of product state testing [HM13], where the goal is to determine whether or not a state $|\psi\rangle$ is a product state, or has fidelity at most $1-\varepsilon$ with any product state. In the former case, the test should always accept, and in the latter, the test should reject with probability at least p, for some $p = \Theta(\varepsilon)$.

Our results shed new light on this problem: the celebrated tester of Harrow and Montanaro [HM13] exhibits a strange behavior, wherein their rejection probability satisfies $p \leqslant 1/2 + o(1)$, even when $\varepsilon \to 1$. That is, for some reason, the tester cannot distinguish the case where $|\psi\rangle$ has overlap roughly $1/2$ with some product state, versus the case where the state has overlap $\ll 1/2$ with any product state. Since our algorithm also produces an estimate of OPT to error ε, it improves upon the best-known guarantees for product state testing [SW22] in this "tolerant" [Can20] regime.

Computational lower bounds. It is natural to ask whether or not one can hope for a running time for this problem which is polynomial in both n and $1/\varepsilon$. We complement our upper bound with a lower bound, demonstrating that our runtime is, in a qualitative sense, close to optimal:

Theorem 1.2 (Hardness of product state approximation, Theorem 7.3)**.** *Suppose there is an efficient quantum algorithm for solving the following problem: given* $\mathrm{poly}(n)$ *copies of an unknown, n-qubit mixed state* ρ*, with probability* $\geqslant 0.01$*, output* $|\psi\rangle$ *satisfying*

$$\langle\psi|\rho|\psi\rangle \geqslant \max_{|\pi\rangle\in\mathcal{P}} \langle\pi|\rho|\pi\rangle - \frac{1}{\mathrm{poly}(n)}\,.$$

Then $\mathsf{BQP} \supseteq \mathsf{NP}$.

In particular, this rules out algorithms with strongly polynomial dependencies on all parameters. We prove this hardness via a straightforward, polynomial-time reduction to an NP-complete problem. Consequently, this also rules out any algorithms that have sub-exponential dependence on $1/\varepsilon$, assuming the quantum analog of the exponential time hypothesis. We interpret this as saying that it is likely challenging to obtain substantial qualitative improvements to the runtime in Theorem 1.1.

We also remark that this hardness result demonstrates an interesting computational-statistical gap for the problem of finding the closest product state. Namely, classical shadow estimation [HKP20] demonstrates that this regime can be solved sample efficiently, but on the other hand, our lower bound demonstrates that this rate cannot be matched by any efficient algorithm.

Approximate tensor optimization. The upper and lower bound are based on a new connection to the classical problem of *approximate tensor optimization*. Here, one is given a d-tensor $T \in$

$(\mathbb{C}^n)^{\otimes d}$, and the goal is to find a unit vector $\vec{x} \in \mathbb{C}^n$ satisfying

$$T(\vec{x}, \ldots, \vec{x}) \geqslant \max_{\|u\|_2=1} T(\vec{u}, \ldots, \vec{u}) - \varepsilon \|T\|_F .$$

Our lower bound proceeds by direct reduction to this problem for $d = 4$, which is known to be NP-hard when $\varepsilon = 1/\operatorname{poly}(n)$ [FL17], and our upper bound works by reducing the problem to many different instances of constrained versions of this problem. This problem itself bears great resemblance to the problem of solving dense CSPs, and indeed, we believe the techniques we develop for constrained tensor optimization here may have applications to that setting as well.

Faster agnostic tomography of product states. In light of our lower bound, we ask whether there are simpler algorithms for agnostic tomography of product states, perhaps under additional assumptions. We show that this is true for three natural settings: (1) when the best product state approximation is quite good; (2) when the number of choices for each qubit is discrete; and (3) when the output is allowed to be a matrix product state.

First, we obtain a linear copy and nearly-quadratic time algorithm for agnostic tomography of product states as long as the fidelity of the optimal solution exceeds a fixed constant (namely, $5/6$):

Theorem 1.3 (High-fidelity learning, Theorem 4.20)**.** *There is an algorithm that takes as input a parameter $\varepsilon > 0$ as well as $N = O(n/\varepsilon)$ copies of an n-qubit state ρ, and has the following guarantees: Provided $\mathsf{OPT} > 5/6 + \varepsilon$, it runs in $O(Nn \log n)$ time and outputs a pure product state $|\psi\rangle$ that satisfies*

$$\langle \psi | \rho | \psi \rangle \geqslant \mathsf{OPT} - \varepsilon ,$$

(except with probability at most .01).

In other words, so long as the quality of the product approximation OPT exceeds $5/6$, there is a strongly polynomial time algorithm for agnostic product state tomography. This stands in stark contrast to the state of affairs for general OPT, where the hardness result demonstrates such an algorithm is impossible. The threshold $5/6$ naturally arises from our analysis, but it is an interesting open question to what extent it can be pushed.

We remark that the runtime dependence of the algorithm is linear in $1/\varepsilon$, even though it is easily seen that *estimating* OPT to $\pm\varepsilon$ requires $\Omega(1/\varepsilon^2)$ samples. For example, this lower bound holds even in the special case when $\rho = \mathsf{OPT}\,|0\rangle\langle 0| + (1 - \mathsf{OPT})\,|1\rangle\langle 1|$ is a biased coin, and we want to distinguish whether (say) $\mathsf{OPT} = 0.9 + \varepsilon$ or $\mathsf{OPT} = 0.9 - \varepsilon$. Our algorithm demonstrates that the task of *finding* a state whose fidelity is within ε of the optimum may be easier.

Second, we give an efficient algorithm for agnostic tomography, when the class of states is the set of product states where each qubit is drawn from a finite set of possible states:

Theorem 1.4 (Learning of a finite class of product states, Theorem A.1)**.** *For $k = 1, \ldots, n$, let $\mathcal{A}_k$ denote a set of single qudit pure states satisfying $|\mathcal{A}_k| \leqslant s$ and $|\langle \phi | \phi' \rangle| \leqslant 1 - \delta$ for all distinct $|\phi\rangle, |\phi'\rangle \in \mathcal{A}_k$. Let $\mathcal{A} = \mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_n$, and for any n-qudit quantum state ρ, let $\mathsf{OPT}_{\mathcal{A}} = \mathsf{OPT}_{\mathcal{A}}(\rho) = \max_{|\pi\rangle \in \mathcal{A}} \langle \pi | \rho | \pi \rangle$. Then there is an algorithm which, given as input $\varepsilon > 0$ and $N = \operatorname{poly}((ns)^{\log(1/\varepsilon)/\delta})$ copies of an n-qudit state ρ, runs in $\operatorname{poly}(N)$ time and outputs the classical description of some $|\psi\rangle \in \mathcal{A}$ satisfying*

$$\langle \psi | \rho | \psi \rangle \geqslant \mathsf{OPT}_{\mathcal{A}} - \varepsilon ,$$

(except with probability at most .01).

Stated plainly, so long as there are a finite set of possible states, and these states are all pairwise separated, then there is an efficient algorithm for agnostic tomography for this class of product states. We note that, similar to Theorem 1.1, our algorithm actually outputs all good solutions. This result also directly generalizes prior work of Grewal, Iyer, Kretschmer, and Liang [GIKL24], which studied the special case where each $\mathcal{A}_k$ is the set of 1-qubit stabilizer states. A very similar result was also obtained independently in [CGYZ24], albeit with quite different techniques.

Third, we give an algorithm for learning a good matrix-product state approximation of a given state ρ. Matrix product states with small bond dimension can be used to efficiently describe systems of multiple particles where particles share a small (but non-zero) amount of entanglement, and are ubiquitous in quantum many-body physics [PVWC07; Sch11]. We give an algorithm for *agnostic* tomography of matrix product states.

Theorem 1.5 (Agnostic (improper) learning of matrix product states, Theorem B.2)**.** *Let n, d, r be positive integers, and let $\mathrm{MPS}_{n,d,r}$ be the class of matrix product states on n qudits of local dimension d with bond dimension r (Definition B.1). For any state $\rho \in (\mathbb{C}^{d\times d})^{\otimes n}$, let $\mathsf{OPT}_r = \mathsf{OPT}_{n,d,r}(\rho) = \max_{|\phi\rangle\in\mathrm{MPS}_{n,d,r}} \langle\phi|\rho|\phi\rangle$ be the maximum fidelity any such MPS has with ρ. There is an algorithm which, given as input $\varepsilon > 0$ and $N = \mathrm{poly}(n, d, r, 1/\varepsilon)$ copies of an unknown n-qudit state ρ, runs in time $\mathrm{poly}(N)$ and outputs the classical description of a matrix product state $|\widehat{\phi}\rangle$ of bond dimension $dn^2 \cdot \mathrm{poly}(r, 1/\varepsilon)$ such that*

$$\langle\widehat{\phi}|\rho|\widehat{\phi}\rangle \geqslant \mathsf{OPT}_r - \varepsilon\,,$$

(except with probability at most .01).

We can relate this task back to learning the closest product state by taking $r = 1$ and $d = 2$; then, $\mathrm{MPS}_{n,d,r}$ is the class of product states over qubits, and our algorithm is able to output a matrix product state with bond dimension $n^2\,\mathrm{poly}(1/\varepsilon)$ whose fidelity with ρ is at least $\mathsf{OPT} - \varepsilon$. This gives an improper learner for product states, "improper" referring to our output not being a product state but instead a low-entanglement state. Our main result Theorem 1.1 is a *proper* learner for product states. In the error regimes of our lower bound, this gives an instance where improper agnostic learning is efficient, but proper agnostic learning is NP-hard. In general, the output of this algorithm is an MPS with a bond dimension of at least rn^2, which achieves a fidelity which is optimal with respect to MPSs with bond dimension r; this dependence on n in particular seems to be what makes this result more straightforward than proper learning of MPSs.

For this task, we recognize that the algorithm of Cramer, Plenio, Flammia, Somma, Gross, Bartlett, Landon-Cardinal, Poulin, and Liu [Cra+10] to learn an MPS also works when the input state is not an MPS, but merely has large constant fidelity with an MPS; our contribution is to generalize it to the agnostic case and perform the necessary analysis. We give a more detailed discussion in Appendix B.

Techniques. All of these results, as well as Theorem 1.1, are all based on a common algorithmic framework, which may have applications more broadly. At a very high level, our algorithms sweep through the qubits one at a time, and generate a set of candidates for good solutions on the qubits seen so far. This cover is then used as the starting point for generating candidates over the subsequent qubits. The main algorithmic challenge is in making extending the cover efficient. In the case of Theorem 1.1, extending this cover is intimately connected to tensor optimization, as mentioned above. To achieve our faster algorithms Theorems 1.3 to 1.5, our

key insight is that there are relatively simple and "greedy" techniques that allow us to extend this cover.

Our algorithms interleave classical computation with a particular quantum subroutine: the only way we access ρ is to perform tomography on various subspaces of subsystems, e.g. to estimate $\Pi \operatorname{tr}_S(\rho)\Pi$ for S some subset of qubits and Π a projector onto a subspace of $\operatorname{poly}(n)$ dimension (which can be represented efficiently with a quantum circuit). Our algorithms apply this subroutine to various choices of Π and S, which are adaptively chosen after classical computation on the output of the previous tomography routines. Since such a tomography subroutine can be performed with single-copy measurements, our algorithm can also be performed with only single-copy measurements. However, the adaptivity of this algorithm appears inherent: the classical shadows formalism [HKP20] is the standard technique to allow algorithms like these to perform all of their measurements up-front, but doing this comes at the cost of exponential running time, which we cannot tolerate.

1.2 Related work

Concurrent work. In independent and concurrent work, Chen, Gong, Ye, and Zhang [CGYZ24] give an algorithm for agnostic tomography of a finite set of product states, attaining a near-identical result to Theorem 1.4 via completely unrelated techniques. They also give an improved algorithm when the product states are stabilizer; we are also able to get a similar improvement in this setting(see Remark A.5 for more details).

Agnostic tomography. The notion of agnostic tomography was introduced by Grewal, Iyer, Kretschmer, and Liang [GIKL24], though similar notions have been considered under the notion "quantum hypothesis selection" [BO24] and in the PAC-learning setting; we refer to the survey [AA23] for a thorough discussion. Recent work has given agnostic tomography algorithms for stabilizer product states [GIKL24] and stabilizer states [CGYZ24]. These algorithms use unrelated techniques.

Product state testing. A notable related algorithm is the product state test, which, using copies of a state ρ is able to distinguish the cases that $\mathsf{OPT} = 1$ from $\mathsf{OPT} \leqslant 1 - \varepsilon$. This algorithm, introduced by [MKB05] and analyzed by Harrow and Montanaro [HM13], plays an important role in complexity theory, being used to prove that $\mathsf{QMA}(2) = \mathsf{QMA}(k)$ for $k > 2$. Though the algorithm suffices for testing, it cannot be used to estimate OPT when OPT is bounded away from 1 [HM13, Appendix B]. For similar reasons, it also does not seem to help with the task of finding good product states.

Product state approximations in other contexts. Our algorithm shows that it is possible to estimate the "geometric measure of entanglement" of a given pure state in polynomial time. This measure of entanglement, defined by Wei and Goldbart [WG03; VPRK97], has seen significant investigation as a measure of multipartite entanglement. This interest comes from this measure's potential to capture aspects of entanglement in condensed matter systems which cannot be captured by the typical, bipartite measures of entanglement [WDMVG05; ODV08; OW10]. See the survey of De Chiara and Sanpera for further discussion [DS18]. However, research has been limited by computational intractability, so our work may give a possible avenue to expand its scope via quantum simulation.

Mean-field approximations also arise naturally in contexts where we want to understand things like ground states of many-body systems, and only have a handle on product states [BH16].

For example, to prepare ground states of many-body systems, current heuristic phase estimation methods have a running time which depends on the fidelity between the ground state and an input product state [Lee+23].

Agnostic learning of product distributions. In some ways, the problem we consider here is the quantum analog of the well-studied problem of agnostic learning of product distributions on the hypercube. In its most basic form, we are given samples from a distribution that is close to a product distribution over the hypercube, and the goal is to learn the optimal product distribution approximation. Efficient algorithms for this problem were given in [DKKLMS19; LRV16]. However, these algorithms only work when their version of OPT is sufficiently large; in classical learning theory, the regime when OPT is small is known as *list learning*, and efficient algorithms for list learning of product distributions are also known; see, e.g. [CSV17; KSS18]. However, the guarantees they obtain are quite incomparable to ours, and their techniques do not have a meaningful parallel in the quantum setting.

Polynomial optimization. Polynomial optimization over the sphere is hard in general. Multiplicative approximations for optimizing low-degree polynomials in the worst case are well-understood (see [BGGLT17] and references therein). However, polynomial optimization has still found prominent applications in classical learning problems in the last decade. The polynomials that naturally appear in these settings do not tend to be worst-case, and admit significantly better approximations. Optimizing low-degree polynomials (often subject to polynomial constraints) has become a key algorithmic primitive in dictionary learning [BKS15], tensor decomposition [HSS15], robustly learning Gaussian mixture models [MV10; BS15; LM21; BDJKKV22] and private and list-decodable learning [HKMN23; KKK19; RY20; BK21]. These techniques have also found applications in quantum tasks, such as best separable state [BKS17] and learning quantum Hamiltonians [BLMT24; Nar24]. An interesting feature of our algorithm, compared to this other work, is that we do not establish uniqueness of some strong structure arising from the underlying parameters. Instead, we output a (non-unique) cover over solutions, and use polynomial optimization as a subroutine to produce such a cover.

Optimizing low-degree polynomials over the hypercube also leads to approximation algorithms for constraint satisfaction problems on dense and low-threshold rank graphs [BRS11; RT12; MR17] and high-dimensional expanders [AGT19]. These results roughly proceed via solving a sum-of-squares relaxation of a polynomial maximization problem, and obtain additive error that scales proportional to ε times the ℓ_2-norm of the coefficients of the polynomial and runs in $n^{\mathrm{poly}(1/\varepsilon)}$ time. Similar techniques have also appeared in the context of refuting random CSP's [RRS17].

A closely related problem is that of optimizing random polynomials over the sphere, which has deep connections to statistical physics and admits an additive-error guarantee under full replica-symmetry breaking [Sub20]. While our optimization problem does not involve random polynomials, we show that we can optimize low-degree polynomials up to small additive-error efficiently.

2 Technical overview

We now cover the key technical ideas of our algorithms. The precise version of the main algorithm can be found in Algorithm 5.5, which describes its outer loop, and Algorithm 5.10,

which describes the main subroutine. Further explanation of the subroutine can be found in Section 5.2.

Why naive approaches fail. Given an n-qubit quantum state with density matrix $\rho \in \mathbb{C}^{2^n \times 2^n}$, we want to find the product state that maximizes fidelity with ρ. The obvious algorithm that one might try to learn the closest product state is to take the best pure state approximation to each of its single-qubit subsystems. This algorithm works if ρ itself is a pure product state. However, the single-qubit subsystems do not contain enough information to deduce the best product state, even when the fidelity of ρ with the best product state is very close to 1. This phenomenon is why many naive approaches give exponentially poor approximations to the optimal value.

An illustrative example is to consider $\rho = |\psi\rangle\langle\psi|$ where $|\psi\rangle$ is the state proportional to

$$|\psi\rangle \propto \sqrt{1-\varepsilon}\,|0^n\rangle + \sqrt{\varepsilon}\,|+^n\rangle$$

for some small constant ε. Because $\langle +^n|0^n\rangle = 2^{-n/2}$, $|\psi\rangle$ as written is exponentially close to normalized. The fidelity with the product state $|0^n\rangle$ can be computed explicitly:[2]

$$\langle 0^n|\rho|0^n\rangle = |\langle\psi|0^n\rangle|^2 = \left(\frac{\sqrt{1-\varepsilon} + \sqrt{\varepsilon/2^n}}{\|\sqrt{1+\varepsilon}\,|0^n\rangle + \sqrt{\varepsilon}\,|+^n\rangle\|_2}\right)^2 \geqslant 1-\varepsilon\,.$$

In the limit of large n, the one-qubit density matrices of $|\psi\rangle$ all approach

$$\rho_i = \begin{bmatrix} 1-\varepsilon/2 & \varepsilon/2 \\ \varepsilon/2 & \varepsilon/2 \end{bmatrix}$$

We will see that there is a distinct state $|\psi'\rangle$ that is also ε-close to a product state, and has identical reduced density matrices, but for which $|0^n\rangle$ is a very bad product state approximation. Take an eigendecomposition of ρ_i as

$$\rho_i = p_1\,|v_1\rangle\langle v_1| + p_2\,|v_2\rangle\langle v_2|\,,$$

with $p_1 > p_2$. The state

$$|\psi'\rangle = \sqrt{p_1}\,|v_1\rangle^{\otimes n} + \sqrt{p_2}\,|v_2\rangle^{\otimes n}$$

also has at least $1-\varepsilon$ fidelity with a product state (namely $|v_1\rangle^{\otimes n}$), and has all its local density matrices equal to ρ_i. However, calculation shows that $|\langle\psi'|0^n\rangle|^2$ decays exponentially to 0 in the limit of large n, because both $|v_1\rangle$ and $|v_2\rangle$ are constant-far from $|0\rangle$. So, there is not enough information in the one-qubit reduced density matrices to learn the best product state approximation.

Barriers to agnostic product tomography. The hard case above illuminates broader challenges inherent to this task. We are concerned with optimizing the fidelity $\langle\pi|\rho|\pi\rangle$ over the class of product states; however, fidelity is typically quite poorly behaved. For example, almost all product states have exponentially small fidelity with ρ, which is too small to detect, and the fidelity landscape can have many local optima which thwart local search algorithms, like those based on convex optimization. This ill-behavedness is a well-established phenomenon related to the "barren plateau" problem in quantum machine learning [MBSBN18].

[2] It follows from one of our later results (Theorem 4.4) that the maximum product state fidelity with $|\psi\rangle$ is exponentially close to $1-\varepsilon$.

The regime where the optimal fidelity is a small constant like 0.1 is particularly challenging since, unlike the case where OPT is near 1, there are many well-separated globally optimal solutions. This lack of uniqueness presents basic issues for us: even if we manage to traverse the fidelity landscape and find many locally-optimal product states with fidelity 0.1, how can we conclude that we are done, and certify that there is no product state with fidelity 0.2?

Maintaining a cover over good product states. Our crucial insight is that we can efficiently maintain a cover over all product states that have large fidelity. This insight is enabled by the following observations:

1. If a product state $|\pi\rangle$ has good fidelity with ρ, then its restriction to a subsystem S has good fidelity with the partial trace of ρ onto the subsystem: $\langle\pi|\rho|\pi\rangle \leqslant \langle\pi_S|\rho_S|\pi_S\rangle$.
2. The number of product states with good fidelity with ρ and which have pairwise small fidelity with each other is small.

The first observation means that we do not have to optimize fidelity over the entire space of product states: just those which are extensions of good product states over a subsystem. In short, we can build a set of good product states qubit by qubit. The second observation means that, instead of maintaining *all* good product states, of which there could be exponentially many, it suffices to maintain a small number of well-separated good product states. In short, it suffices to maintain a cover.

A priori, it is even unclear whether a small cover over such product states exists. Our main technical contribution is to establish the existence of such a cover and demonstrate that it can be computed efficiently. Our algorithm starts with a cover over good product states for $\rho_{[1]}$, the state on qubit 1, and iteratively expands the cover a single qubit at a time. In particular, we show that given a cover for qubits $1, 2, \ldots, m-1$, extending it to qubit m can be reduced to polynomial optimization problems over the sphere with a dimension-independent number of ℓ_2 and ℓ_∞ constraints.

For the remainder of the section, we outline our approach to efficiently maintain a cover.

Fidelity and tangent distance. We begin by introducing a parametrization of product states which is used throughout the paper. For a n-dimensional vector of complex numbers, $\vec{z} \in \mathbb{C}^n$, we denote its corresponding product state by

$$|\pi_{\vec{z}}\rangle = \bigotimes_{i=1}^{n} \frac{|0\rangle + z_i |1\rangle}{\sqrt{1+|z_i|^2}}.$$

Looking ahead, we ultimately want to optimize over these z_i's, so we want a notion of cover which behaves nicely with respect to this parametrization. Fidelity is well-known to be an unwieldy notion of distance between quantum states and is typically hard to analyze. So, instead of constructing a cover directly using fidelity, we introduce an alternate measure between product states:

Definition 2.1 (Tangent distance, Definition 3.4)**.** Given two product states $|\pi_{\vec{z}}\rangle$ and $|\pi_{\vec{a}}\rangle$, the tangent distance between them is defined as

$$\mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}}\rangle) = \left\| \frac{\vec{z}-\vec{a}}{1+\vec{z}^*\vec{a}} \right\|_2 = \left(\sum_{i=1}^{n} \left| \frac{z_i - a_i}{1+z_i^* a_i} \right|^2 \right)^{1/2}.$$

We call it "tangent distance" because, for a single qubit, this measure corresponds to $|\tan(\theta)|$, where θ is the angle between the two states on the Bloch sphere. This notion of distance satisfies several desiderata, including being invariant under single-qubit unitaries and being equal to $\|\vec{z}\|_2$ when $\vec{a} = \vec{0}$ (see Section 3.1 for details). Importantly, tangent distance can be related to fidelity as follows (see Lemma 3.6 for a proof):

$$\log\left(\frac{1}{|\langle\pi_{\vec{z}}|\pi_{\vec{a}}\rangle|^2}\right) \leqslant \mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}}\rangle)^2 \leqslant \frac{1}{|\langle\pi_{\vec{z}}|\pi_{\vec{a}}\rangle|^2} - 1\,. \tag{2}$$

Now, we can introduce our notion of cover under tangent distance[3] (Definition 5.1): a cover $\mathcal{C}$ over product states is (η, ε)-good for a state ρ if the following hold

- **Good fidelity:** For all product states $|\pi\rangle \in \mathcal{C}$, $\langle\pi|\rho|\pi\rangle \geqslant \eta - \varepsilon$;
- **Separation:** For all distinct $|\pi\rangle, |\pi'\rangle \in \mathcal{C}$, $\mathrm{d}_{\tan}(|\pi\rangle, |\pi'\rangle) \geqslant 2/\eta$;
- **Coverage:** For any product state $|\phi\rangle$ such that $\langle\phi|\rho|\phi\rangle \geqslant \eta$, there exists a $|\pi\rangle \in \mathcal{C}$ such that $\mathrm{d}_{\tan}(|\phi\rangle, |\pi\rangle) \leqslant 3/\eta$.

We design an algorithm which, given η and $\varepsilon < \eta/3$, outputs a (η, ε)-good cover, where every product state $|\pi_{\vec{z}}\rangle$ in the cover is described by its parametrization $\vec{z}$. In particular, this gives a product state with fidelity $\geqslant \eta - \varepsilon$, assuming a product state with fidelity η exists. By performing binary search on η, one can use this to find a product state with fidelity $\geqslant \mathsf{OPT} - \varepsilon$, as stated in Theorem 1.1.

Existence of small covers. Our first step is to show that the size of an (η, ε)-good cover is at most $6/\eta$ (see Claim 5.4). Let $\mathcal{C} = \{\langle\pi^{(i)}\rangle\}_i$ be an (η, ε)-good cover. For intuition, suppose the product states in the cover were not just well-separated but orthogonal. Then each captures a different part of the "mass" of ρ. That is,

$$1 = \mathrm{tr}(\rho) \geqslant \sum_i \langle\pi^{(i)}|\rho|\pi^{(i)}\rangle \geqslant |\mathcal{C}|(\eta - \varepsilon) \geqslant |\mathcal{C}|(2\eta/3),$$

where the last two inequalities use the good fidelity property of the cover and that $\varepsilon < \eta/3$, respectively. In general, $\sum_i \langle\pi^{(i)}|\rho|\pi^{(i)}\rangle$ is equal to $\mathrm{tr}(MM^\dagger\rho)$ for M the matrix whose columns are the states in the cover $|\pi^{(i)}\rangle$. Then,

$$|\mathcal{C}|(2\eta/3) \leqslant \mathrm{tr}(MM^\dagger\rho) \leqslant \|MM^\dagger\|_{\mathrm{op}} = \|M^\dagger M\|_{\mathrm{op}} \leqslant 1 + |\mathcal{C}|(\eta/2),$$

giving the bound $|\mathcal{C}| \leqslant 6/\eta$. In the last step, we used the well-separated condition: the diagonal entries of $M^\dagger M$ are 1, the off-diagonal ones have magnitude at most $\eta/2$ by Eq. (2), and by the Gershgorin circle theorem the operator norm of $MM^\dagger$ is bounded by the maximum sum of magnitudes of any of its rows.

We can further show how to construct an (η, ε)-good cover algorithmically (Algorithm 5.5). We do this by iteratively forming an (η, ε)-good cover for $\rho_{[m]}$, the partial trace of ρ onto qubits 1 through m, for m from 1 to n. We can construct a good cover for $\rho_{[m]}$ greedily: start with $\mathcal{C}_m$ empty, and look for a violation of the coverage property. When we find one, add the corresponding $|\phi\rangle$ to $\mathcal{C}_m$, and repeat. Because we know an (η, ε)-good cover on $m-1$ qubits $\mathcal{C}_{m-1}$, we can restrict our search to just look over product states $|\phi\rangle$ whose first $m-1$ qubits are

[3] Though our algorithm naturally produces a cover with respect to tangent distance, one can use (2) to convert the guarantees to those involving fidelity.

close in tangent distance to an element of $\mathcal{C}_{m-1}$. This makes the problem of finding a violation tractable, since we only have to search in the neighborhood of some "root" product state. In particular, we show that it suffices to solve the following optimization problem (see Claim 5.7).

$$\begin{aligned} \underset{\vec{z}\in\mathbb{C}^m}{\text{maximize}} \quad & \langle \pi_{\vec{z}} | \rho | \pi_{\vec{z}} \rangle \\ \text{subject to} \quad & \mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}}\rangle) \geqslant 2/\eta \text{ for all } |\pi_{\vec{a}}\rangle \in \mathcal{C}_m, \\ & \mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{0}}\rangle) \leqslant 4/\eta. \end{aligned} \tag{Tangent Cover}$$

The precise soundness and completeness guarantees needed are shown in Algorithm 5.5; the constraints allow for significant slack. Note that the second constraint is equivalent to $\|\vec{z}\|_2 \leqslant 4/\eta$.

Constructing covers and polynomial optimization. Now, we consider the task of solving (Tangent Cover). Solving this even in the simplest case is not straightforward. An example to keep in mind is the following: suppose we are adding our first state to $\mathcal{C}_n$, which is currently empty. So, there are no "farness" constraints, the first kind of constraint in the program. Then, let $\rho = |\psi\rangle\langle\psi|$, where $|\psi\rangle$ is a superposition over computational basis strings with Hamming weight 0 and d:

$$|\psi\rangle = \sqrt{\gamma}\,|0^n\rangle + \sqrt{\frac{1-\gamma}{\binom{n}{d}}} \sum_{\substack{x\in\{0,1\}^n \\ |x|=d}} |x\rangle\,.$$

We are imagining, say, $\gamma = 0.9\eta$. Then $\langle 0^n|\rho|0^n\rangle = \gamma$, so our root state has good fidelity, but not quite enough to be a violation as we desire. (This can indeed happen; though $|0^n\rangle$ comes from the cover $\mathcal{C}_{n-1}$, so $|0^{n-1}\rangle$ has fidelity at least $\eta - \varepsilon$, it is extended by one qubit, which can drop the fidelity to γ or lower.) However, for $\vec{z} = \frac{1}{\sqrt{n}}\vec{1}$,

$$\begin{aligned} \langle \pi_{\vec{z}}|\rho|\pi_{\vec{z}}\rangle &= (1+1/n)^{-n}\left(\sqrt{\gamma} + \sqrt{\frac{\binom{n}{d}(1-\gamma)}{n^d}}\right)^2 \\ &\underset{n \text{ large}}{\approx} \frac{1}{e}\left(\sqrt{\gamma} + \sqrt{(1-\gamma)/d!}\right)^2, \end{aligned}$$

which can be larger than η even for $d = \Theta(\log(1/\eta)/\log\log(1/\eta))$. Note that $\|\vec{z}\|_2 = 1 \leqslant 4/\eta$, so it is close enough to the root in (Tangent Cover), and our algorithm must be able to recognize this better solution of $|\pi_{\vec{z}}\rangle$. This demands knowledge of ρ in a (quite large) Hamming ball around the root product state. Further, by changing the signs of the $|x\rangle$'s in $|\psi\rangle$, (Tangent Cover) can encode dense d-CSP instances. This suggests that the right approach is a reduction to polynomial optimization.

First, we consider solving (Tangent Cover) when $\mathcal{C}_m$ is empty. We can reduce this to low-degree polynomial optimization over the sphere. We start by observing that it suffices to consider the projection of ρ on basis states of low Hamming weight. Concretely, let $\Pi_{\geqslant d}$ be the projection onto the subspace of Hamming weight at least $d = O(1/\eta + \log(1/\varepsilon))$. Then, we show that, provided $\|\vec{z}\| \leqslant 4/\eta$ as in (Tangent Cover),

$$\|\Pi_{\geqslant d}\,|\pi_{\vec{z}}\rangle\| \leqslant \varepsilon\,.$$

This is a Chernoff bound, since the squared norm is the probability that the n Bernoulli random variables sums to at least d, where the probabilities come from the $\vec{z}$'s (see Lemma 3.10 with

$\mu = 4/\eta$). So, it suffices to perform state tomography for ρ on the space of low Hamming weight vectors $\rho_d = \Pi_{<d}\rho\Pi_{<d}$, which is computationally efficient because this subspace has dimension $O(n^d)$ (Lemma 5.9). We can use ρ_d in place of ρ in the objective because

$$\begin{aligned}
|\langle\pi_{\vec{z}}|\,\rho\,|\pi_{\vec{z}}\rangle - \langle\pi_{\vec{z}}|\,\rho_d\,|\pi_{\vec{z}}\rangle| &= \left|\mathrm{tr}\Big(\rho(|\pi_{\vec{z}}\rangle\langle\pi_{\vec{z}}| - \Pi_{<d}\,|\pi_{\vec{z}}\rangle\langle\pi_{\vec{z}}|\,\Pi_{<d})\Big)\right| \\
&\leqslant \||\pi_{\vec{z}}\rangle\langle\pi_{\vec{z}}| - \Pi_{<d}\,|\pi_{\vec{z}}\rangle\langle\pi_{\vec{z}}|\,\Pi_{<d}\|_{\mathrm{op}} \\
&\leqslant 2\||\pi_{\vec{z}}\rangle - \Pi_{<d}\,|\pi_{\vec{z}}\rangle\|_2 \\
&= 2\|\Pi_{\geqslant d}\,|\pi_{\vec{z}}\rangle\|_2 \leqslant \varepsilon\,.
\end{aligned}$$

Further, once we have our estimate of ρ_d, the objective function is fully specified explicitly; the rest of the algorithm is classical. Because ρ_d is only supported on low Hamming weight, $\langle\pi_{\vec{z}}|\,\rho_d\,|\pi_{\vec{z}}\rangle$ is a low-degree polynomial up to a normalization factor:

$$\langle\pi_{\vec{z}}|\,\rho_d\,|\pi_{\vec{z}}\rangle = \underbrace{\frac{1}{\prod_{i\in[m]}(1+|z_i|^2)}}_{(3).(1)} \underbrace{\sum_{x,x'\in\{0,1\}^m} \langle x|\,\rho_d\,|x'\rangle\,(\vec{z}^*)^x(\vec{z})^{x'}}_{(3).(2)}\,. \tag{3}$$

(3).(2) is a degree-$2d$ polynomial in the z_i's and their complex conjugates. Further, when the $|z_i|$'s are small, we can approximate term (3).(1) by $\exp\left(-\|z\|_2^2\right)$, which is a bounded scalar variable that we can hardcode into our constraints. While the $|z_i|$'s won't always be small, we can guess the ones that are large, fix their value and use the above approximation on the rest. This is where we pick up an ℓ_∞ constraint on the z_i's, since we must enforce that entries which we do not guess are small. This reduces the algorithm to solving problems of the following form.

$$\max_{\substack{\|\vec{z}\|_2=1\\ \|\vec{z}\|_\infty\leqslant\mu}} p(\vec{z}) = \max_{\substack{\|\vec{z}\|_2=1\\ \|\vec{z}\|_\infty\leqslant\mu}} \sum_{x,x'\in\{0,1\}^m} \langle x|\,\rho_d\,|x'\rangle\,(\vec{z}^*)^x(\vec{z})^{x'}.$$

Optimizing low-degree polynomials over the sphere is known to be hard to approximate up to polynomial factors in the worst-case, even when the degree is 4 [BBHKSZ12; BGGLT17]. However, in our case, $p(\vec{z})$ is not an arbitrary low-degree polynomial, but is quite 'small': the ℓ_2 norm of the coefficients $\langle x|\rho_d|x'\rangle$ is bounded, since it is $\|\rho_d\|_F \leqslant \|\rho\|_F \leqslant \mathrm{tr}(\rho) = 1$. We will show that, in this case, obtaining additive error that scales with ε admits a polynomial time algorithm. Additive-error approximations to max k-CSPs also admit a similar guarantee.

In the general case, we must also deal with the farness constraints in (Tangent Cover), $\mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}}\rangle) \geqslant 2/\eta$ for a small number of $|\pi_{\vec{a}}\rangle$'s. Recall that $\mathrm{d}_{\tan}(|\pi_{\vec{a}}\rangle, |\pi_{\vec{z}}\rangle)^2 = \sum_{i\in[n]}|\frac{z_i-a_i}{1+z_i^*a_i}|^2$ by definition. We will not try to directly optimize over these constraints. We use a similar trick as before and show that when the $|z_i|$'s are small, $\mathrm{d}_{\tan}(|\pi_{\vec{a}}\rangle, |\pi_{\vec{z}}\rangle) \approx \|\vec{a}-\vec{z}\|_2$ (see Lemma 3.9). These constraints essentially only enforce what $\vec{z}$ can be in the low-dimensional subspace spanned by the $\vec{a}$'s. So, we can guess the choice of $\vec{z}$ on this subspace (in addition to the coordinates for which $|z_i|$ is large), and for each guess, solve the problem with the guess hardcoded into the constraints. Putting all the steps together, we show that we can reduce the problem of extending the cover to a polynomial optimization problem, where the ℓ_2 norm of the coefficients is bounded by 1, subject to ℓ_2 and ℓ_∞ constraints.

Optimizing low-degree polynomials over the sphere. We now focus on the algorithmic problem of optimizing a low-degree polynomial over the sphere subject to ℓ_2 and ℓ_∞ constraints.

Our results for polynomial optimization can be thought of as analogs of maximizing dense k-CSPs, only the domain is the sphere instead of the hypercube. The underlying algorithms for

max k-CSPs are either based on brute-force search over a dimension-independent number of variables followed by greedily completing the solution or global correlation rounding [MS08; Yar14; MR17; AGT19]. One may expect the correlation rounding algorithms for max k-CSPs to generalize straightforwardly to optimize low-degree polynomials over the sphere up to additive error. However, the existing analysis [MR17] would merely translate to outputting a product state with fidelity $\Omega(\mathsf{OPT}) - \varepsilon$, as opposed to $\mathsf{OPT} - \varepsilon$. One can also try to extend the correlation rounding algorithm of Alev, Jeronimo and Tulsiani [AGT19] to the sphere, but their algorithm obtains a doubly-exponential dependence on k. In contrast, our approach is closer in spirit to the brute-force style algorithm for max k-CSPs, and allows for additional ℓ_2 and ℓ_∞ constraints. Translating our algorithm back to optimizing dense polynomials over the hypercube, we can show that we obtain yet another algorithm that achieves additive error guarantees.

To get the key ideas across, we first consider the unconstrained polynomial optimization problem, reformulated as maximizing the injective norm of a tensor:

$$\max_{x \in \mathbb{C}^m, \|\vec{x}\|_2 = 1} \langle T, \vec{x}^{\otimes k} \rangle$$

for a tensor T with $\|T\|_F \leqslant 1$. While it is hard to obtain a multiplicative approximation to tensor optimization problems, we show that we can obtain an additive $\varepsilon \cdot \|T\|_F$ approximation in $n^{\mathrm{poly}(1/\varepsilon)}$ time (see Theorem 6.3 for a formal statement). We begin by observing that there is a $\mathrm{poly}(1/\varepsilon)$-dimensional subspace such that projecting x onto this subspace suffices to obtain an $\varepsilon\|T\|_F$ approximation to the optimum objective value. To see this, we can unfold the tensor T along the first mode to obtain a $m \times m^{k-1}$ matrix M. We can then compute the singular value decomposition of M and let $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_m \geqslant 0$ be the resulting singular values. There are at most $r = \lceil 1/\varepsilon^2 \rceil$ singular values larger than $\varepsilon \cdot \|T\|_F$. Let $\Pi^{(1)}_{>\varepsilon}$ be the projection on the top-r subspace of M. Then,

$$\left| \langle \vec{x}, M(\mathsf{vec}(\vec{x}^{\otimes k-1})) \rangle - \langle \Pi^{(1)}_{>\varepsilon} \vec{x}, M(\mathsf{vec}(\vec{x}^{\otimes k-1})) \rangle \right| \leqslant \varepsilon \|T\|_F.$$

Now, we can repeat the same argument for the remaining modes to obtain projectors $\Pi^{(2)}_{>\varepsilon}, \ldots, \Pi^{(k)}_{>\varepsilon}$. Setting $\Pi_{>\varepsilon}$ to project onto the union of the spans of $\Pi^{(1)}, \ldots, \Pi^{(k)}$, we conclude that

$$\left| \langle T, \vec{x}^{\otimes k} \rangle - \langle T, \Pi_{>\varepsilon} \vec{x}^{\otimes k} \rangle \right| \leqslant k\varepsilon\,.$$

In short, *the polynomial can be approximated by projecting $\vec{x}$ onto a small-sized subspace.* To solve this tensor optimization problem, it suffices to brute-force over a fine-enough net on this constant-dimensional subspace and pick the vector that obtains the maximal value.

In general, we need to deal with an optimization problem that involves a dimension-independent number of ℓ_2 constraints and an ℓ_∞ constraint (see Definition 6.2). We handle the ℓ_2 constraints by simply projecting onto the union of the subspace Π and the subspace corresponding to the span of the ℓ_2 constraints. It remains to handle the ℓ_∞ constraint, which takes the form $\|\vec{x}\|_\infty \leqslant \mu$ for some constant $\mu > 0$. Only $1/\mu^2$ of the coordinates can saturate the ℓ_∞ constraint. Since $\mu > 0$ is a constant, we can brute-force over which coordinates saturate the ℓ_∞ constraint. Ultimately, we can still reduce the constrained optimization problem to checking over a net in a constant-dimensional subspace. We refer the reader to Section 6 for more details.

Hardness for agnostic product tomography. Our lower bound starts from the hardness of computing (asymmetric) tensor spectral norm for 4-tensors (see Theorem 7.2). In particular,

for an $n \times n \times n \times n$ tensor T, computing the spectral norm to additive error $\|T\|_F / \operatorname{poly}(n)$ is NP-hard. We attain our result by reducing tensor optimization to product state learning, essentially inverting the reduction discussed earlier. The main idea is to set the unknown state $\rho = |\psi\rangle\langle\psi|$ where

$$|\psi\rangle = \frac{1}{\|T\|_F} \sum_{i,j,k,l \in [m]} T_{ijkl} |e_i\rangle |e_j\rangle |e_k\rangle |e_l\rangle \ .$$

Then, we can show that finding the $4m$-qubit product state $|\pi_{\vec{x}}\rangle |\pi_{\vec{y}}\rangle |\pi_{\vec{u}}\rangle |\pi_{\vec{v}}\rangle$ with optimal fidelity is essentially equivalent to maximizing the tensor form $\langle T, \vec{x} \otimes \vec{y} \otimes \vec{u} \otimes \vec{v} \rangle$. The only additional difficulty is that this equivalence only holds if T is sufficiently flat; our reduction thus requires an additional step where we embed our input T in a larger space and randomly rotate it to make all its entries small without changing the optimal value.

Faster algorithms. In light of the lower bound, one can still ask what additional structure yields faster algorithms. We consider three additional settings: the high-fidelity regime (high overlap with a product state), a bounded number of choices for each qubit, and matrix-product states. In all of these settings, we follow the same overall strategy of sweeping over the qubits, but maintaining a cover becomes significantly easier:

1. In the high-fidelity regime, the cover can be made to be only *one* state;
2. In the finite-choices setting, we can simply maintain *all* good product states in the class, instead of a cover over them;
3. In the MPS setting, we can make our cover one state, though one with a large bond dimension, in some sense capturing many good product states.

For this overview, we focus on the high-fidelity setting. Here, the optimal solution is unique and we do not require maintaining a net. Instead, we only need to maintain a single candidate product state as we sweep across the qubits, performing local updates until convergence.

To illustrate how and why local optimization works, suppose for simplicity that $\rho = |\psi\rangle\langle\psi|$ is a pure state; the mixed state case is similar but involves some additional parameters. Imagine that $|0^n\rangle$ is the current candidate product state (in some appropriately chosen basis), and consider what happens when we express $|\psi\rangle$ in the low-Hamming weight subspace:

$$|\psi\rangle = \alpha |0^n\rangle + \delta |v_1\rangle + \beta |v_{\geqslant 2}\rangle \ .$$

Above, we assume without loss of generality that $|v_1\rangle$ is a normalized state on the subspace of Hamming weight 1, $|v_{\geqslant 2}\rangle$ is a normalized state on the subspace of Hamming weight at least 2, and α, β, δ are all nonnegative reals. It is helpful to express $|\psi\rangle$ this way because δ quantifies local updates that we can make to improve fidelity. By rotating qubit i of our candidate product state away from $|0\rangle$, we can increase the product state fidelity from $|\langle 0^n|\psi\rangle|^2 = \alpha^2$ to $\alpha^2 + \delta^2 |\langle e_i|v_1\rangle|^2$, where $|e_i\rangle$ is the string with 1 in position i and 0s elsewhere. Our goal, then, will be to establish that α^2 is close to OPT whenever δ is small, because it implies that local optimization converges efficiently: either δ is large, in which case we can increase the candidate fidelity by a substantial amount, or δ is small, in which case we are near the global optimum.

We prove this by bounding the contributions to product state fidelity from the Hamming weight 0, 1, and $\geqslant 2$ subspaces separately. Consider an arbitrary product state $|\pi\rangle$ that, when measured in the computational basis, gives $|0^n\rangle$ with probability $1 - p$ and anything else with probability p. We observe (Corollary 4.3) that $|\pi\rangle$ places at most $O(p^2)$ probability on Hamming

weight 2 or larger, and use this to upper bound the overlap between $|\pi\rangle$ and $|\psi\rangle$ (this is a simplified version of Equation (5)):

$$|\langle\pi|\psi\rangle| \leqslant \alpha(1 - \Omega(p)) + \delta\sqrt{p} + O(\beta p)\,.$$

Working out the constants, we find that so long as $\alpha^2 \geqslant 2/3$, the $-\Omega(\alpha p)$ term dominates the $O(\beta p)$ term, leaving us with $|\langle\pi|\psi\rangle| \leqslant \alpha + \delta\sqrt{p} \leqslant \alpha + \delta$. So, *every* product state satisfies $|\langle\pi|\psi\rangle|^2 \leqslant (\alpha + \delta)^2$, and therefore $\mathsf{OPT} \leqslant (\alpha + \delta)^2$.

The above analysis straightforwardly gives rise to a polynomial-time but suboptimal algorithm for finding the closest product state. We briefly summarize the additional tricks that are required to reduce the sample complexity to linear in n.

First, we observe that divide-and-conquer is more efficient than sweeping through one additional qubit at a time. So, the basic structure of the learning algorithm (Algorithm 4.19) is:

1. Recursively run the algorithm on the left and right halves of ρ, obtaining product states $|\pi_L\rangle$ and $|\pi_R\rangle$ that have fidelity at least $5/6$ with the respective halves.
2. Take $|\pi\rangle = |\pi_L\rangle \otimes |\pi_R\rangle$, which satisfies $\langle\pi|\rho|\pi\rangle \geqslant 2/3$.
3. Run local optimization on $|\pi\rangle$ until convergence.

Second, we improve the bound on $|\langle\pi|\psi\rangle|$ when α^2 is much larger than $2/3$. Namely, we show the alternative bound

$$|\langle\pi|\psi\rangle| \leqslant \alpha + O\left(\frac{\delta^2}{\alpha^2 - 2/3}\right),$$

which ultimately implies that local optimization needs fewer iterations to converge.

Third, we find that one can make larger improvements to the fidelity by updating all n qubits simultaneously, rather than one at a time. We take $\vec{z} \in \mathbb{C}^n$ to be the vector defined by $z_i = \langle e_i|\rho|0^n\rangle$, which captures the mass that ρ places on Hamming weight 1 that is coherent with $|0^n\rangle$. Then we show that a step from $|0^n\rangle$ to $|\pi_{\vec{z}/10}\rangle$ increases the fidelity with ρ by $\Omega\big(\|\vec{z}\|_2^2\big)$. Note that in the pure state case $\rho = |\psi\rangle\langle\psi|$, this $\vec{z}$ is precisely $\alpha\delta\,|v_1\rangle$, and therefore the fidelity improvement is $\Omega(\delta^2)$. For comparison, recall that the improvement from a single-qubit update was only $\delta^2 \max_{i\in[n]} |\langle e_i|v_1\rangle|^2$, which can be as small as δ^2/n.

Finally, we establish that it suffices to obtain a relative ℓ^2-error estimate of $\vec{z}$ in order to perform these local updates. In symbols, if we can produce an estimate $\vec{a}$ satisfying $\|\vec{a} - \vec{z}\|_2 \leqslant \|\vec{z}\|_2/3$ (say), then we show that $|\pi_{\vec{a}/10}\rangle$ also increases the fidelity by $\Omega\big(\|\vec{z}\|_2^2\big)$. This allows us to cut down some of the cost of the tomography subroutine by varying the error parameter throughout the algorithm, because we can afford to be sloppier when the step size is large.

2.1 Notation

Throughout, $\log_b$ is the logarithm base b, and log is shorthand for the natural logarithm base $e \approx 2.718$. The complex conjugate of $z \in \mathbb{C}$ is denoted z^*. For a function $f : \mathbb{R} \to \mathbb{R}$, $f'(x)$ denotes its derivative at x. An unspecified polynomially-bounded function of n may be written as $\mathrm{poly}(n)$.

We use standard shorthand for tensor products of quantum basis states. So, for example, $|0\rangle \otimes |1\rangle$ can be written either as $|0\rangle\,|1\rangle$ or $|01\rangle$.

$[n]$ is defined as the set of integers $\{1, 2, \ldots, n\}$. If $S \subseteq [n]$, $\overline{S}$ denotes its complement in $[n]$. When ρ is an n-qubit mixed state, we write $\rho_S := \mathrm{tr}_{\overline{S}}(\rho)$ for the reduced state on the qubits indexed by S.

The Hamming weight of a binary string $x \in \{0,1\}^n$ is denoted by $|x|$. We write $|e_i\rangle$ for the Hamming weight-1 n-qubit computational basis state that has $|1\rangle$ in the ith position and $|0\rangle$ everywhere else, leaving n implicit from context.

Vectors always have arrows over them (e.g. $\vec{v}$), unless they represent a (pure) quantum state, in which case we use ket notation (e.g. $|\psi\rangle$). If $\vec{v} \in \mathbb{C}^n$, $\vec{v}^* \in \mathbb{C}^n$ is its entrywise complex conjugate. For matrices $A \in \mathbb{C}^{m\times n}$, the conjugate transpose is $A^\dagger \in \mathbb{C}^{n\times m}$. The Euclidean (or ℓ^2) norm of a vector $\vec{v} \in \mathbb{C}^n$ is denoted by $\|\vec{v}\|_2 := \sqrt{\sum_i |v_i|^2}$, and the ℓ^∞ norm is denoted by $\|\vec{v}\|_\infty := \max_i |v_i|$. The norms that we use for matrices $A \in \mathbb{C}^{m\times n}$ are the operator norm $\|A\|_{\text{op}} := \sup_{\vec{v}\neq 0} \frac{\|A\vec{v}\|_2}{\|\vec{v}\|_2}$, the trace norm $\|A\|_1 := \text{tr}\left(\sqrt{A^\dagger A}\right)$, and the Frobenius norm $\|A\|_F := \sqrt{\text{tr}(A^\dagger A)}$. We also use $\|A\|_F$ for the Frobenius norm of a tensor A, which is the square root of the sum of the squared magnitudes of the entries.

3 Parametrization of product states

Definition 3.1 (Product state parametrization)**.** A product state $|\pi\rangle$ can be described by n parameters $\vec{z} \in (\mathbb{C} \cup \{\infty\})^n$ in the extended complex plane. We use the notation

$$|\pi_{\vec{z}}\rangle = \bigotimes_{i=1}^{n} \frac{|0\rangle + z_i |1\rangle}{\sqrt{1+|z_i|^2}}.$$

Note that this parametrization normalizes global phase such that the $|0^n\rangle$ amplitude, or more generally, the non-zero amplitude with lowest Hamming weight, is real. Though we define this parametrization for $\vec{z}$ with entries in the extended complex plane, we will generally work with $\vec{z} \in \mathbb{C}^n$ for simplicity: our algorithms incur error, so we will always be able to work with vectors with finite (but possibly very large) entries. However, this limitation is not necessary, and the diligent reader can extend our results to hold even with parameters at infinity.

3.1 Tangent distance

We now define a distance measure between product states which behaves nicely with respect to our parametrization, which we call the *tangent distance*. For the sake of building intuition, we first consider the case when $n = 1$:

Definition 3.2 (Tangent distance, $n = 1$)**.** For $z, a \in \mathbb{C} \cup \{\infty\}$, we define the following distance between 1-qubit states:

$$\text{d}_{\text{tan}}(|\pi_z\rangle, |\pi_a\rangle) = \left|\frac{z_i - a_i}{1 + z_i^* a_i}\right|.$$

We also have that $\text{d}^2_{\text{tan}}(|\pi_z\rangle, |\pi_a\rangle) = \tan^2(\theta/2)$ where $\theta \in [0, \pi]$ is the angle between the Bloch sphere points for $|\pi_z\rangle$ and $|\pi_a\rangle$.

Remark 3.3. Presumably the Bloch sphere interpretation is well known but we could not find a reference. Here is a quick proof: The Bloch sphere point for $|\pi_z\rangle$ is the inverse stereographic projection of z; namely $\vec{p}_z := \frac{1}{1+|z|^2}(z + z^*, z - z^*, 1 - |z|^2)$. From this it follows that $\cos\theta = \vec{p}_z \cdot \vec{p}_a = 1 - \frac{2|z-a|^2}{(1+|z|^2)(1+|a|^2)}$. Now use $\tan^2(\theta/2) = \frac{1-\cos\theta}{1+\cos\theta}$.

Definition 3.4 (Tangent distance, general n)**.** For two product states $|\pi_{\vec{z}}\rangle$ and $|\pi_{\vec{a}}\rangle$, we define the distance measure $\mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}}\rangle)$ via

$$\mathrm{d}^2_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}}\rangle) = \sum_{i=1}^{n} \mathrm{d}_{\tan}(|\pi_{z_i}\rangle, |\pi_{a_i}\rangle)^2 = \left\| \frac{\vec{z} - \vec{a}}{1 + \vec{z}^* \vec{a}} \right\|_2^2 .$$

Above, we abuse notation by using product and quotient of vectors to denote their entry-wise product and quotient.

This distance measure has several nice properties: it satisfies $\mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{z}}\rangle) = 0$; it is symmetric; and it is invariant under applying single-qubit unitaries. This last condition is evident from the Bloch sphere interpretation. However, it is not a metric: the triangle inequality does not hold, since the distance measure is infinity for orthogonal product states:

$$\mathrm{d}_{\tan}(|+\rangle, |0\rangle) + \mathrm{d}_{\tan}(|0\rangle, |-\rangle) = 2 < \mathrm{d}_{\tan}(|+\rangle, |-\rangle) = \infty .$$

Nevertheless, we can think of this as being approximately a metric when the denominator is "close to constant", i.e. when the two product states, and therefore their parameters, are close.

Remark 3.5 (Generalizing to qudits)**.** We can take the following parametrization over qudits, which takes $(d-1)n$ parameters $\boldsymbol{z} \in (\mathbb{C} \cup \{\infty\})^{(d-1)n}$:

$$|\pi_{\boldsymbol{z}}\rangle = \bigotimes_{i=1}^{n} \frac{|0\rangle + \sum_{j=1}^{d} z_{i,j} |j\rangle}{\sqrt{1 + \|\vec{z}_i\|_2^2}} .$$

Here we imagine partitioning $\boldsymbol{z}$ into n many vectors $\vec{z}_i$ of size $d-1$. The qudit version of tangent distance becomes

$$\mathrm{d}_{\tan}(|\pi_{\boldsymbol{z}}\rangle, |\pi_{\boldsymbol{a}}\rangle) = \Big(\sum_{i=1}^{n} \frac{\|\vec{z}_i - \vec{a}_i\|^2 + \|\vec{z}_i\|_2^2 \|\vec{a}_i\|_2^2 - |\langle \vec{z}_i, \vec{a}_i \rangle|^2}{|1 + \langle \vec{z}_i, \vec{a}_i \rangle|^2} \Big)^{1/2},$$

which we can see by rotating every qudit to reduce to the qubit definition. Notice that this reduces to the qubit definition when $d = 2$. Though not immediate, we anticipate no barriers in generalizing our results to qudits. Generalized versions of the lemmas to follow hold, and we can run the algorithms and analyses accordingly.

We define tangent distance to be a version of fidelity which behaves more nicely with respect to the parametrization. Tangent distance is related to fidelity in the following way.

Lemma 3.6 (Relationship between tangent distance and fidelity)**.** *For all product states $|\pi_{\vec{z}}\rangle$ and $|\pi_{\vec{a}}\rangle$, the following holds:*

$$\log\left(\frac{1}{|\langle \pi_{\vec{z}} | \pi_{\vec{a}} \rangle|^2} \right) \leqslant \mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}}\rangle)^2 \leqslant \frac{1}{|\langle \pi_{\vec{z}} | \pi_{\vec{a}} \rangle|^2} - 1 .$$

Both inequalities are tight.

Proof. First, because both fidelity and $\mathrm{d}_{\tan}$ are invariant under single-qubit unitaries, it suffices to show the inequalities for $|\pi_{\vec{a}}\rangle = |0^n\rangle$. Then, $|\langle \pi_{\vec{z}} | 0^n \rangle|^2 = \prod_{i=1}^{n} \frac{1}{1+|z_i|^2}$ and $\mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |0^n\rangle)^2 = \|\vec{z}\|_2^2$. Consequently,

$$\begin{aligned} \log\left(\frac{1}{|\langle \pi_{\vec{z}} | 0^n \rangle|^2} \right) = \sum_{i=1}^{n} \log(1 + |z_i|^2) &\leqslant \sum_{i=1}^{n} |z_i|^2 = \mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |0^n\rangle)^2 \\ &\leqslant \prod_{i=1}^{n} (1 + |z_i|^2) - 1 = \frac{1}{|\langle \pi_{\vec{z}} | 0^n \rangle|^2} - 1 . \end{aligned}$$

The first inequality is tight for $\vec{z} = 0^n$ and the right inequality is tight for $\vec{z} = (\infty, 0^{n-1})$. ♦

The approximation in Lemma 3.6 can be made tighter when $\vec{z}$ and $\vec{a}$ are closer together. For simplicity, we state the following lemma when $\vec{a} = 0^n$.

Lemma 3.7 (Approximation of $\mathrm{d}_{\tan}$ at small distances). *For a vector $\vec{z} \in \mathbb{C}^n$,*

$$e^{-\mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |0^n\rangle)^2} \leqslant \prod_{i=1}^{n} \frac{1}{1+|z_i|^2} \leqslant e^{-\mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |0^n\rangle)^2 + \sum_{i=1}^{n} |z_i|^4}.$$

Note that $\prod_{i=1}^{n} \frac{1}{1+|z_i|^2} = |\langle \pi_{\vec{z}} | 0^n \rangle|^2$.

Proof. The first inequality is equivalent to the first inequality in Lemma 3.6. For the second inequality, we use that, for $x \geqslant 0$,

$$e^{-x} \leqslant \frac{1}{1+x} \leqslant 1 - x + x^2 \leqslant e^{-x+x^2}.$$

So, this gives us that

$$\prod_{i=1}^{n} \frac{1}{1+|z_i|^2} \leqslant \prod_{i=1}^{n} e^{-|z_i|^2 + |z_i|^4} = e^{-\mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |0^n\rangle)^2 + \sum_{i=1}^{n} |z_i|^4}. \qquad \blacklozenge$$

As a corollary, we can also relate tangent distance to trace distance.

Corollary 3.8. *For $\vec{z}, \vec{a} \in \mathbb{C}^n$, $\frac{1}{2} \| |\pi_{\vec{z}}\rangle\langle\pi_{\vec{z}}| - |\pi_{\vec{a}}\rangle\langle\pi_{\vec{a}}| \|_1 = \| |\pi_{\vec{z}}\rangle\langle\pi_{\vec{z}}| - |\pi_{\vec{a}}\rangle\langle\pi_{\vec{a}}| \|_{\mathrm{op}} \leqslant \mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}}\rangle)$.*

Proof. The equality holds because $|\pi_{\vec{z}}\rangle\langle\pi_{\vec{z}}| - |\pi_{\vec{a}}\rangle\langle\pi_{\vec{a}}|$ is traceless and rank two. The inequality holds because

$$\begin{aligned} \frac{1}{2} \| |\pi_{\vec{z}}\rangle\langle\pi_{\vec{z}}| - |\pi_{\vec{a}}\rangle\langle\pi_{\vec{a}}| \|_1 &\leqslant \sqrt{1 - |\langle \pi_{\vec{z}} | \pi_{\vec{a}} \rangle|^2} \\ &\leqslant \sqrt{1 - e^{-\mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}}\rangle)^2}} \\ &\leqslant \mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}}\rangle), \end{aligned}$$

where the first inequality is Fuchs–van de Graaf, the second is Lemma 3.6, and the third is $1 + x \leqslant e^x$. $\blacklozenge$

We can further simplify $\mathrm{d}_{\tan}$ to a simple ℓ^2 norm when $\vec{z}$ and $\vec{a}$ are close entry-wise.

Lemma 3.9. *Let $\vec{z}, \vec{a} \in \mathbb{C}^n$. Then*

$$\left| \mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}}\rangle) - \|\vec{z} - \vec{a}\|_2 \right| \leqslant \mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}}\rangle)(\max_i |z_i||a_i|).$$

Proof. Consider a single qubit i. Then,

$$\left| \frac{z_i - a_i}{1 + z_i^* a_i} - (z_i - a_i) \right| = \left| \frac{z_i - a_i}{1 + z_i^* a_i} \right| |1 - (1 + z_i^* a_i)| = \left| \frac{z_i - a_i}{1 + z_i^* a_i} \right| |z_i||a_i|. \tag{4}$$

So, we can conclude

$$\left| \mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}}\rangle) - \|\vec{z} - \vec{a}\|_2 \right| \leqslant \left\| \frac{\vec{z} - \vec{a}}{1 + \vec{z}^* \vec{a}} - (\vec{z} - \vec{a}) \right\|_2 \leqslant \mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}}\rangle)(\max_i |z_i||a_i|). \qquad \blacklozenge$$

3.2 Approximation lemmas

Lemma 3.10 (Low-weight truncation of product states)**.** *For a product state $|\pi_{\vec{z}}\rangle$, let $\mu = \sum_{i=1}^{n} \frac{|z_i|^2}{1+|z_i|^2}$. Then for $c \leqslant \mu$ and $d \geqslant \mu$,*

$$\|\Pi_{\geqslant d} |\pi_{\vec{z}}\rangle\|_2^2 \leqslant e^{-d\log(d/\mu)+(d-\mu)},$$
$$\|\Pi_{\leqslant c} |\pi_{\vec{z}}\rangle\|_2^2 \leqslant e^{-(2\mu-c)\log(2-c/\mu)+(\mu-c)},$$

where $\Pi_{\geqslant d}$ is the projection onto computational basis states $|b\rangle$ such that $|b| \geqslant d$, and similarly for $\Pi_{\leqslant c}$.

Proof. Let $p_i = \frac{|z_i|^2}{1+|z_i|^2}$ be the probability that qubit i of $|\pi_{\vec{z}}\rangle$ outputs $|1\rangle$ when measured in the computational basis. Consider n independent Bernoulli random variables $Z_1, \ldots, Z_n$, where $\mathbf{Pr}[Z_i = 1] = p_i$. Then, because the qubits of $\pi_{\vec{z}}$ are uncorrelated,

$$\|\Pi_{\geqslant k} |\pi_{\vec{z}}\rangle\|_2^2 = \mathbf{Pr}[Z_1 + \cdots + Z_n \geqslant k].$$

Let $\mu = \mathbb{E}[Z_1 + \cdots + Z_n] = p_1 + \cdots + p_n$. By Bennett's inequality [BLM13, Theorem 2.9], for any $t \geqslant 0$,

$$\mathbf{Pr}\Big[\sum_{i=1}^{n}(Z_i - p_i) \geqslant t\Big] \leqslant e^{-\mu((1+t/\mu)\log(1+t/\mu)-t/\mu)} = e^{-(t+\mu)\log(1+t/\mu)+t}.$$

The same bound holds on the other tail:

$$\mathbf{Pr}\Big[\sum_{i=1}^{n}(Z_i - p_i) \leqslant -t\Big] \leqslant e^{-(t+\mu)\log(1+t/\mu)+t}.$$

The statement follows upon taking $d = \mu + t$ in the first inequality and $c = \mu - t$ in the second. ♦

4 High-fidelity product state learning

In this section, we give a simple polynomial-time algorithm for product state agnostic learning in the high signal-to-noise regime, where the fidelity of the unknown state ρ with the closest product state $|\pi\rangle$ is sufficiently close to 1. The algorithm operates by local optimization, where a candidate product state approximation is updated on the Hamming weight-1 subspace until convergence. We argue this algorithm's correctness by showing that once all of the local updates are sufficiently small, the algorithm must be close to a global optimum. The analysis involves bounding the optimal product state fidelity $\max_{\text{product } |\pi\rangle} \langle\pi|\rho|\pi\rangle$ in terms of the projection of ρ onto the subspace of Hamming weight 0 or 1.

4.1 Properties of product distributions

We begin by showing some simple concentration bounds on the Hamming weight of a product distribution.

Lemma 4.1. *Let π be a product distribution over $\{0,1\}^n$, and let $p_0 = \mathbf{Pr}_{x\sim\pi}[x = 0^n]$. Then*

$$\mathbf{Pr}_{x\sim\pi}[|x| = 1] \geqslant -p_0 \log p_0,$$

with the understanding that $0 \log 0 = 0$.

Proof. Let $a_i = \mathbf{Pr}_{x\sim\pi}[x_i = 0]$. If any a_i are equal to 0, we have $p_0 = 0$ and the lemma is trivial. Otherwise, assuming all a_i are strictly positive,

$$\begin{aligned}
\mathbf{Pr}_{x\sim\pi}[|x| = 1] &= \sum_{i=1}^{n}(1-a_i)\prod_{j\neq i} a_j \\
&= \sum_{i=1}^{n}\frac{1-a_i}{a_i}\prod_{j=1}^{n} a_j \\
&= p_0\sum_{i=1}^{n}\frac{1-a_i}{a_i} \\
&\geqslant p_0\sum_{i=1}^{n}\log(1/a_i) && (x-1\geqslant\log x) \\
&= -p_0\log\left(\prod_{i=1}^{n} a_i\right) \\
&= -p_0\log p_0\,. && \blacklozenge
\end{aligned}$$

Next, we prove a useful upper bound on the quantity $p_0 \log p_0$ appearing in the previous lemma.

Lemma 4.2. *For all $p \in [0,1]$,*

$$1 - p + p\log p \leqslant 2(1-\sqrt{p})^2\,,$$

with the understanding that $0\log 0 = 0$.

Proof. Define

$$f(p) \coloneqq 2(1-\sqrt{p})^2 - 1 + p - p\log(p)\,.$$

Notice that $f(1) = 0$. Moreover, the derivative of f satisfies:

$$f'(p) = 2 - \frac{2}{\sqrt{p}} - \log p \leqslant 0\,,$$

because $1 - \frac{1}{x} \leqslant \log(x)$ for all $x > 0$ (substituting $x = \sqrt{p}$). So, f is decreasing on $[0,1]$, and therefore $f(p) \geqslant 0$ for all $p \in [0,1]$, which implies the lemma. $\blacklozenge$

Combining the previous two lemmas gives an upper bound on the probability assigned by a product distribution to strings of Hamming weight at least 2.

Corollary 4.3. *Let π be a product distribution over $\{0,1\}^n$, and let $p_0 = \mathbf{Pr}_{x\sim\pi}[x = 0^n]$. Then*

$$\mathbf{Pr}_{x\sim\pi}[|x| \geqslant 2] \leqslant 2(1-\sqrt{p_0})^2\,.$$

Proof. Follows from Lemma 4.1 and Lemma 4.2. $\blacklozenge$

4.2 Characterizing optimal product approximations

Proved in this section is the theorem below, which is the heart of the algorithm's correctness.

Theorem 4.4. *Consider an arbitrary quantum state $|\psi\rangle$, which we express in the form*

$$|\psi\rangle = \left(\alpha\,|0^n\rangle + \delta\,|v_1\rangle + \beta\,|v_{\geqslant 2}\rangle\right)|g\rangle + \gamma\,|\perp\rangle\,.$$

Assume without loss of generality that

- α, β, δ, *and* γ *are all real and nonnegative,*
- $|v_1\rangle$, $|v_{\geqslant 2}\rangle$, $|g\rangle$, *and* $|\perp\rangle$ *are unit vectors,*
- $|v_1\rangle$ *is supported only on strings of Hamming weight* 1*,*
- $|v_{\geqslant 2}\rangle$ *is supported only on strings of Hamming weight* 2 *or larger, and*
- $|\perp\rangle$ *is orthogonal to all states beginning with* $|0^n\rangle$ *or ending with* $|g\rangle$*.*

Suppose further that $\alpha^2 = 2/3 + c$ *for some* $c \geqslant 0$*. Then for all product states* $|\pi\rangle$ *on* n *qubits:*

$$\|(\langle\pi| \otimes I)\,|\psi\rangle\|_2^2 \leqslant \left(\alpha + \min\left\{\delta, \sqrt{\frac{2}{27}}\frac{\delta^2}{c}\right\}\right)^2.$$

Before establishing this theorem, let us explain how to interpret it. Suppose we are searching for the best product state approximation to the leftmost n qubits of some state $|\psi\rangle$. Our current candidate for the best product state is $|0^n\rangle$ (in some appropriately chosen product basis), whose fidelity with the leftmost n qubits is currently $\alpha^2 = 2/3 + c$.

First consider the special case where $|\psi\rangle$ itself is an n-qubit pure state, in which case we may take $|g\rangle = 1$ and $\gamma = 0$ to write

$$|\psi\rangle = \alpha\,|0^n\rangle + \delta\,|v_1\rangle + \beta\,|v_{\geqslant 2}\rangle$$

as a sum over basis states of Hamming weight 0, 1, and 2 or greater. Then, the theorem says that if $|\psi\rangle$'s support on Hamming weight 1 is small (as captured by δ), $|0^n\rangle$ must approximately maximize the product state fidelity with $|\psi\rangle$.

In the general case, where $|\psi\rangle$ has more than n qubits, the theorem similarly shows that $|0^n\rangle$ is an approximate maximizer of product fidelity, but under a slightly different assumption: that $|\psi\rangle$ places small support on states of Hamming weight 1 *that are coherent with* $|0^n\rangle$ *on the leftmost* n *qubits*. In other words, $|\psi\rangle$ may place large mass on Hamming weight 1 in $|\perp\rangle$, but this does not affect the bound on product state fidelity.

Now we proceed towards the proof. We first establish two quantitative bounds, whose importance will become clear in the proof of Theorem 4.4.

Lemma 4.5. *Let* p*,* β*,* γ*, and* r *be nonnegative reals satisfying* $0 \leqslant p \leqslant 1$*,* $\beta^2 + \gamma^2 \leqslant 1/3$*, and* $r \geqslant \sqrt{2/3}$*. Then:*

$$\left(r\left(1 - \frac{p}{2}\right) + \frac{\beta\sqrt{2}}{2}p\right)^2 + \gamma^2 p \leqslant \left(r\left(1 - \frac{p}{2}\right) + \frac{\sqrt{\beta^2+\gamma^2}\sqrt{2}}{2}p\right)^2.$$

Proof. Expanding out, the desired inequality becomes

$$r^2\left(1 - \frac{p}{2}\right)^2 + \beta\sqrt{2}rp\left(1 - \frac{p}{2}\right) + \frac{\beta^2}{2}p^2 + \gamma^2 p \leqslant r^2\left(1 - \frac{p}{2}\right)^2 + \sqrt{\beta^2+\gamma^2}\sqrt{2}rp\left(1 - \frac{p}{2}\right) + \frac{\beta^2+\gamma^2}{2}p^2,$$

which considerably simplifies to

$$\gamma^2 p\left(1 - \frac{p}{2}\right) \leqslant \sqrt{\beta^2+\gamma^2}\sqrt{2}rp\left(1 - \frac{p}{2}\right) - \beta\sqrt{2}rp\left(1 - \frac{p}{2}\right).$$

Factoring out $p(1-\frac{p}{2}) \geqslant 0$, it suffices to show that

$$\gamma^2 \leqslant r\sqrt{2}\left(\sqrt{\beta^2+\gamma^2}-\beta\right).$$

We can assume henceforth that β and γ are both strictly positive, because the inequality above is easily verified if either are zero. Multiplying both sides by $\sqrt{\beta^2+\gamma^2}+\beta$ is equivalent to

$$\gamma^2\left(\sqrt{\beta^2+\gamma^2}+\beta\right) \leqslant r\sqrt{2}\gamma^2,$$

and then we divide by γ^2 to obtain

$$\sqrt{\beta^2+\gamma^2}+\beta \leqslant r\sqrt{2}.$$

This is implied by the assumptions of the lemma because the left side is at most $\frac{2}{\sqrt{3}}$, and the right side is at least $\frac{2}{\sqrt{3}}$. ♦

Lemma 4.6. *Suppose that $\alpha^2 = 2/3 + c$ for some $c \geqslant 0$. Then:*

$$\alpha - \sqrt{2-2\alpha^2} \geqslant c\sqrt{\frac{27}{8}}.$$

Proof. In other words, we wish to show that

$$f(\alpha) := \alpha - \sqrt{2-2\alpha^2} - \sqrt{\frac{27}{8}}\left(\alpha^2 - \frac{2}{3}\right) \geqslant 0.$$

Consider the first and second derivatives of f:

$$f'(\alpha) = 1 + \frac{2\alpha}{\sqrt{2-2\alpha^2}} - \sqrt{\frac{27}{2}}\alpha, \qquad f''(\alpha) = \frac{\sqrt{2}}{(1-\alpha^2)^{3/2}} - \sqrt{\frac{27}{2}}.$$

A simple calculation shows that $f''(\sqrt{2/3}) > 0$, and $f''(\alpha)$ is clearly increasing in α, so $f''(\alpha) > 0$ for all $\alpha \geqslant \sqrt{2/3}$. $f'(\sqrt{2/3}) = 0$, and $f'(\alpha)$ is increasing in α as a consequence of the positive second derivative, so $f'(\alpha) \geqslant 0$ for all $\alpha \geqslant \sqrt{2/3}$. Hence, f is non-decreasing for $\alpha \geqslant \sqrt{2/3}$. Since $f(\sqrt{2/3}) = 0$, the lemma follows. ♦

Proof of Theorem 4.4. Write

$$|\pi\rangle = \sqrt{p_0}\,|0^n\rangle + \sqrt{p_1}\,|w_1\rangle + \sqrt{p_{\geqslant 2}}\,|w_{\geqslant 2}\rangle\,,$$

where p_0, p_1, and $p_{\geqslant 2}$ are probabilities summing to 1, $|w_1\rangle$ is supported over strings of weight 1, $|w_{\geqslant 2}\rangle$ is supported over strings of weight at least 2, and both $|w_1\rangle$ and $|w_{\geqslant 2}\rangle$ are unit vectors. Then:

$$\begin{aligned}
\|(\langle\pi|\otimes I)\,|\psi\rangle\|_2^2 &= \|(\alpha\sqrt{p_0} + \delta\sqrt{p_1}\,\langle w_1|v_1\rangle + \beta\sqrt{p_{\geqslant 2}}\,\langle w_{\geqslant 2}|v_{\geqslant 2}\rangle)\,|g\rangle + \gamma(\langle\pi|\otimes I)\,|\perp\rangle\|_2^2 \\
&= \|(\alpha\sqrt{p_0} + \delta\sqrt{p_1}\,\langle w_1|v_1\rangle + \beta\sqrt{p_{\geqslant 2}}\,\langle w_{\geqslant 2}|v_{\geqslant 2}\rangle)\,|g\rangle\|_2^2 + \|\gamma(\langle\pi|\otimes I)\,|\perp\rangle\|_2^2 \\
&= |\alpha\sqrt{p_0} + \delta\sqrt{p_1}\,\langle w_1|v_1\rangle + \beta\sqrt{p_{\geqslant 2}}\,\langle w_{\geqslant 2}|v_{\geqslant 2}\rangle|^2 + \gamma^2\|(\langle\pi|\otimes I)\,|\perp\rangle\|_2^2 \\
&\leqslant (\alpha\sqrt{p_0} + \delta\sqrt{p_1} + \beta\sqrt{p_{\geqslant 2}})^2 + \gamma^2(1-p_0)\,,
\end{aligned}$$

where in the second line we used the Pythagorean theorem which is valid because $|\perp\rangle$ has no support on $|g\rangle$, and in the last line we applied the triangle inequality and the assumption that

$|\perp\rangle$ has no support on $|0^n\rangle$. We first turn our attention to bounding (the square root of) the left term. Let $p = p_1 + p_{\geqslant 2}$ be the probability that the measurement distribution of $|\pi\rangle$ assigns to strings of Hamming weight 1 or more. Then:

$$\begin{aligned}
\alpha\sqrt{p_0} + \delta\sqrt{p_1} + \beta\sqrt{p_{\geqslant 2}} &\leqslant \alpha\sqrt{p_0} + \delta\sqrt{p_1} + \beta\sqrt{2}(1-\sqrt{p_0}) && \text{(Corollary 4.3)} \\
&= (\alpha - \beta\sqrt{2})\sqrt{p_0} + \delta\sqrt{p_1} + \beta\sqrt{2} \\
&\leqslant (\alpha - \beta\sqrt{2})\sqrt{1-p} + \delta\sqrt{p} + \beta\sqrt{2} && (p_0 = 1-p, p_1 \leqslant p) \\
&\leqslant (\alpha - \beta\sqrt{2})\left(1 - \frac{p}{2}\right) + \delta\sqrt{p} + \beta\sqrt{2} && (\sqrt{1-p} \leqslant 1 - \frac{p}{2}, \alpha \geqslant \beta\sqrt{2}) \\
&= \alpha\left(1 - \frac{p}{2}\right) + \frac{\beta\sqrt{2}}{2}p + \delta\sqrt{p}\,.
\end{aligned}$$

Substituting, we find that

$$\begin{aligned}
\|(\langle\pi| \otimes I)\,|\psi\rangle\|_2^2 &\leqslant \left(\alpha\left(1 - \frac{p}{2}\right) + \frac{\beta\sqrt{2}}{2}p + \delta\sqrt{p}\right)^2 + \gamma^2 p \\
&\leqslant \left(\alpha\left(1 - \frac{p}{2}\right) + \frac{\sqrt{\beta^2+\gamma^2}\sqrt{2}}{2}p + \delta\sqrt{p}\right)^2 && \text{(Lemma 4.5)} \\
&\leqslant \left(\alpha - \frac{p}{2}\left(\alpha - \sqrt{2-2\alpha^2}\right) + \delta\sqrt{p}\right)^2 && (\alpha^2+\beta^2+\gamma^2 \leqslant 1) \\
&\leqslant \left(\alpha - p\sqrt{\frac{27}{32}}c + \delta\sqrt{p}\right)^2. && \text{(Lemma 4.6)} \qquad (5)
\end{aligned}$$

The use of Lemma 4.5 in the second line is by choosing $r = \frac{\alpha\left(1-\frac{p}{2}\right)+\delta\sqrt{p}}{1-\frac{p}{2}} \geqslant \alpha \geqslant \sqrt{2/3}$. (To give some interpretation for this use of Lemma 4.5, it effectively says that we can assume without loss of generality that $\gamma = 0$, by placing all of its amplitude on β instead.) To bound this last quantity, we first observe that

$$\begin{aligned}
p\sqrt{\frac{27}{32}}c &\leqslant \sqrt{\frac{3}{32}} && (0 \leqslant p \leqslant 1, 0 \leqslant c \leqslant 1/3) \\
&< \sqrt{\frac{2}{3}} \\
&\leqslant \alpha\,,
\end{aligned}$$

and therefore the expression inside the parentheses is always positive. So, it suffices to upper bound what is in the parentheses. Next, we note that

$$\sqrt{\frac{2}{27}}\frac{\delta^2}{c} - \delta\sqrt{p} + p\sqrt{\frac{27}{32}}c \geqslant 0\,,$$

because the discriminant of the left side (as a quadratic function of δ) is 0. Plugging into the bound obtained in Equation (5) gives

$$\|(\langle\pi| \otimes I)\,|\psi\rangle\|_2^2 \leqslant \left(\alpha + \sqrt{\frac{2}{27}}\frac{\delta^2}{c}\right)^2.$$

Alternatively, we can take

$$\|(\langle\pi| \otimes I)\,|\psi\rangle\|_2^2 \leqslant (\alpha + \delta)^2$$

by using $0 \leqslant p \leqslant 1$. ♦

We conclude this subsection by proving a version of Theorem 4.4 for mixed states.

Corollary 4.7. *For an n-qubit density matrix ρ, define $\vec{z} \in \mathbb{C}^n$ by*

$$z_i = \langle e_i|\rho|0^n\rangle \ .$$

If $\langle 0^n|\rho|0^n\rangle = 2/3 + c$ for some $c > 0$, then for all product states $|\pi\rangle$:

$$\langle \pi|\rho|\pi\rangle \leqslant \langle 0^n|\rho|0^n\rangle + \min\left\{3\|\vec{z}\|_2, \frac{\|\vec{z}\|_2^2}{c}\right\}.$$

Proof. Take $|\psi\rangle$ to be a purification of ρ in the form of Theorem 4.4:

$$|\psi\rangle = (\alpha\,|0^n\rangle + \delta\,|v_1\rangle + \beta\,|v_{\geqslant 2}\rangle)\,|g\rangle + \gamma\,|\perp\rangle \ .$$

Observe that

$$\delta^2 = \sum_{i=1}^{n} \frac{|\langle e_i|\rho|0^n\rangle|^2}{\alpha^2} = \frac{\|\vec{z}\|_2^2}{\alpha^2},$$

because $z_i = \langle e_i|\rho|0^n\rangle = \alpha\delta\,\langle e_i|v_1\rangle$. Hence:

$$\begin{aligned}
\langle \pi|\rho|\pi\rangle &= \|(\langle\pi| \otimes I)\,|\psi\rangle\|_2^2 \\
&\leqslant \left(\alpha + \frac{\|\vec{z}\|_2}{\alpha}\right)^2 && \text{(Theorem 4.4)} \\
&= \alpha^2 + 2\|\vec{z}\|_2 + \frac{\|\vec{z}\|_2^2}{\alpha^2} \\
&= \alpha^2 + \left(2 + \frac{\delta}{\alpha}\right)\|\vec{z}\|_2 \\
&\leqslant \alpha^2 + 3\|\vec{z}\|_2 \,. && (\delta \leqslant \sqrt{1/3}, \alpha \geqslant \sqrt{2/3})
\end{aligned}$$

We can also use the other half of Theorem 4.4 to obtain:

$$\begin{aligned}
\langle \pi|\rho|\pi\rangle &= \|(\langle\pi| \otimes I)\,|\psi\rangle\|_2^2 \\
&\leqslant \left(\alpha + \sqrt{\frac{2}{27}\frac{\|\vec{z}\|_2^2}{\alpha^2 c}}\right)^2 && \text{(Theorem 4.4)} \\
&\leqslant \left(\alpha + \frac{\|\vec{z}\|_2^2}{3\alpha c}\right)^2 && (\alpha \geqslant \sqrt{2/3}) \\
&= \alpha^2 + \frac{2\|\vec{z}\|_2^2}{3c} + \frac{\|\vec{z}\|_2^4}{9\alpha^2 c^2} \,.
\end{aligned}$$

To complete the proof, assume without loss of generality that $\frac{\|z\|_2^2}{c} \leqslant 1/3$, as otherwise the statement is trivial. Then:

$$\begin{aligned}
\langle \pi|\rho|\pi\rangle &\leqslant \alpha^2 + \frac{2\|\vec{z}\|_2^2}{3c} + \frac{\|\vec{z}\|_2^2}{27\alpha^2 c} \\
&\leqslant \alpha^2 + \frac{2\|\vec{z}\|_2^2}{3c} + \frac{\|\vec{z}\|_2^2}{18c} && (\alpha^2 \geqslant 2/3) \\
&\leqslant \alpha^2 + \frac{\|z\|_2^2}{c} \,.
\end{aligned}$$

♦

4.3 Bounding local updates

Corollary 4.7 shows that the maximum product fidelity can be bounded in terms of the ℓ^2 norm of a certain vector $\vec{z}$, where $\vec{z}$ captures mass that ρ places coherently between $|0^n\rangle$ and strings of Hamming weight 1. In this subsection, we show a sort of converse: that there always exists a product state $|\pi\rangle$ whose increase in fidelity compared to $|0^n\rangle$ is proportional to $\|\vec{z}\|_2^2$. So, we can use $\vec{z}$ to guide our local optimization algorithm until convergence.

We first need a claim showing that $\|\vec{z}\|_2$ is bounded:

Claim 4.8. For an n-qubit density matrix ρ, define $\vec{z} \in \mathbb{C}^n$ by

$$z_i = \langle e_i|\rho|0^n\rangle \ .$$

Then $\|\vec{z}\|_2 \leqslant 1/2$.

Proof. Following the proof of Corollary 4.7, let $|\psi\rangle$ to be a purification of ρ in the form of Theorem 4.4:

$$|\psi\rangle = (\alpha\,|0^n\rangle + \delta\,|v_1\rangle + \beta\,|v_{\geqslant 2}\rangle)\,|g\rangle + \gamma\,|\perp\rangle \ .$$

Recall that

$$\delta^2 = \sum_{i=1}^{n} \frac{|\langle e_i|\rho|0^n\rangle|^2}{\alpha^2} = \frac{\|\vec{z}\|_2^2}{\alpha^2} \ ,$$

and therefore $\|\vec{z}\|_2 = \alpha\delta$. Since $\alpha^2 + \delta^2 \leqslant 1$, $\|\vec{z}\|_2$ is maximized when $\alpha = \delta = \sqrt{2}$, and the claim follows. ♦

Now we show how to make local updates based on $\vec{z}$. To understand the theorem below, think of $\vec{a}$ as a "good enough" approximation to $\vec{z}$. (We require an approximation because quantum tomography algorithms are necessarily non-exact.) Then the theorem shows that, using the product state parametrization from Definition 3.1, moving in the direction of $\vec{a}$ always increases the fidelity by an amount proportional to $\|\vec{a}\|_2^2$.

Theorem 4.9. *For an n-qubit density matrix ρ, define $\vec{z} \in \mathbb{C}^n$ by*

$$z_i = \langle e_i|\rho|0^n\rangle \ .$$

Then if $\|\vec{a} - \vec{z}\|_2 \leqslant \|\vec{a}\|_2/2$,

$$\langle \pi_{\vec{a}/10}|\rho|\pi_{\vec{a}/10}\rangle \geqslant \langle 0^n|\rho|0^n\rangle + \frac{\|\vec{a}\|_2^2}{20} \ .$$

Proof. Consider a purification $|\psi\rangle$ of ρ in a form similar to Theorem 4.4:

$$|\psi\rangle = \left(\alpha\,|0^n\rangle + \sum_{i=1}^{n} \frac{\langle e_i|\rho|0^n\rangle}{\alpha}\,|e_i\rangle + \beta\,|v_{\geqslant 2}\rangle\right)|g\rangle + \gamma\,|\perp\rangle \ .$$

We assume without loss of generality that

- $\alpha = \sqrt{\langle 0^n|\rho|0^n\rangle}$ and β and γ are nonnegative reals,
- $|v_{\geqslant 2}\rangle$, $|g\rangle$, and $|\perp\rangle$ are unit vectors,
- $|v_{\geqslant 2}\rangle$ is supported only on strings of Hamming weight 2 or larger, and
- $|\perp\rangle$ is orthogonal to all states beginning in $|0^n\rangle$ or ending in $|g\rangle$.

Let us express $|\pi_{\vec{a}/10}\rangle$ in the form:

$$|\pi_{\vec{a}/10}\rangle = \sqrt{p_0}\,|0^n\rangle + \sqrt{p_1}\,|w_1\rangle + \sqrt{p_{\geqslant 2}}\,|w_{\geqslant 2}\rangle\ ,$$

where p_0, p_1, and $p_{\geqslant 2}$ are probabilities summing to 1, $|w_1\rangle$ is supported over strings of weight 1, $|w_{\geqslant 2}\rangle$ is supported over strings of weight at least 2, and both $|w_1\rangle$ and $|w_{\geqslant 2}\rangle$ are unit vectors. We note that:

$$\langle \pi_{\vec{a}/10}|e_i\rangle = \frac{a_i^*/10}{\prod_{j=1}^n \sqrt{1+|a_i/10|^2}} = \frac{a_i^*\sqrt{p_0}}{10}\,.$$

So, the fidelity is equal to:

$$\begin{aligned}
\langle \pi_{\vec{a}/10}|\rho|\pi_{\vec{a}/10}\rangle &= \|(\langle \pi_{\vec{a}/10}| \otimes I)\,|\psi\rangle\|_2^2 \\
&= \left\| \left(\alpha\sqrt{p_0} + \sum_{i=1}^n \frac{a_i^*\sqrt{p_0}}{10}\frac{\langle e_i|\rho|0^n\rangle}{\alpha} + \beta\sqrt{p_{\geqslant 2}}\,\langle w_{\geqslant 2}|v_{\geqslant 2}\rangle \right)|g\rangle + \gamma(\langle \pi| \otimes I)\,|\perp\rangle \right\|_2^2 \\
&= \left\| \left(\alpha\sqrt{p_0} + \frac{\sqrt{p_0}\,\langle \vec{a},\vec{z}\rangle}{10\alpha} + \beta\sqrt{p_{\geqslant 2}}\,\langle w_{\geqslant 2}|v_{\geqslant 2}\rangle \right)|g\rangle + \gamma(\langle \pi| \otimes I)\,|\perp\rangle \right\|_2^2 \\
&= \left\| \left(\alpha\sqrt{p_0} + \frac{\sqrt{p_0}\,\langle \vec{a},\vec{z}\rangle}{10\alpha} + \beta\sqrt{p_{\geqslant 2}}\,\langle w_{\geqslant 2}|v_{\geqslant 2}\rangle \right)|g\rangle \right\|_2^2 + \|\gamma(\langle \pi| \otimes I)\,|\perp\rangle\|_2^2
\end{aligned}$$

by the Pythagorean theorem, because $|\perp\rangle$ has no support on $|g\rangle$. We can lower bound this by:

$$\begin{aligned}
\langle \pi_{\vec{a}/10}|\rho|\pi_{\vec{a}/10}\rangle &\geqslant \left| \alpha\sqrt{p_0} + \frac{\sqrt{p_0}\,\langle \vec{a},\vec{z}\rangle}{10\alpha} + \beta\sqrt{p_{\geqslant 2}}\,\langle w_{\geqslant 2}|v_{\geqslant 2}\rangle \right|^2 \\
&= \left| \alpha\sqrt{p_0} + \frac{\sqrt{p_0}\big(\|\vec{a}\|_2^2 - \langle \vec{a},\vec{a}-\vec{z}\rangle\big)}{10\alpha} + \beta\sqrt{p_{\geqslant 2}}\,\langle w_{\geqslant 2}|v_{\geqslant 2}\rangle \right|^2 \\
&\geqslant \left(\alpha\sqrt{p_0} + \frac{\sqrt{p_0}\big(\|\vec{a}\|_2^2 - |\langle \vec{a},\vec{a}-\vec{z}\rangle|\big)}{10\alpha} - \beta\sqrt{p_{\geqslant 2}} \right)^2 && \text{(Triangle inequality)} \\
&\geqslant \left(\alpha\sqrt{p_0} + \frac{\sqrt{p_0}\big(\|\vec{a}\|_2^2 - \|\vec{a}\|_2\|\vec{a}-\vec{z}\|_2\big)}{10\alpha} - \beta\sqrt{p_{\geqslant 2}} \right)^2 && \text{(Cauchy-Schwarz)} \\
&\geqslant \left(\alpha\sqrt{p_0} + \frac{\sqrt{p_0}\|\vec{a}\|_2^2}{20\alpha} - \beta\sqrt{p_{\geqslant 2}} \right)^2 && (\|\vec{a}-\vec{z}\|_2 \leqslant \|\vec{a}\|_2/2) \\
&\geqslant \left(\alpha\sqrt{p_0} + \frac{\sqrt{p_0}\|\vec{a}\|_2^2}{20\alpha} - \beta\sqrt{2}(1-\sqrt{p_0}) \right)^2 && \text{(Corollary 4.3)} \\
&\geqslant \left(\alpha p_0 + \frac{p_0\|\vec{a}\|_2^2}{20\alpha} - \beta\sqrt{2}(1-p_0) \right)^2. && (p_0 \leqslant 1)
\end{aligned}$$

Let us now obtain some bounds on p_0. From Lemma 3.7, we know that

$$p_0 = \prod_{i=1}^n \frac{1}{1+|a_i/10|^2} \geqslant e^{-\mathrm{d}_{\tan}(|\pi_{\vec{a}/10}\rangle,|0^n\rangle)^2} = e^{-\|\vec{a}/10\|_2^2} \geqslant 1 - \frac{\|\vec{a}\|_2^2}{100}\,.$$

By Claim 4.8, because $\|\vec{z}\|_2 \leqslant \frac{1}{2}$, we know that

$$\|\vec{a}\|_2 \leqslant \|\vec{z}\|_2 + \|\vec{a}-\vec{z}\|_2 \leqslant \frac{1+\|\vec{a}\|_2}{2}\,,$$

and therefore $\|\vec{a}\|_2 \leqslant 1$. This further implies $p_0 \geqslant 0.99$. Using these two bounds on p_0, we obtain the desired lower bound:

$$\begin{aligned}
\langle \pi_{\vec{a}/10} | \rho | \pi_{\vec{a}/10} \rangle &\geqslant \left(\alpha \left(1 - \frac{\|\vec{a}\|_2^2}{100} \right) + \frac{0.99 \|\vec{a}\|_2^2}{20\alpha} - \beta\sqrt{2} \frac{\|\vec{a}\|_2^2}{100} \right)^2 \\
&\geqslant \left(\alpha + \frac{0.0495}{\alpha} \|\vec{a}\|_2^2 - \frac{1+\sqrt{2}}{100} \|\vec{a}\|_2^2 \right)^2 && (\alpha, \beta \leqslant 1) \\
&\geqslant \left(\alpha + \frac{0.0395 - \sqrt{2}/100}{\alpha} \|\vec{a}\|_2^2 \right)^2 && (\alpha \leqslant 1) \\
&\geqslant \left(\alpha + \frac{1}{40\alpha} \|\vec{a}\|_2^2 \right)^2 \\
&\geqslant \alpha^2 + \frac{\|\vec{a}\|_2^2}{20} \\
&= \langle 0^n | \rho | 0^n \rangle + \frac{\|\vec{a}\|_2^2}{20} .
\end{aligned}$$

♦

4.4 Local optimization algorithms

Let's briefly take a step back to show the power of the combination of Corollary 4.7 and Theorem 4.9. Say that (in some appropriately chosen basis), we've found that

$$\langle 0^n | \rho | 0^n \rangle = 2/3 + c .$$

Suppose that the largest fidelity achievable with ρ by any product state is $2/3 + D$. Then Corollary 4.7 shows that

$$2/3 + D \leqslant 2/3 + c + \min\left\{ 3\|\vec{z}\|_2, \frac{\|\vec{z}\|_2^2}{c} \right\},$$

or equivalently

$$\|\vec{z}\|_2^2 \geqslant \max\left\{ \frac{(D-c)^2}{9}, c(D-c) \right\},$$

which further implies

$$\|\vec{z}\|_2^2 \geqslant \frac{D(D-c)}{10} .$$

To simplify things, imagine that we were able to learn $\vec{z}$ *exactly*. Then Theorem 4.9 shows that we can find a product state $|\pi\rangle$ whose fidelity with ρ is at least

$$\langle \pi | \rho | \pi \rangle \geqslant 2/3 + c + \frac{\|z\|_2^2}{20} + .$$

Combining these two shows that after a single local update according to Theorem 4.9, c increases at least as fast as

$$c \to c + \frac{D(D-c)}{200} .$$

If we make such local updates repeatedly, the fidelity of our product state with ρ converges towards $2/3 + D$ exponentially quickly. This is the high-level idea behind our local optimization procedure, Algorithm 4.11 below.

We first take a small detour to show how to learn an approximation of $\vec{z}$ in a sample- and time-efficient manner, via a simple modification of the classical shadows protocol [HKP20].

Lemma 4.10. *For an n-qubit density matrix ρ, define $\vec{z} \in \mathbb{C}^n$ by*

$$z_i = \langle e_i|\rho|0^n\rangle \ .$$

Then there is a procedure to find $\vec{a}$ satisfying $\|\vec{a} - \vec{z}\|_2 \leqslant \varepsilon$ with probability $1 - \delta$ that uses $O(\frac{n}{\varepsilon^2} \log \frac{1}{\delta})$ copies of ρ and $O\left(\frac{n^2 \log n}{\varepsilon^2} \log \frac{1}{\delta} + n \log^2 \frac{1}{\delta}\right)$ time.

Proof. Note that the real part of $\vec{z}$ is

$$\mathrm{Re}(z_i) = \mathrm{tr}\left(\frac{|0^n\rangle\langle e_i| + |e_i\rangle\langle 0^n|}{2}\rho\right),$$

and the imaginary part is

$$\mathrm{Im}(z_i) = \mathrm{tr}\left(\rho\frac{i\,|0^n\rangle\langle e_i| - i\,|e_i\rangle\langle 0^n|}{2}\rho\right).$$

So, it suffices to obtain an ℓ^2-error estimate of the $2n$ observables

$$\left\{\frac{|0^n\rangle\langle e_i| + |e_i\rangle\langle 0^n|}{2}, \frac{i\,|0^n\rangle\langle e_i| - i\,|e_i\rangle\langle 0^n|}{2}\right\}_{i\in[n]},$$

which we will call $O_1, \ldots, O_{2n}$.

For some N and K that we choose later, we will use NK copies of ρ to produce estimates $\hat{o}_i(N, K)$ of each $o_i = \mathrm{tr}(O_i\rho)$. Our success criterion will then be

$$\sum_{i=1}^{2n}(o_i - \hat{o}_i(N, K))^2 \leqslant \varepsilon^2 .$$

To do so, we use the classical shadows framework of [HKP20]. Consider measuring ρ in a random Clifford basis U, obtaining outcome $|\hat{b}\rangle$. Following [HKP20, Eqs. (S16) and (S5)], we define

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

to be the classical shadow, and

$$\hat{o}_i = \mathrm{tr}(O_i\hat{\rho})$$

the estimator corresponding to O_i. The key fact shown in [HKP20, Lemma 1 and Eq. (S16)] is

$$\mathbf{E}[\hat{o}_i] = o_i \qquad \text{and} \qquad \mathbf{Var}[\hat{o}_i] \leqslant 3\,\mathrm{tr}(O_i)^2 = \frac{3}{2}.$$

So, the expected sum of squared deviations is at most:

$$\mathbb{E}\left[\sum_{i=1}^{2n}(o_i - \hat{o}_i)^2\right] = \sum_{i=1}^{2n}\mathbb{E}\left[(o_i - \hat{o}_i)^2\right] \leqslant 3n\,.$$

If we take N classical shadows $\hat{\rho}_1, \ldots, \hat{\rho}_N$ and compute the mean of their estimators, as in [HKP20, Eq. (S11)]:

$$\hat{o}_i(N, 1) := \frac{1}{N}\sum_{j=1}^{N}\mathrm{tr}(O_i\hat{\rho}_j)\,,$$

the expected sum of squared deviations is decreased by a factor of N:

$$\mathbb{E}\left[\sum_{i=1}^{2n}(o_i - \hat{o}_i(N, 1))^2\right] \leqslant \frac{3n}{N}\,.$$

Choosing $N = \frac{n}{27\varepsilon^2}$, by Markov's inequality we will have

$$\mathbf{Pr}\left[\sum_{i=1}^{2n}(o_i - \hat{o}_i(N,1))^2 \leqslant \frac{\varepsilon^2}{3}\right] \geqslant \frac{2}{3}.$$

To boost the success probability from $2/3$ to $1-\delta$, we can use a median-of-means trick by letting $\vec{\delta}(N,K)$ be the "median" of $K = O(\log \frac{1}{\delta})$ independent samples from $\vec{\delta}(N,1)$; see for example [HKOT23, Proposition 2.4]. (Note: it is important that the "median" in this case is performed with respect to the entire vectors $\vec{\delta}(N,1)$ rather than entrywise on $\hat{o}_i(N,1)$; see [HKOT23] for details. The important feature is that the measure of error, ℓ^2 distance, is a metric.)

It remains to bound the runtime. Naively, operating on the classical shadows takes exponential time. However, because all of the observables are supported only the subspace of Hamming weight 0 or 1, we can encode the information about this subspace into a register of just $O(\log n)$ qubits and perform the classical shadows there. For each sample of ρ, we append a register of $1 + \lceil \log_2(n+1) \rceil$ qubits initialized to $|0\rangle$. We coherently set the first appended qubit to $|1\rangle$ conditioned on having Hamming weight 0 or 1 in the n-qubit register. Then, conditioned on having Hamming weight 0 or 1, we populate the remainder of the register with the binary representation of the qubit that was set to $|1\rangle$, or zero if no such qubit exists. Now we can take the classical shadows entirely within this $O(\log n)$-qubit register.

Mapping ρ into this $O(\log n)$-qubit register takes $O(n \log n)$ time per classical shadow. Next, sampling the random Clifford U and then measuring to get $|\hat{b}\rangle$ takes $O(\log^2 n)$ time per classical shadow [Van21]. We then compute the entire amplitude vector of $U^\dagger |\hat{b}\rangle$ in $\mathbb{C}^{O(n)}$, which takes time $O(n \log n)$ per classical shadow [SSY23, Algorithm 2]. Since the O_i's are sparse, computing each $\hat{o}_i = \mathrm{tr}(O_i \hat{\rho})$ takes $O(1)$ time by using $U^\dagger |\hat{b}\rangle$ as a lookup table. Finally, as noted in [HKOT23, Proposition 2.4], the "median-of-means procedure" takes $O(K^2)$ times the cost of computing the distance between two vectors $\vec{\delta}(N,1)$, for a total of $O(K^2 n)$ time.[4] The overall runtime is therefore:

$$O(NK(n \log n) + K^2 n) = O\left(\frac{n^2 \log n}{\varepsilon^2} \log \frac{1}{\delta} + n \log^2 \frac{1}{\delta}\right). \qquad \blacklozenge$$

We remark that, instead of the lemma above, one could simply appeal to [HKP20, Theorem 1] as a black box to estimate each z_i to additive error $\sqrt{\frac{\varepsilon}{n}}$. However, our approach saves a $\log(n)$ factor in the sample complexity because we only need to approximate $\vec{z}$ in ℓ^2 distance, rather than ℓ^∞.

We can now state our algorithm for optimizing toward a global maximizer of product state fidelity.

[4] $O(K^2 n)$ is a worst-case bound, but the average-case time complexity is easily seen to be $O(Kn)$. We believe that a more careful accounting of this factor over the iterations of Algorithm 4.11 could remove the $\log^2 \frac{1}{\delta}$ factors that appear in the runtime bounds later in this section.

Algorithm 4.11. (Local product state optimization).

Input: Copies of an n-qubit state ρ, an n-qubit product state $|\pi\rangle$, $\varepsilon \in (0, 1/3]$, $\delta \in (0,1)$, $C \in [0, 1/3]$

Promise: $\langle\pi|\rho|\pi\rangle \geqslant 2/3$, and $\max_{\text{product } |\pi'\rangle} \langle\pi'|\rho|\pi'\rangle \geqslant 2/3 + C$

Output: A product state $|\pi\rangle$ satisfying $\langle\pi|\rho|\pi\rangle \geqslant \max_{\text{product } |\pi'\rangle} \langle\pi'|\rho|\pi'\rangle - \varepsilon$ with probability at least $1 - \delta$

Procedure:

1: $C' := \max\{\varepsilon, C\}$
2: $m := \left\lceil \frac{1}{2} \log\left(\frac{90}{C'\varepsilon}\right) \right\rceil$
3: **loop**
4: Find a product unitary U such that $U|\pi\rangle = |0^n\rangle$
5: Define $\vec{z}$ by $z_i = \langle e_i | U\rho U^\dagger | 0^n\rangle$ for $i \in [n]$
6: **for** $\lambda = 1, \ldots, m$ **do**
7: $\ell_\lambda := m + 1 - \lambda$
8: $\delta_\lambda := \delta 2^{-\ell_\lambda} \left(\left\lceil 5\varepsilon e^{2m} \right\rceil + \left\lceil \frac{900\ell_\lambda}{C'} \right\rceil \right)^{-1}$
9: Use Lemma 4.10 to find $\vec{a}$ satisfying $\|\vec{a} - \vec{z}\|_2 \leqslant e^{-\lambda}$ with prob. $1 - \delta_\lambda$
10: **if** $\|\vec{a}\|_2 \geqslant 2e^{-\lambda}$ **then**
11: $|\pi\rangle := U^\dagger |\pi_{\vec{a}/10}\rangle$
12: Exit for-loop
13: **else if** $\lambda = m$ **then**
14: **return** $|\pi\rangle$

We note that the parameter C is effectively optional: one can always set $C = 0$ and the algorithm will be correct. However, the algorithm becomes more efficient when C is larger.

We will show the correctness of Algorithm 4.11 in a series of smaller steps. First we lower bound the improvement from updates:

Claim 4.12. In a given non-terminating iteration of the outer loop of Algorithm 4.11, if Line 9 does not err, then $\langle\pi|\rho|\pi\rangle$ increases by at least $\frac{\|\vec{a}\|_2^2}{20}$.

Proof. $|\pi\rangle$ only changes in Line 11, so consider an iteration in which the algorithm reaches Line 11, and therefore $\|\vec{a}\|_2 \geqslant 2e^{-\lambda} \geqslant 2\|\vec{a} - \vec{z}\|_2$ for some λ. Then we may appeal to Theorem 4.9 to conclude that

$$\begin{aligned} \langle\pi_{\vec{a}/10}|U\rho U^\dagger|\pi_{\vec{a}/10}\rangle - \langle\pi|\rho|\pi\rangle &= \langle\pi_{\vec{a}/10}|U\rho U^\dagger|\pi_{\vec{a}/10}\rangle - \langle 0^n|U\rho U^\dagger|0^n\rangle \\ &\geqslant \frac{\|\vec{a}\|_2^2}{20}. \end{aligned}$$

♦

The helper lemma below will be used to show that the product state fidelity converges to within ε of the optimum in roughly $O\left(\log \frac{1}{\varepsilon}\right)$ iterations of the outer loop.

Lemma 4.13. *Let $\{x_i\}_{i \in \mathbb{N}}$ be a sequence satisfying $x_0 \geqslant c \geqslant 0$ and*

$$x_{i+1} \geqslant \min\left\{ x_i + \frac{D(D - x_i)}{r}, D - \varepsilon \right\}$$

for some $r > D$. If we define

$$k := \frac{r}{D} \log\left(\frac{D-c}{\varepsilon}\right),$$

then for all $i \geqslant k$, $x_i \geqslant D - \varepsilon$.

Proof. We assume without loss of generality that $c \leqslant D - \varepsilon$, because otherwise k is negative and the statement clearly holds.

We first note that $x_i \geqslant t$ implies

$$x_{i+1} \geqslant \min\left\{t + \frac{D(D-t)}{r}, D - \varepsilon\right\},$$

because $f(t) = t + \frac{D(D-t)}{r}$ is increasing on $[0, D]$. So, if we define the sequence $\{y_i\}_{i\in\mathbb{N}}$ by $y_0 = c$ and

$$y_{i+1} = y_i + \frac{D(D-y_i)}{r},$$

then $x_i \geqslant \min\{y_i, D - \varepsilon\}$, and thus it suffices to show that $y_i \geqslant D - \varepsilon$ for all $i \geqslant k$.

We can write the definition of y equivalently as

$$D - y_{i+1} = D - y_i - \frac{D(D-y_i)}{r} = \left(1 - \frac{D}{r}\right)(D - y_i).$$

In other words, $D - y_i$ decays exponentially as

$$D - y_i = \left(1 - \frac{D}{r}\right)^i (D - c).$$

Thus choosing

$$k' = \frac{\log\left(\frac{\varepsilon}{D-c}\right)}{\log\left(1 - \frac{D}{r}\right)}$$

guarantees that $y_i \geqslant D - \varepsilon$ for all $i \geqslant k'$. Because

$$\begin{aligned} k' &= -\frac{\log\left(\frac{D-c}{\varepsilon}\right)}{\log\left(1 - \frac{D}{r}\right)} \\ &\leqslant \frac{r}{D} \log\left(\frac{D-c}{\varepsilon}\right) && \left(\log(1+x) \leqslant x, \log\left(\frac{D-c}{\varepsilon} \geqslant 0\right)\right) \\ &\leqslant k, \end{aligned}$$

the lemma follows. ♦

Lemma 4.14. *Suppose that*

$$\max_{\text{product } |\pi'\rangle} \langle\pi'|\rho|\pi'\rangle = 2/3 + D,$$

and suppose $\langle\pi|\rho|\pi\rangle \geqslant 2/3 + c \geqslant 2/3$ in some iteration of Line 9's outer loop. Then conditioned on Line 9 never erring thereafter, Algorithm 4.11 returns a $|\pi\rangle$ satisfying the output condition (i.e., $\langle\pi|\rho|\pi\rangle \geqslant 2/3 + D - \varepsilon$) within at most

$$\lceil 5\varepsilon e^{2m} \rceil + \max\left\{0, \left\lceil \frac{450}{D} \log\left(\frac{D-c}{\varepsilon}\right) \right\rceil\right\}$$

additional iterations of the outer loop.

Proof. Claim 4.12 shows that in each non-terminating iteration, $\langle\pi|\rho|\pi\rangle$ increases by at least $\frac{\|\vec{a}\|_2^2}{20} \geqslant \frac{e^{-2m}}{5}$. So, if $c \geqslant D - \varepsilon$, then the algorithm must halt within $\lceil 5\varepsilon e^{2m} \rceil$ additional iterations (as the fidelity can never exceed $2/3 + D$). Conversely, if $c < D - \varepsilon$, then it suffices to show that the algorithm achieves $\langle\pi|\rho|\pi\rangle \geqslant 2/3 + D - \varepsilon$ within the initial $\lceil \frac{450}{D} \log\left(\frac{D-c}{\varepsilon}\right) \rceil$ iterations of the outer loop, because thereafter the algorithm must halt within $\lceil 5\varepsilon e^{2m} \rceil$ additional iterations. So, we assume henceforth that $c < D - \varepsilon$.

For $i \in \mathbb{N}$, define x_i so that $\langle\pi|\rho|\pi\rangle = 2/3 + x_i$ immediately after the end of i additional iterations of the outer loop, with the convention that $\langle\pi|\rho|\pi\rangle = 2/3 + x_0 \geqslant 2/3 + c$ at the start of the algorithm.

Suppose that $x_i \leqslant D - \varepsilon$, and consider what happens in iteration $i + 1$. We know that $x_i \geqslant c \geqslant 0$ by Claim 4.12. Then Corollary 4.7 tells us that

$$2/3 + D \leqslant 2/3 + x_i + \min\left\{3\|\vec{z}\|_2, \frac{\|\vec{z}\|_2^2}{x_i}\right\},$$

or equivalently

$$\|\vec{z}\|_2^2 \geqslant \max\left\{\frac{(D - x_i)^2}{9}, x(D - x_i)\right\},$$

which then gives

$$\|\vec{z}\|_2^2 \geqslant \frac{D(D - x_i)}{10}. \tag{6}$$

We claim that within this $(i+1)$th iteration of the outer loop, the algorithm finds $\vec{a}$ satisfying $\|\vec{a}\|_2 \geqslant e^{1-\lambda} \geqslant 2\|\vec{a} - \vec{z}\|_2$ for *some* $\lambda \leqslant m$, as otherwise upon reaching Line 14 we would have

$$\begin{aligned}
\sqrt{\frac{C'\varepsilon}{10}} &\leqslant \sqrt{\frac{D\varepsilon}{10}} && (D \geqslant C \text{ and } D \geqslant x_i + \varepsilon \geqslant \varepsilon \text{ by supposition}) \\
&\leqslant \sqrt{\frac{D(D - x_i)}{10}} && (D \geqslant x_i + \varepsilon \text{ by supposition})) \\
&\leqslant \|\vec{z}\|_2 && (\text{Equation (6)}) \\
&\leqslant \|\vec{a}\|_2 + \|\vec{a} - \vec{z}\|_2 && (\text{Triangle inequality}) \\
&< 2e^{-m} + e^{-m} \\
&\leqslant \sqrt{\frac{C'\varepsilon}{10}},
\end{aligned}$$

a contradiction.

We know that $\|\vec{a}\|_2 \geqslant \|\vec{z}\|_2 - \|\vec{a} - \vec{z}\|_2$ by the triangle inequality, and since $\|\vec{a} - \vec{z}\|_2 \leqslant \|\vec{a}\|_2/2$, we get that

$$\begin{aligned}
\|\vec{a}\|_2 &\geqslant \frac{2}{3}\|\vec{z}\|_2 \\
&\geqslant \frac{2}{3}\sqrt{\frac{D(D - x_i)}{10}}. && (\text{Equation (6)})
\end{aligned}$$

Plugging into Corollary 4.7,

$$\begin{aligned}
\langle\pi_{\vec{a}/10}|U\rho U^\dagger|\pi_{\vec{a}/10}\rangle - \langle\pi|\rho|\pi\rangle &= \langle\pi_{\vec{a}/10}|U\rho U^\dagger|\pi_{\vec{a}/10}\rangle - \langle 0^n|U\rho U^\dagger|0^n\rangle \\
&\geqslant \frac{\|\vec{a}\|_2^2}{20} \\
&\geqslant \frac{D(D - x_i)}{450}.
\end{aligned}$$

We have effectively established that

$$x_{i+1} \geqslant \min\left\{x_i + \frac{D(D-x_i)}{450}, D-\varepsilon\right\}.$$

Lemma 4.13 shows that for all $i \geqslant k$, $x_i \geqslant D-\varepsilon$, where

$$k = \frac{450}{D}\log\left(\frac{D-c}{\varepsilon}\right),$$

which proves the lemma. ♦

Lemma 4.14 is almost sufficient to establish Algorithm 4.11's correctness; we only need to show that the total error probability is at most δ. For that, we bound the total number of calls to the tomography subroutine (Line 9).

Lemma 4.15. *Consider a run of Algorithm 4.11 where Line 9 never errs. Then for a given λ, Line 9 is executed at most*

$$\left\lceil 5\varepsilon e^{2m}\right\rceil + \left\lceil \frac{900\ell_\lambda}{C'}\right\rceil$$

times before Algorithm 4.11 halts.

Proof. Define $D := \max_{\text{product } |\pi'\rangle} \langle\pi'|\rho|\pi'\rangle - 2/3$ as in Lemma 4.14. We break into cases depending on D and λ.

Case 1: $D < \varepsilon$. Then the total number of calls to Line 9 is bounded by the number of iterations of the outer loop, which by Lemma 4.14 is at most

$$\left\lceil 5\varepsilon e^{2m}\right\rceil + \max\left\{0, \left\lceil \frac{450}{D}\log\left(\frac{D}{\varepsilon}\right)\right\rceil\right\} = \left\lceil 5\varepsilon e^{2m}\right\rceil.$$

Case 2: $D \geqslant \varepsilon$, $\lambda = 1$. Again we bound the number of calls to Line 9 by the number of iterations of the outer loop, using Lemma 4.14 and that $1/3 \geqslant D \geqslant \max\{C, \varepsilon\} = C'$:

$$\begin{aligned}\left\lceil 5\varepsilon e^{2m}\right\rceil + \max\left\{0, \left\lceil \frac{450}{D}\log\left(\frac{D}{\varepsilon}\right)\right\rceil\right\} &\leqslant \left\lceil 5\varepsilon e^{2m}\right\rceil + \left\lceil \frac{450}{C'}\log\left(\frac{1}{3\varepsilon}\right)\right\rceil \\ &\leqslant \left\lceil 5\varepsilon e^{2m}\right\rceil + \left\lceil \frac{900m}{C'}\right\rceil \\ &= \left\lceil 5\varepsilon e^{2m}\right\rceil + \left\lceil \frac{900\ell_\lambda}{C'}\right\rceil.\end{aligned}$$

Case 3: $D \geqslant \varepsilon$, $\lambda \geqslant 2$. Consider an iteration of the algorithm in which $\langle\pi|\rho|\pi\rangle = 2/3 + x$. In order for the for-loop (Line 6) to reach iteration λ, by the triangle inequality we must have

$$\begin{aligned}\|\vec{z}\|_2 &\leqslant \|\vec{a}\|_2 + \|\vec{a}-\vec{z}\|_2 \\ &\leqslant 2e^{-(\lambda-1)} + e^{-(\lambda-1)} \\ &= 3e^{\ell_\lambda - m} \\ &\leqslant \sqrt{\frac{C'\varepsilon}{10}}e^{\ell_\lambda}.\end{aligned}$$

We know from Corollary 4.7 that

$$2/3 + D \leqslant 2/3 + x + \min\left\{3\|\vec{z}\|_2, \frac{\|\vec{z}\|_2^2}{x}\right\},$$

or equivalently

$$\|\vec{z}\|_2^2 \geqslant \max\left\{\frac{(D-x)^2}{9}, x(D-x)\right\},$$

which then gives

$$\|\vec{z}\|_2^2 \geqslant \frac{D(D-x)}{10}.$$

Combining, we find that

$$\frac{D-x}{\varepsilon} \leqslant \frac{C'}{D} e^{2\ell_\lambda} \leqslant e^{2\ell_\lambda},$$

because $D \geqslant \max\{\varepsilon, C\} = C'$.

We have shown that Line 9 is executed for this particular λ only when $x \geqslant D - \varepsilon e^{2\ell_\lambda}$. Since x never drops below 0 (Claim 4.12), we can let $c := \max\{0, D - \varepsilon e^{2\ell_\lambda}\}$ and appeal to Lemma 4.14 to bound the number of additional iterations for which the algorithm can run by

$$\lceil 5\varepsilon e^{2m} \rceil + \max\left\{0, \left\lceil \frac{450}{D} \log\left(\frac{D-c}{\varepsilon}\right) \right\rceil\right\}.$$

The lemma follows by substituting $c \geqslant D - \varepsilon e^{2\ell_\lambda}$ and $C' \leqslant D$. ♦

Corollary 4.16. *The total failure probability of Algorithm 4.11 is at most δ.*

Proof. By a union bound, we can use Lemma 4.15 to bound the contribution of each run of Line 9 for a given λ:

$$\begin{aligned}
\sum_{\lambda=1}^{m} \delta_\lambda \left(\lceil 5\varepsilon e^{2m} \rceil + \left\lceil \frac{900\ell_\lambda}{C'} \right\rceil \right) &= \delta \sum_{\lambda=1}^{m} 2^{-\ell_\lambda} \\
&= \delta \sum_{\ell=1}^{m} 2^{-\ell} \qquad (\ell := m+1-\lambda) \\
&< \delta \sum_{\ell=1}^{\infty} 2^{-\ell} \\
&= \delta .
\end{aligned}$$

♦

Corollary 4.17. *Assuming Line 9 never errs, the total sample complexity Algorithm 4.11 is at most*

$$O\left(\frac{n}{\varepsilon C'^2} \log \frac{1}{\delta C'}\right)$$

and the runtime is

$$O\left(\frac{n^2 \log n}{\varepsilon C'^2} \log \frac{1}{\delta C'} + \frac{n}{C'} \log^2 \frac{1}{\varepsilon C'} \log^2 \frac{1}{\delta C'}\right),$$

recalling that $C' = \max\{\varepsilon, C\}$.

Proof. For a given λ, a single run of Line 9 has sample complexity $O\left(ne^{2\lambda} \log \frac{1}{\delta_\lambda}\right)$ according to Lemma 4.10. A simple calculation shows that

$$\log \frac{1}{\delta_\lambda} \leqslant O\left(\ell_\lambda + \log \frac{\ell_\lambda}{\delta C'}\right) \leqslant O\left(\ell_\lambda + \log \frac{1}{\delta C'}\right).$$

Using Lemma 4.15, Line 9 is called a total of

$$\lceil 5\varepsilon e^{2m} \rceil + \left\lceil \frac{900\ell_\lambda}{C'} \right\rceil \leqslant O\left(\frac{\ell_\lambda}{C'}\right)$$

times. The total sample complexity is therefore

$$\begin{aligned}
\sum_{\lambda=1}^{m} O\left(ne^{2\lambda}\log\left(\frac{1}{\delta_\lambda}\right)\frac{\ell_\lambda}{C'}\right) &\leqslant \sum_{\lambda=1}^{m} O\left(ne^{2\lambda}\left(\ell_\lambda+\log\frac{1}{\delta C'}\right)\frac{\ell_\lambda}{C'}\right)\\
&\leqslant \sum_{\ell=1}^{m} O\left(ne^{2m+2-2\ell}\left(\ell+\log\frac{1}{\delta C'}\right)\frac{\ell}{C'}\right) \qquad (\ell := m+1-\lambda)\\
&\leqslant \frac{n}{\varepsilon C'^2}\sum_{\ell=1}^{m} O\left(e^{-2\ell}\left(\ell+\log\frac{1}{\delta C'}\right)\ell\right)\\
&\leqslant O\left(\frac{n}{\varepsilon C'^2}\log\frac{1}{\delta C'}\right). \qquad (7)
\end{aligned}$$

We turn to the time complexity. By Lemma 4.10, the runtime of a single call to Line 9 with chosen λ is

$$(\text{Sample complexity})\cdot O(n\log n)+O\left(n\log^2\frac{1}{\delta_\lambda}\right).$$

Since the $O(n\log n)$ is independent of λ, it is easy to bound the contribution of the left term to the runtime: we multiply the total sample complexity (Equation (7)) by $O(n\log n)$, yielding $O\left(\frac{n^2\log n}{\varepsilon C'^2}\log\frac{1}{\delta C'}\right)$. So, we focus on bounding the contribution of the right term, which is

$$\begin{aligned}
\sum_{\lambda=1}^{m} O\left(n\log^2\left(\frac{1}{\delta_\lambda}\right)\frac{\ell_\lambda}{C'}\right) &\leqslant \sum_{\lambda=1}^{m} O\left(n\left(\ell_\lambda+\log\frac{1}{\delta C'}\right)^2\frac{\ell_\lambda}{C'}\right)\\
&\leqslant \sum_{\ell=1}^{m} O\left(n\left(\ell+\log\frac{1}{\delta C'}\right)^2\frac{\ell}{C'}\right) \qquad (\ell := m+1-\lambda)\\
&\leqslant \frac{n}{C'}\sum_{\ell=1}^{m} O\left(\left(\ell+\log\frac{1}{\delta C'}\right)^2\ell\right)\\
&\leqslant \frac{n}{C'}\sum_{\ell=1}^{m} O\left(\ell^3+\ell\log^2\frac{1}{\delta C'}\right) \qquad ((a+b)^2\leqslant O(a^2+b^2))\\
&\leqslant \frac{n}{C'}O\left(m^4+m^2\log^2\frac{1}{\delta C'}\right)\\
&\leqslant O\left(\frac{n}{\varepsilon C'^2}\right)+O\left(\frac{n}{C'}\log^2\frac{1}{\varepsilon C'}\log^2\frac{1}{\delta C'}\right). \quad \left(m=O\left(\log\frac{1}{\varepsilon C'}\right)\right) \qquad (8)
\end{aligned}$$

The left part of Equation (8) gets absorbed into the $O\left(\frac{n^2\log n}{\varepsilon C'^2}\log\frac{1}{\delta C'}\right)$. ♦

4.5 Divide and conquer

As written, Algorithm 4.11 assumes that we begin with a product state $|\pi\rangle$ having fidelity at least $2/3$ with ρ. This is not a true learning algorithm, then, because such a state $|\pi\rangle$ might not be known in advance. Nevertheless, we can straightforwardly generalize Algorithm 4.11 to a learning algorithm that only takes copies of ρ as input, using a divide-and-conquer approach. This works as a consequence of the following lemma:

Lemma 4.18. *Let ρ_{AB} be a state on a systems A and B. Suppose that $\langle\phi|\rho_A|\phi\rangle \geqslant 1-\varepsilon_1$ and $\langle\pi|\rho_B|\pi\rangle \geqslant 1-\varepsilon_2$. Then*

$$\langle\phi\pi|\rho_{AB}|\phi\pi\rangle \geqslant 1-\varepsilon_1-\varepsilon_2\,.$$

Proof. Extend $|\pi\rangle =: |\pi_1\rangle$ to an orthonormal basis $\{|\pi_i\rangle \mid i \in [d]\}$ for B. Since the partial trace can be computed by summing over any orthonormal basis, we have

$$\begin{aligned}
1 - \varepsilon_1 &\leqslant \langle \phi | \rho_A | \phi \rangle \\
&= \sum_{i=1}^{d} \langle \phi \pi_i | \rho_{AB} | \phi \pi_i \rangle \\
&\leqslant \langle \phi \pi | \rho_{AB} | \phi \pi \rangle + \sum_{i=2}^{d} \langle \pi_i | \rho_B | \pi_i \rangle \\
&= \langle \phi \pi | \rho_{AB} | \phi \pi \rangle + 1 - \langle \pi | \rho_B | \pi \rangle \\
&\leqslant \langle \phi \pi | \rho_{AB} | \phi \pi \rangle + \varepsilon_2 \,.
\end{aligned}$$

The lemma follows by rearranging. ♦

The full learning algorithm is below; its correctness is self-explanatory. Note that we must assume the existence of a product state with fidelity above $5/6$, instead of $2/3$ in the previous algorithm. This is because of the loss incurred from combining the two halves via Lemma 4.18.

Algorithm 4.19. (High-fidelity product state agnostic learning).

Input: Copies of an n-qubit state ρ, $\varepsilon \in (0, 1/6]$, $\delta \in (0, 1)$

Promise: There exists a product state $|\pi\rangle$ satisfying $\langle \pi | \rho | \pi \rangle \geqslant 5/6 + \varepsilon$

Output: A product state $|\pi\rangle$ satisfying $\langle \pi | \rho | \pi \rangle \geqslant \max_{\text{product } |\pi'\rangle} \langle \pi' | \rho | \pi' \rangle - \varepsilon$ with probability at least $1 - \delta$

Procedure:

▷ *First find a product state $|\pi\rangle$ such that $\langle \pi | \rho | \pi \rangle \geqslant 2/3$*
if $n = 1$ **then**
  Use tomography to find a $|\pi\rangle$ satisfying $\langle \pi | \rho | \pi \rangle \geqslant 2/3$ with prob. $1 - \delta/2$
else
  $\rho_L :=$ the left half of ρ
  $|\pi_L\rangle :=$ Algorithm 4.19$(\rho_L, \varepsilon, \delta/4)$ ▷ $\langle \pi_L | \rho_L | \pi_L \rangle \geqslant 5/6$
  $\rho_R :=$ the right half of ρ
  $|\pi_R\rangle :=$ Algorithm 4.19$(\rho_R, \varepsilon, \delta/4)$ ▷ $\langle \pi_R | \rho_R | \pi_R \rangle \geqslant 5/6$
  $|\pi\rangle := |\pi_L\rangle \otimes |\pi_R\rangle$ ▷ $\langle \pi | \rho | \pi \rangle \geqslant 2/3$ *by Lemma 4.18*
return Algorithm 4.11$(\rho, |\pi\rangle, \varepsilon, \delta/2, 1/3 + \varepsilon)$ ▷ *Error* $\leqslant \delta/4 + \delta/4 + \delta/2 = \delta$

Theorem 4.20. *Algorithm 4.19 has sample complexity*

$$O\left(\frac{n}{\varepsilon} \log \frac{1}{\delta}\right)$$

and runs in time

$$O\left(\frac{n^2 \log n}{\varepsilon} \log \frac{1}{\delta} + n \log n \log^2 \frac{1}{\varepsilon} \log^2 \frac{1}{\delta}\right) \leqslant O\left(\frac{n^2 \log n}{\varepsilon} \log^2 \frac{1}{\delta}\right).$$

Proof. For the sample complexity, observe that copies of ρ can be shared between the two recursive calls to Algorithm 4.19, because a copy of ρ is both a copy of ρ_L and a copy of ρ_R.

When $n = 1$, it is clear that Line 3 can be performed using $O\left(\log \frac{1}{\delta}\right)$ time and samples. For example, one can estimate the coordinates of ρ on the Bloch sphere to some small constant precision by measuring repeatedly in the X, Y, and Z bases. Therefore, the contribution of Line 3 to sample complexity and runtime across recursive calls to Algorithm 4.19 are, respectively, at most $O\left(\log \frac{n^2}{\delta}\right) \leqslant O\left(\log n \log \frac{1}{\delta}\right)$ and $O\left(n \log \frac{n^2}{\delta}\right) \leqslant O\left(n \log n \log \frac{1}{\delta}\right)$. These are dominated by the other terms.

Line 10 accounts for the remaining algorithmic complexity. By Corollary 4.17, because we pick $C > 1/3$, the overall sample complexity due to Algorithm 4.11 across recursive calls is at most

$$\sum_{i=0}^{\lceil \log_2 n \rceil} O\left(\frac{n2^{-i}}{\varepsilon} \log \frac{4^i}{\delta}\right) \leqslant \frac{n}{\varepsilon} \log \frac{1}{\delta} \sum_{i=0}^{\lceil \log_2 n \rceil} O\left(\frac{i}{2^i}\right) \leqslant O\left(\frac{n}{\varepsilon} \log \frac{1}{\delta}\right).$$

The time complexity bound is

$$\sum_{i=0}^{\lceil \log_2 n \rceil} 2^i \cdot O\left(\frac{\left(n2^{-i}\right)^2 \log(n2^{-i})}{\varepsilon} \log \frac{4^i}{\delta} + n2^{-i} \log^2 \frac{1}{\varepsilon} \log^2 \frac{4^i}{\delta}\right).$$

We handle the two terms within the summation separately. The left term is bounded by:

$$\frac{n^2 \log n}{\varepsilon} \log \frac{1}{\delta} \sum_{i=0}^{\lceil \log_2 n \rceil} O\left(\frac{i}{2^i}\right) \leqslant O\left(\frac{n^2 \log n}{\varepsilon} \log \frac{1}{\delta}\right). \tag{9}$$

The right term is bounded by:

$$\begin{aligned} n \log^2 \frac{1}{\varepsilon} \sum_{i=0}^{\lceil \log_2 n \rceil} O\left(\log^2 \frac{4^i}{\delta}\right) &\leqslant n \log^2 \frac{1}{\varepsilon} \sum_{i=0}^{\lceil \log_2 n \rceil} O\left(i^2 + \log^2 \frac{1}{\delta}\right) \qquad ((a+b)^2 \leqslant O(a^2+b^2)) \\ &\leqslant O\left(n \log^3 n \log^2 \frac{1}{\varepsilon} + n \log n \log^2 \frac{1}{\varepsilon} \log^2 \frac{1}{\delta}\right). \end{aligned}$$

The theorem follows because $n \log^3 n \log^2 \frac{1}{\varepsilon}$ is bounded by Equation (9). ♦

5 Agnostic learning of product states

In this section, we give our general algorithm for finding product states which have good fidelity with an input state ρ. Our output will take the form of a "good" product state cover, as given below.

Definition 5.1 (Good product state cover)**.** A collection of pure product states on m qubits, $\mathcal{C} = \{|\pi_i\rangle\}_i$ is a $(\eta, \varepsilon, b, B)$-*good cover for a state* ρ if

1. For all $|\pi\rangle \in \mathcal{C}$, $\langle\pi|\rho|\pi\rangle \geqslant \eta - \varepsilon$.
2. For all distinct $|\pi\rangle, |\pi'\rangle \in \mathcal{C}$, $\mathrm{d}_{\tan}(|\pi\rangle, |\pi'\rangle) \geqslant b$.
3. For all product states on m qubits, $|\phi\rangle$, such that $\langle\phi|\rho|\phi\rangle \geqslant \eta$, there exists $|\pi\rangle \in \mathcal{C}$ such that $\mathrm{d}_{\tan}(|\phi\rangle, |\pi\rangle) \leqslant B$.

We will eventually take $b = 2/\eta$ and $B = 3/\eta$, so for brevity a (η, ε)-good cover refers to a $(\eta, \varepsilon, 2/\eta, 3/\eta)$-good cover.[5]

[5]The properties our parameters must satisfy are that $\eta - \varepsilon - \frac{1}{b} > 0$ (for Claim 5.4) and that $B > b$ (for the greedy algorithm to succeed).

It may not yet be clear that a good product state cover for ρ even exists. When $B \geqslant b$, a greedy approach works here: start with $\mathcal{C} = \varnothing$, and while property 3 does not hold for $\mathcal{C}$, add the violating $|\phi\rangle$ to $\mathcal{C}$. This approach will be used and adapted to be more tractable in Algorithm 5.5. We also soon show that the size of good product state convers is small (Claim 5.4). First, we state the theorem we will prove in this section.

Theorem 5.2 (Agnostic learning of product states)**.** *Let ρ be an n-qubit state and suppose we are given error parameters $\eta \in (0,1)$, $\varepsilon \in (0, \eta/3)$, and $\delta \in (0,1)$. Then there is an algorithm which, with probability $\geqslant 1-\delta$, outputs an (η,ε)-good product state cover for ρ. The algorithm uses $N \leqslant (\mathrm{poly}(n))^{1/\eta^2 + \log\frac{1}{\varepsilon}} \mathrm{poly}(\log\frac{1}{\delta})$ copies of ρ, $\mathrm{poly}(n, 1/\eta, \log\frac{1}{\varepsilon})$ quantum gates per copy of ρ, and $n^{\mathrm{poly}(1/\varepsilon)} \mathrm{poly}(\log\frac{1}{\delta})$ additional classical overhead.*

Remark 5.3 (Applications of Theorem 5.2)**.** We use this remark to note some immediate corollaries of the above algorithm.

First, it can be used to deduce $\mathsf{OPT} = \max_{|\pi\rangle} \langle\pi|\rho|\pi\rangle$, the maximum fidelity a product state has with ρ, to a specified error 2ε. This is because the cover $\mathcal{C}$ output by the algorithm contains at least one product state with fidelity at least $\eta - \varepsilon$, if a product state with fidelity at least η exists. So, we can start with $\eta = 1/2$ and perform binary search on the choice of η, reducing η when the output cover is empty and increasing it when it is non-empty. After $O(\log(1/\varepsilon))$ iterations (and with an appropriate choice of δ), η will be an ε-good estimate. This also gives a product state $|\pi\rangle$ such that $\langle\pi|\rho|\pi\rangle \geqslant \mathsf{OPT} - 2\varepsilon$. The running time for this algorithm is also $n^{\mathrm{poly}(1/\varepsilon)} \mathrm{poly}(\log\frac{1}{\delta})$, since if η ever drops below ε, 0 is a suitable output, and the number of copies used is at most $(\mathrm{poly}(n))^{1/(\max(\mathsf{OPT},\varepsilon))^2 + \log\frac{1}{\varepsilon}} \mathrm{poly}(\log\frac{1}{\delta})$.

Now, we show that a good product state cover cannot be too large.

Claim 5.4 (Size of a good cover)**.** Let $\mathcal{C}$ be a (η,ε,b,B)-good cover for a state with density matrix ρ. Then $|\mathcal{C}| \leqslant \frac{1}{\eta - \varepsilon - \frac{1}{b}}$, provided $\eta - \varepsilon - \frac{1}{b} > 0$.

Proof. Let $\mathcal{C} = \{|\pi^{(i)}\rangle\}_i$, and let M be the matrix whose ith column is $|\pi^{(i)}\rangle$. Consider the Gram matrix $M^\dagger M$; its (i,j)th entry is $\langle\pi^{(i)}|\pi^{(j)}\rangle$. When $i = j$, this entry is 1, and otherwise, the entry has bounded magnitude:

$$|\langle\pi^{(i)}|\pi^{(j)}\rangle| \leqslant \frac{1}{\mathrm{d_{tan}}(|\pi^{(i)}\rangle, |\pi^{(j)}\rangle)} \leqslant \frac{1}{b},$$

where the first step follows from Lemma 3.6 and the second from property 2 of the definition of a good product state cover. So, every row has one diagonal entry of value 1 and the other $|\mathcal{C}| - 1$ entries are bounded by $\frac{1}{b}$. A consequence of the Gershgorin circle theorem is that the operator norm of a symmetric matrix is bounded by the maximum sum of absolute values of all the entries in a single column. So,

$$\|M\|_{\mathrm{op}}^2 = \|M^\dagger M\|_{\mathrm{op}} \leqslant 1 + \frac{|C|}{b}.$$

This gives us the upper bound $\leqslant 1 + |\mathcal{C}|/b$. We can similarly lower bound the operator norm of M:

$$\|M\|_{\mathrm{op}}^2 = \|MM^\dagger\|_{\mathrm{op}} \geqslant \mathrm{tr}[MM^\dagger\rho] = \sum_i \langle\pi_i|\rho|\pi_i\rangle \geqslant |\mathcal{C}|(\eta - \varepsilon).$$

Here we use the definition of the operator norm and property 1 of the definition of a good product state cover. Putting both bounds together, we have that

$$|\mathcal{C}| \leqslant \frac{1}{\eta - \varepsilon - \frac{1}{b}}$$

as desired. ⧫

5.1 Finding a good product state cover

Now we present an iterative algorithm that builds a good product state cover as it sweeps along the registers of the input state.

Algorithm 5.5 (Extending a good product state cover)**.**

Input: Copies of a state ρ; parameters $\eta \in (0,1)$ and $\varepsilon \in (0, \frac{\eta}{3})$

Output: $\mathcal{C}$, a (η, ε) good product state cover for ρ.

Subroutine: We assume the existence of an oracle which, given a set of product states $\mathcal{C}$, copies of the state ρ, and a "root" product state $|\varphi\rangle$, either outputs a classical description of a product state $|\pi\rangle$ or $\perp$. The output is guaranteed to satisfy

(a) $\langle\pi|\rho|\pi\rangle \geqslant \eta - \varepsilon$;

(b) For all $|\pi'\rangle \in \mathcal{C}_k$, $\mathrm{d}_{\tan}(|\pi\rangle, |\pi'\rangle) \geqslant 2/\eta$;

If there is a product state $|\pi\rangle$ such that

(a′) $\langle\pi|\rho|\pi\rangle \geqslant \eta$;

(b′) For all $|\pi'\rangle \in \mathcal{C}_k$, $\mathrm{d}_{\tan}(|\pi\rangle, |\pi'\rangle) \geqslant 3/\eta$;

(c′) $\mathrm{d}_{\tan}(|\pi\rangle, |\varphi\rangle) \leqslant 4/\eta$;

then the oracle is guaranteed to not output $\perp$.

Procedure:

Let $\mathcal{N} \subset \mathbb{C}^2$ be a 1-tangent distance net over qubits;
Let $\mathcal{C}_0 = \{1\}$;
for k from 1 to n **do**
    ▷ *Create an (η, ε)-good product state cover for $\rho_{[k]}$* ◁
    Let $\mathcal{C}_k = \varnothing$;
    loop
        Call the oracle on $\mathcal{C}_k$ and $\rho_{[k]}$ and all $|\varphi\rangle \in \mathcal{C}_{k-1} \otimes \mathcal{N}$;
        If any of the calls return a product state $|\pi\rangle \in (\mathbb{C}^2)^{\otimes k}$, add it to $\mathcal{C}_k$;
        Otherwise, exit the loop;
Output $\mathcal{C}_n$;

Remark 5.6 (An explicit 1-tangent distance net)**.** In the algorithm, we need a net over qubit states $\mathcal{N}$ such that, for every state $|\phi\rangle$, there is a $|\varphi\rangle \in \mathcal{N}$ such that $\mathrm{d}_{\tan}(|\phi\rangle, |\varphi\rangle) < 1$. By Definition 3.2, considering the states on the Bloch sphere, this means that the angle θ between

$|\phi\rangle$ and $|\varphi\rangle$ is less than $\pi/2$. So, we can form such a net just by picking the states on the axes of the Bloch sphere: $\mathcal{N} = \{|0\rangle, |1\rangle, |+\rangle, |-\rangle, |+\mathrm{i}\rangle, |-\mathrm{i}\rangle\}$.

Claim 5.7. Algorithm 5.5 outputs a (η, ε) good product state cover for ρ, requiring $O(\frac{n}{\eta^2})$ runs of the subroutine and $O(\frac{n^2}{\eta^2})$ classical overhead.

Proof. We will show that $\mathcal{C}_k$ is an (η, ε)-good product state cover for $\rho_{[k]}$, by induction on k. We consider the oracle's behavior in Line 7, when run on $\mathcal{C}_k$ and $\rho_{[k]}$. First, we observe that, for all k, because conditions (a) and (b) of the subroutine guarantee are identical to properties 1 and 2 of Definition 5.1 with respect to $\rho_{[k]}$, $\mathcal{C}_k$ will always obey such properties. So, it suffices to show that $\mathcal{C}_k$ also obeys property 3 at the end of the "repeat" loop.

We first show for $k = 1$. Consider some product state $|\phi\rangle$ such that $\langle\phi|\rho_{[k]}|\phi\rangle \geqslant \eta$. Then $|\phi\rangle$ satisfies condition (a') and (c') in Line 7 for some element of $\mathcal{C}_0 \otimes \mathcal{N}$: since $\mathcal{N}$ is a net, there is some $|\varphi\rangle \in \mathcal{C}_0 \otimes \mathcal{N} = \mathcal{N}$ such that $\mathrm{d}_{\tan}(|\varphi\rangle, |\phi\rangle) < 1 \leqslant 4/\eta$. Because the oracle output $\perp$ when run on $\mathcal{C}_k$, this means that (b') must not be satisfied for $|\phi\rangle$. In other words, some $|\pi\rangle \in \mathcal{C}_1$ satisfies $\mathrm{d}_{\tan}(|\pi\rangle, |\phi\rangle) < \frac{3}{\eta}$. This shows that $\mathcal{C}_1$ satisfies property 3 of the product state cover.

For $k > 1$, again consider a product state $|\phi\rangle = |\phi_1\rangle \ldots |\phi_k\rangle$ such that $\langle\phi|\rho_{[k]}|\phi\rangle \geqslant \eta$. Then $|\phi\rangle$ satisfies condition (a'). Further, it satisfies condition (c') for some $|\varphi\rangle \in \mathcal{C}_{k-1} \otimes \mathcal{N}$: since $\mathcal{C}_{k-1}$ is a good product state cover and $(\langle\phi_1| \ldots \langle\phi_{k-1}|)\rho_{[k-1]}(|\phi_1\rangle \ldots |\phi_{k-1}\rangle) \geqslant \eta$, there exists a product state $|\nu\rangle$ in $\mathcal{C}_{k-1}$ such that $\mathrm{d}_{\tan}(|\phi_1\rangle \ldots |\phi_{k-1}\rangle, |\nu\rangle) \leqslant 3/\eta$. Then, there is some $|\varphi_k\rangle$ such that $\mathrm{d}_{\tan}(|\phi\rangle, |\nu\rangle|\varphi_k\rangle) \leqslant \sqrt{(\frac{3}{\eta})^2 + 1} \leqslant \frac{4}{\eta}$ as claimed.

Thus, because of the guarantee of the oracle, after the repeat loop terminates, condition (b') cannot be true for $|\phi\rangle$. So, there is a $|\pi\rangle \in \mathcal{C}_k$ such that $\mathrm{d}_{\tan}(|\phi\rangle, |\pi\rangle) < 3/\eta$. This shows that $\mathcal{C}_k$ satisfies property 3 of the product state cover.

Every product state cover satisfies $|\mathcal{C}_k| \leqslant 1/(\eta - \varepsilon - \eta/2) \leqslant 6/\eta$ by Claim 5.4, so the subroutine only needs to be run at most $n(6/\eta)^2|\mathcal{N}|$ times. The only additional overhead is the task of storing the cover, which takes $O(n)$ time with a classical computer per element added. ♦

5.2 Finding candidate product states

Now, we specify how to perform Line 7 in Algorithm 5.5. We restate the goal of that subroutine here.

Lemma 5.8. *Suppose we are given a set of r product state constraints $\{(\vec{a}^{(s)}, b)\}_{s\in[r]}$ where $\vec{a}^{(s)} \in \mathbb{C}^m$ and $b > 0$, a description of a known "root" product state $|\varphi\rangle \in (\mathbb{C}^2)^{\otimes m}$, and error parameters $\eta \in (0,1)$, $\varepsilon \in (0, \eta/3)$, and $\delta \in (0,1)$. Then there is an algorithm which, with probability $\geqslant 1 - \delta$, successfully performs the subroutine as specified in Algorithm 5.5: it outputs either $\perp$ or a $\vec{z} \in \mathbb{C}^m$ such that*

(a) $\langle\pi_{\vec{z}}|\rho|\pi_{\vec{z}}\rangle \geqslant \eta - \varepsilon$;

(b) For all $s \in [r]$, $\mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}^{(s)}}\rangle) \geqslant b$.

If there is a product state $|\pi\rangle$ such that

(a') $\langle\pi|\rho|\pi\rangle \geqslant \eta$;

(b') For all $s \in [r]$, $\mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}^{(s)}}\rangle) \geqslant 1.5b$;

(c') $\mathrm{d}_{\tan}(|\pi\rangle, |\varphi\rangle) \leqslant B$;

then the output is guaranteed to not be $\perp$*. The algorithm uses* $N \leqslant (\mathrm{poly}(m))^{(B^2+\log\frac{1}{\varepsilon})} \log\frac{1}{\delta}$ *copies of* ρ*,* $\mathrm{poly}(m, B, \log\frac{1}{\varepsilon})$ *quantum gates per copy of* ρ*, and* $m^{\mathrm{poly}(r,B,b,1/b,1/\varepsilon)} \mathrm{poly}(\log\frac{1}{\delta})$ *additional classical overhead.*

From this, we can immediately conclude our main result:

Proof of Theorem 5.2. By Claim 5.7, to construct the cover, it suffices to call Lemma 5.8 $\mathrm{poly}(n, 1/\eta)$ times with parameters $m \leqslant n$, $B = 4/\eta$, $b = 3/\eta$, and a set of product state constraints where $r = O(1/\eta)$ by Claim 5.4. There is a failure of probability associated to each run of the subroutine, but the failure probability parameter can be rescaled such that the probability of all calls succeeding is at least $\geqslant 1 - \delta$. This gives the stated running time, and associated quantum complexities. ♦

The rest of this section is devoted to proving Lemma 5.8. The algorithm is given in Algorithm 5.10; we spend the rest of this subsection describing the intuition for this algorithm. We prove the desired complexity bounds in Claim 5.11, and then we prove the above guarantees in Claim 5.18 and Claim 5.19. The complexity bound requires the use of a polynomial optimization routine, which is described and proved later, in Section 6.

Algorithm intuition. Our goal is to search within tangent distance B of the root state for a product state that has good overlap with ρ, if one exists. Further, we have additional constraints that the state we find be far away from a collection of r product states in tangent distance. Without loss of generality, we can take the root state to be $|\pi_{\vec{0}}\rangle = |0^n\rangle$. Then it suffices to find an (approximate) solution to the following optimization problem, which finds the product state with the best fidelity under these constraints:

$$\begin{aligned} \underset{\vec{z}\in\mathbb{C}^m}{\text{maximize}} \quad & \langle\pi_{\vec{z}}|\,\rho\,|\pi_{\vec{z}}\rangle \\ \text{subject to} \quad & \mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}^{(s)}}\rangle) \geqslant b \text{ for all } s \in [r], \\ & \mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{0}}\rangle) \leqslant B. \end{aligned} \tag{P1}$$

We would like to reduce this task to one of optimizing a low-degree polynomial under simple constraints. The criterion $\mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{0}}\rangle) \leqslant B$ is equivalent to $\|\vec{z}\|_2 \leqslant B$. However, the other tangent distance constraints do not simplify so easily. We can simplify them for small coordinates, and because $\vec{z}$ is bounded norm, most coordinates are small; we encode this into the program:

$$\begin{aligned} \underset{\vec{z},\, S\subseteq[m]}{\text{maximize}} \quad & \langle\pi_{\vec{z}}|\,\rho\,|\pi_{\vec{z}}\rangle \\ \text{subject to} \quad & |\vec{z}_i| \leqslant \mu \text{ for all } i \notin S \\ & |S| \leqslant B^2/\mu^2 \\ & \mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}^{(s)}}\rangle) \geqslant b \text{ for all } s \in [r], \\ & \|\vec{z}\|_2 \leqslant B. \end{aligned} \tag{P2}$$

This program gives an equivalent solution to (P1). Now we can approximate the farness

constraints by an ℓ^2 constraint for the coordinates outside of S.

$$\begin{aligned} \underset{\vec{z},\, S\subset[m]}{\text{maximize}} \quad & \langle \pi_{\vec{z}} |\, \rho \,| \pi_{\vec{z}} \rangle \\ \text{subject to} \quad & |\vec{z}_i| \leqslant \mu \text{ for all } i \notin S \\ & |S| \leqslant B^2/\mu^2 \\ & \mathrm{d}_{\tan}(|\pi_{\vec{z}_S}\rangle, |\pi_{\vec{a}_S^{(s)}}\rangle)^2 + \|\vec{z}_{\overline{S}} - \vec{a}_{\overline{S}}^{(s)}\|_2^2 \geqslant 1.5b^2 \text{ for all } s \in [r], \\ & \|\vec{z}\|_2 \leqslant B. \end{aligned} \tag{P3}$$

Our first key lemma (Lemma 5.16) shows that this is a stronger constraint than that of (P2), but still satisfies for completeness, so the program will still find a good product state under the desired circumstances.

Finally, we would like to approximate the objective function $\langle \pi_{\vec{z}} | \rho | \pi_{\vec{z}} \rangle$ by a low-degree polynomial. We do this by replacing ρ with a truncation: for $d \in [m]$, let $\rho_d = \Pi_{\leqslant d} \rho \Pi_{\leqslant d}$, where $\Pi_{\leqslant d}$ is the projector onto computational basis strings with Hamming weight less than or equal to d. Then $\langle \pi_{\vec{z}} | \rho_d | \pi_{\vec{z}} \rangle$ is a degree-d polynomial multiplied by the normalization $\prod_{i\in[m]} \frac{1}{1+|z_i|^2}$. Because most coordinates of $\vec{z}$ are small, this normalization can be captured by a simpler quantity: $\prod_{i\in\overline{S}} \frac{1}{1+|z_i|^2} \approx e^{-\|\vec{z}_{\overline{S}}\|_2^2}$. After taking these approximations, we have the following optimization problem.

$$\begin{aligned} \underset{\vec{z},\, S\subset[m]}{\text{maximize}} \quad & p_{\vec{z}_S}(\vec{z}_{\overline{S}}) = \frac{e^{-\|\vec{z}_{\overline{S}}\|_2^2}}{\prod_{i\in S}(1+|z_i|^2)} \sum_{x,x'\in\{0,1\}^m} \langle x | \, \rho_d \, | x' \rangle \, (\vec{z}^*)^x (\vec{z})^{x'} \\ \text{subject to} \quad & |z_i| \leqslant \mu \text{ for all } i \notin S \\ & |S| \leqslant B^2/\mu^2 \\ & \mathrm{d}_{\tan}(|\pi_{\vec{z}_S}\rangle, |\pi_{\vec{a}_S^{(s)}}\rangle)^2 + \|\vec{z}_{\overline{S}} - \vec{a}_{\overline{S}}^{(s)}\|_2^2 \geqslant 1.5b^2 \text{ for all } s \in [r], \\ & \|\vec{z}\|_2 \leqslant B. \end{aligned} \tag{P4}$$

Above, for a vector $\vec{z} \in \mathbb{C}^m$ and string $x \in \{0,1\}^m$, we use the notation $(\vec{z})^x = \prod_{x_i=1} z_i$. Our second key lemma (Lemma 5.17) shows that the objective functions of (P4) and (P3) are close on the domain.

(P4) is the optimization program implicitly being run by the eventual algorithm (Algorithm 5.10). There are three major differences. First, we do not know ρ_d, so we perform tomography on ρ to get some estimate ϱ_d which we use instead (Lemma 5.9). This is efficient because we are performing tomography on a small subspace Π_d. Second, the above program still has poorly behaved elements. However, these are all located in the subspace $V^{(S)}$ spanned by the coordinates of S and the $\vec{a}^{(s)}$'s. If we knew the true value of $\vec{z}$ on $V^{(S)}$ (and the true value of $\|\vec{z}_{\overline{S}}\|_2$), then the objective function becomes a degree-d polynomial and the constraints all become simple ℓ_2 and ℓ_∞ constraints, which is a form we can solve (Theorem 6.3). However, this is a low-dimensional subspace, so we can simply guess the value of $\vec{z}$ on this subspace, and pick the best solution among all guesses. Finally, we need to allow for some tolerance ε_{tol} in our argument, because of the error in guessing and the error of the final polynomial optimization algorithm.

5.3 Algorithm and showing running time

We begin with a lemma for the form of state tomography we need for the algorithm. We did not attempt to optimize this, but we note that this is the only "quantum" part of the algorithm,

and any tomography algorithm could be used here. Here, we choose one which only performs single-copy Clifford measurements. This is also the only part of the algorithm which is random, and so this is the only place where there is a probability of failure. For our correctness proofs, we will assume that this tomography step always completes successfully.

Lemma 5.9 (Computationally efficient state tomography)**.** *There is an algorithm which, given copies of ρ and a parameter d, outputs a positive semi-definite matrix $\hat{\rho}$ with trace at most 1 and supported on the strings of Hamming weight $\leqslant d$, such that $\|\hat{\rho} - \Pi_{\leqslant d}\rho\Pi_{\leqslant d}\|_{\mathrm{op}} < \varepsilon$. This algorithm uses $N = O(\frac{(10n)^{2d}}{\varepsilon^2}\log\frac{1}{\delta})$ copies of ρ, along with $\mathrm{poly}(n,d)$ quantum gates per copy of ρ and $\mathrm{poly}(N)$ classical processing.*

Proof. First, similarly to Lemma 4.10, for every copy of ρ, we can attach $d\lceil\log_2(n+1)\rceil + O(\log d)$ ancilla qubits, and then apply a $\mathrm{poly}(n,d)$-sized circuit which performs the following unitary. For a set $J \subseteq [n]$ with elements $i_1 < i_2 < \cdots < i_{|J|}$, the circuit maps $|0^d\rangle|e_J\rangle \mapsto |i_1\rangle|i_2\rangle\ldots|i_{|J|}\rangle|0^{d-|J|}\rangle|e_J\rangle$ when $|J| \leqslant d$, and does nothing to $|0^d\rangle|e_J\rangle$ when $|J| > d$. Here, we use e_J to denote the n-qubit state which is $|1\rangle$ on the qubits in J and $|0\rangle$ otherwise. After applying this circuit and discarding the n-qubit state, the density matrix of the $d\lceil\log_2(n+1)\rceil$ ancilla qubits contains $\Pi_{\leqslant d}\rho\Pi_{\leqslant d}$ as a submatrix. So, it suffices to perform tomography on this density matrix of size $D \leqslant (10n)^d$, which we denote σ, and output the corresponding matrix, which is $\Pi\sigma\Pi$ for some projector Π on some subset of computational basis states.

We use a simple, gate-efficient tomography algorithm. Consider the process that measures a random Clifford circuit, C, performs $C^\dagger$ on the σ and measures in the computational basis to get $|i\rangle$, and then taking the estimator $((D+1)C|i\rangle\langle i|C^\dagger - I)$. If we perform this process N times and then average the estimator to form $\hat{\sigma}$, it satisfies the following [Low21, Section 1.5.2]:

$$\mathbb{E}\|\hat{\sigma} - \sigma\|_F^2 \leqslant \frac{D^2 + D - 1}{N}.$$

In the language of Flammia and O'Donnell [FO24], this is a state estimation algorithm with Frobenius-squared rate $\frac{D^2+D-1}{N}$, and so by [FO24, Proposition 3.10], with only a constant factor loss in the guarantee, we can modify the algorithm such that it always outputs a matrix $\hat{\sigma}$ which is positive semi-definite and satisfies $\mathrm{tr}(\hat{\rho}) = 1$. Further, by [FO24, Proposition 3.11], the guarantee can be upgraded to the guarantee that $\|\hat{\sigma} - \sigma\|_F^2 \leqslant \frac{c(D^2+D-1)}{N}\log\frac{1}{\delta}$ with probability $\geqslant 1-\delta$, for some sufficiently large constant. Since $\|X\|_{\mathrm{op}} \leqslant \|X\|_F$, for some choice of $N = O(\frac{D}{\varepsilon}\log\frac{1}{\delta})$, we have that $\|\hat{\sigma} - \sigma\|_{\mathrm{op}} < \varepsilon$ with probability $\geqslant 1-\delta$. Finally, this operator norm bound also bounds the operator norm distance for every submatrix, so we can conclude that for $\hat{\rho}$ formed by re-labeling the rows and columns of $\Pi\hat{\sigma}\Pi$, $\|\hat{\rho} - \Pi_{\leqslant d}\rho\Pi_{\leqslant d}\|_{\mathrm{op}} < \varepsilon$. ♦

For the following algorithm, and throughout this section, for a vector $\vec{v}$ we use the notation $\vec{v}_S \in \mathbb{C}^{|S|}$ to denote the vector restricted to the coordinates of S, and for two vectors $\vec{u}_S, \vec{v}_{\overline{S}}$, the notation $(\vec{u}_S, \vec{v}_{\overline{S}})$ denotes the vector which is $\vec{u}$ on the coordinates corresponding to S and $\vec{v}$ on the coordinates corresponding to $\overline{S}$.

Algorithm 5.10 (Adding a new element to the cover)**.**

Input: A set of product state constraints $\{(\vec{a}^{(s)}, b)\}_{s\in[r]}$; copies of an m-qubit state ρ; an explicit known "root" product state $|\varphi\rangle \in (\mathbb{C}^2)^{\otimes m}$; parameters $B, \eta, \varepsilon, \delta$.

Output: Either $\perp$ or a $\vec{z} \in \mathbb{C}^m$ with guarantees as in Lemma 5.8

Procedure:

1: Recenter $|\varphi\rangle$ to be $|0^m\rangle$;
2: Let $\varepsilon_{\text{approx}} = \varepsilon/100$; ▷ *$\varepsilon_{\text{approx}}$ is for approximation errors.*
3: Choose $\mu \leqslant \frac{1}{10}\min(b, \frac{1}{b}, \frac{\sqrt{\varepsilon_{\text{approx}}}}{B})$; ▷ *$\mu$ is the eventual ℓ_∞ bound; we need the b upper bound for Lemma 5.16, and the $\sqrt{\varepsilon}/B^2$ upper bound for Lemma 5.17*
4: Let $d = 10B^2 + \log\frac{2}{\varepsilon_{\text{approx}}}$; ▷ *$d$ is the degree of the eventual polynomial*
5: Let $\varepsilon_{\text{tol}} \leqslant \min(\gamma, \mu^4/\sqrt{m}, 0.01\varepsilon) = \text{poly}(1/m, \mu, \varepsilon)$; ▷ *Here, γ is the tolerance parameter coming from Theorem 6.3; the other parts are used in the proof of soundness (Claim 5.18).*
6: Perform tomography on $\rho_d = \Pi_{\leqslant d}\rho\Pi_{\leqslant d}$ to get a description of ϱ_d such that $\|\varrho_d - \rho_d\|_{\text{op}} \leqslant \varepsilon_{\text{approx}}$ with probability $\geqslant 1-\delta$, where $\Pi_{\leqslant d} = \sum_{|x|\leqslant d} |x\rangle\langle x|$ (Lemma 5.9); ▷ *Note that we will use the additional properties of ϱ_d guaranteed by Lemma 5.9. In particular, ϱ_d is supported on the image of $\Pi_{\leqslant d}$.*
7: **for all** subsets $S \subseteq [m]$ of size $\leqslant B^2/\mu^2$ **do**
8: Let $V^{(S)}$ be the subspace spanned by $\{\vec{a}^{(s)}\}_{s\in[r]}$ and the computational basis vectors associated to S;
9: Let $\mathcal{N}_S$ be an ε_{tol}-net over the space $\mathcal{F}_S$, where

$$\mathcal{F}_S = \Big\{\vec{v} \in V^{(S)},\, \nu \in [0, B] \;\Big|\; \|\vec{v}_S\|_2^2 + \nu^2 \leqslant B^2,\ \|\vec{v}\|_2 \leqslant B,$$
$$\nu^2 - \|\vec{v}_{\overline{S}}\|_2^2 + \|\vec{v}_{\overline{S}} - \vec{a}^{(s)}_{\overline{S}}\|_2^2 \geqslant 1.5b^2 - \mathrm{d}_{\tan}(|\pi_{\vec{v}_S}\rangle, |\pi_{\vec{a}^{(s)}_S}\rangle)^2 \text{ for all } s \in [r]\Big\}. \quad (10)$$

▷ *The net enforces that the output is far from the $\vec{a}^{(s)}$'s; the error parameter needs to be at most ε_{tol}, and is used to show completeness (Claim 5.19). The formal criteria the net satisfies is given in Claim 5.15.*
10: **for** $(\vec{v}, \nu) \in \mathcal{N}_S$ **do**
11: Consider the domain

$$\mathcal{D}_{\varepsilon_{\text{tol}}} = \{\vec{z}_{\overline{S}} \in \mathbb{C}^{|\overline{S}|} \mid |\|\vec{z}_{\overline{S}}\|_2 - \nu| \leqslant \varepsilon_{\text{tol}},\ \|\Pi_{V^{(S)}}\vec{z} - \vec{v}\|_2 \leqslant \varepsilon_{\text{tol}},$$
$$\|\vec{z}_{\overline{S}}\|_\infty \leqslant \mu + \varepsilon_{\text{tol}} \text{ for } \vec{z} = (\vec{v}_S, \vec{z}_{\overline{S}})\}; \quad (11)$$

12: Use Theorem 6.3 to find a $\vec{y}_{\overline{S}} \in \mathcal{D}_{2\varepsilon_{\text{tol}}}$ such that

$$p_{\vec{v}_S,\nu}(\vec{y}_{\overline{S}}) \geqslant \max_{\vec{z}_{\overline{S}}\in\mathcal{D}_{\varepsilon_{\text{tol}}}} p_{\vec{v}_S,\nu}(\vec{z}_{\overline{S}}) - \varepsilon_{\text{approx}}$$
$$\text{for } p_{\vec{z}_S,\nu}(\vec{z}_{\overline{S}}) = \frac{e^{-\nu^2}}{\prod_{i\in S}(1+|z_i|^2)} \sum_{x,x'\in\{0,1\}^m} \langle x|\,\varrho_d\,|x'\rangle\,(\vec{z}^*)^x\vec{z}^{x'}, \quad \text{(P0)}$$

13: Add $\vec{y} = (\vec{v}_S, \vec{y}_{\overline{S}})$ to the list of candidate solutions, along with its objective value $p_{\vec{y}} \leftarrow p_{\vec{v}_S,\nu}(\vec{y}_{\overline{S}})$;
14: Let $\vec{u}$ be the candidate solution achieving the largest objective value $p_{\vec{u}}$;
15: If $p_{\vec{u}} \geqslant \eta - \varepsilon/2$, output it; output $\perp$ otherwise.

Claim 5.11. For some sufficiently large $C > 1$, Algorithm 5.10 requires $N \leqslant m^{C(B^2+\log\frac{1}{\varepsilon})}\log\frac{1}{\delta}$ copies of ρ, $\mathrm{poly}(m, B, \log\frac{1}{\varepsilon})$ quantum gates per copy of ρ, $(\mathrm{poly}(B,b,m)/\varepsilon_{\mathrm{tol}})^{B^2/\mu^2+r+1}$ calls to the optimization problem (P0), and $(N + (\mathrm{poly}(B,b,m)/\varepsilon_{\mathrm{tol}})^{B^2/\mu^2+r})^C$ additional classical overhead.

Proof. The only step of the algorithm which is quantum is the tomography step, so the quantum complexities follow from Lemma 5.9 with $d = 10B^2 + \log\frac{2}{\varepsilon_{\mathrm{approx}}}$.

The number of calls to the optimization problem is at most the number of subsets iterated over times a bound on the size of every net $\mathcal{N}_S$. The number of subsets of cardinality at most B^2/μ^2 is at most $(m+1)^{B^2/\mu^2}$. By Claim 5.15, the cardinality of $\mathcal{N}_S$ is at most $(\frac{1}{c\varepsilon_{\mathrm{tol}}})^{|S|+r+1}$ for some $c = \mathrm{poly}(1/B, 1/b, m)$. Multiplying the two together, we get $(m+1)^{B^2/\mu^2}(\frac{1}{c\varepsilon_{\mathrm{tol}}})^{B^2/\mu^2+r+1}$ as desired.

The running time is dominated by the tomography algorithm and the construction of the nets $\mathcal{N}_S$; both take polynomial time, by Lemma 5.9 and Claim 5.15. ♦

Claim 5.12. A $y_{\overline{S}}$ as in (P0) can be solved in $(B/\varepsilon_{\mathrm{tol}})^{\mathrm{poly}(d,r,1/\varepsilon_{\mathrm{approx}},1/\mu)}$ time. Thus, the classical overhead of Algorithm 5.10 is $m^{\mathrm{poly}(r,B,b,1/b,1/\varepsilon)}\,\mathrm{poly}(\log\frac{1}{\delta})$.

Proof. This is a corollary of Theorem 6.3. Recall that we are considering the objective function

$$\begin{aligned} p_{\vec{z}_S,\nu}(\vec{z}_{\overline{S}}) &= \frac{e^{-\nu^2}}{\prod_{i\in S}(1+|z_i|^2)} \sum_{x,x'\in\{0,1\}^m} \langle x|\, \varrho_d\, |x'\rangle\, (\vec{z}^*)^x \vec{z}^{x'}. \\ &= e^{-\nu^2} \sum_{x,x'\in\{0,1\}^{|\overline{S}|}} \langle x|\, \Big((I_{\overline{S}} \otimes \langle \pi_{\vec{z}_S}|_S)\varrho_d(I_{\overline{S}} \otimes |\pi_{\vec{z}_S}\rangle_S)\Big)\, |x'\rangle\, (\vec{z}^*_{\overline{S}})^x (\vec{z}_{\overline{S}})^{x'} \end{aligned}$$

Defining $\sigma = (I_{\overline{S}} \otimes \langle \pi_{\vec{z}_S}|_S)\varrho_d(I_{\overline{S}} \otimes |\pi_{\vec{z}_S}\rangle_S)$, we can express $p_{\vec{z}_S,\nu}(z_{\overline{S}})$ as

$$p_{\vec{z}_S,\nu}(\vec{z}_{\overline{S}}) = e^{-\nu^2}\Big(\prod_{i\in\overline{S}}(1+|z_i|^2)\Big)\, \langle \pi_{\vec{z}_{\overline{S}}}|\, \sigma\, |\pi_{\vec{z}_{\overline{S}}}\rangle\,,$$

We can further write it in terms of tensors $T^{(k)} \in (\mathbb{C}^n)^{\otimes 2k}$.

$$\begin{aligned} p_{\vec{z}_S,\nu}(\vec{z}_{\overline{S}}) &= e^{-\nu^2}\Big(T^{(0)} + \langle T^{(1)}, z^*_{\overline{S}} \otimes z_{\overline{S}}\rangle + \cdots + \langle T^{(d)}, (z^*_{\overline{S}})^{\otimes d} \otimes z_{\overline{S}}^{\otimes d}\rangle\Big), \\ T^{(k)}_{i_1\ldots i_k j_1\ldots j_k} &= \begin{cases} \frac{1}{(k!)^2}\, \langle e_{i_1,\ldots,i_k}|\, \sigma\, |e_{j_1,\ldots,j_k}\rangle & i_1,\ldots,i_k \text{ are distinct, } j_1,\ldots,j_k \text{ are distinct} \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Note that $\|T^{(k)}\|_F \leqslant \frac{1}{(k!)}\|\sigma\|_F$, since for every entry of σ there are $(k!)^2$ corresponding entries in $T^{(k)}$ containing it, scaled down by a factor of $(k!)^2$. In order to apply Theorem 6.3, we can further break this up into two cases. When $\nu \leqslant 1$, we optimize $\frac{1}{10}p_{\vec{z}_S,\nu}(z_{\overline{S}})$ (that is, the function $f_{\frac{e^{-\nu^2}}{10}T}(\vec{z}_{\overline{S}})$) to $\varepsilon_{\mathrm{approx}}/10$ error, since with this choice the tensors satisfy

$$\sum_{k=0}^{d} \frac{e^{-\nu^2}}{10}\|T^{(k)}\|_F \leqslant \sum_{k=0}^{d} \frac{1}{10(k!)} \leqslant 1.$$

The domain we want to optimize over is, recalling (11),

$$\begin{aligned} \mathcal{D}_{\varepsilon_{\mathrm{tol}}} = \{\vec{z}_{\overline{S}} \in \mathbb{C}^{|\overline{S}|} \mid\, & |\|\vec{z}_{\overline{S}}\|_2 - \nu| \leqslant \varepsilon_{\mathrm{tol}},\ \|\Pi_{V^{(S)}}\vec{z} - \vec{v}\|_2 \leqslant \varepsilon_{\mathrm{tol}}, \\ & \|\vec{z}_{\overline{S}}\|_\infty \leqslant \mu + \varepsilon_{\mathrm{tol}} \text{ for } \vec{z} = (\vec{v}_S, \vec{z}_{\overline{S}})\}. \end{aligned}$$

This domain is almost of the form needed for Theorem 6.3, Definition 6.2; all we need is to make the following adjustment:

$$\|\Pi_{V^{(S)}}\vec{z} - \vec{v}\|_2 \leqslant \varepsilon_{\text{tol}} \iff \|W^\dagger(\vec{z}_{\overline{S}} - \vec{v}_{\overline{S}})\|_2 \leqslant \varepsilon_{\text{tol}},$$

where $W \in \mathbb{C}^{|\overline{S}|\times r}$ is a matrix with $\|W\|_{\text{op}} \leqslant 1$ such that $(WW^\dagger)\vec{z}_{\overline{S}} = \Pi_{V^{(S)}}(\vec{0}_S, \vec{z}_{\overline{S}})$. This is possible since $\Pi_{V^{(S)}}$ is rank at most r over the subspace spanned by the vectors $(\vec{0}_S, \vec{z}_{\overline{S}})$. Then, Theorem 6.3 outputs a $\vec{y} \in \mathbb{C}^{|\overline{S}|}$ such that $\frac{1}{10}p_{\vec{z}_S,\nu}(\vec{y}) \geqslant \max_{\vec{z}_{\overline{S}}\in\mathcal{D}^{2\varepsilon_{\text{tol}}}} \frac{1}{10}p_{\vec{z}_S,\nu}(\vec{z}_{\overline{S}}) - \frac{\varepsilon_{\text{approx}}}{10}$. This is the desired bound after rescaling.

When $\nu > 1$, we optimize the tensors $e^{-\nu^2}\nu^{2k}T^{(k)}$ with respect to the variables $\vec{t} = \vec{z}_{\overline{S}}/\nu$. That is, we write

$$p_{\vec{z}_S,\nu}(\vec{z}_{\overline{S}}) = \left(e^{-\nu^2}T^{(0)} + \cdots + \langle e^{-\nu^2}\nu^{2d}T^{(d)}, (\vec{t}^*)^{\otimes d} \otimes (\vec{t})^{\otimes d}\rangle\right)$$

Then, the sum of the norms of the tensors is

$$\sum_{k=0}^{d} e^{-\nu^2}\nu^{2k}\|T^{(k)}\|_F \leqslant e^{-\nu^2}\sum_{k=0}^{d}\frac{\nu^{2k}}{k!} \leqslant 1.$$

So, if we apply Theorem 6.3 to with error parameter $\varepsilon_{\text{approx}}$, we are done. Finally, we can write the domain in terms of the variables $\vec{t}$:

$$\begin{aligned} |\|\vec{z}_{\overline{S}}\|_2 - \nu| \leqslant \varepsilon_{\text{tol}} &\iff |\|\vec{t}\|_2 - 1| \leqslant \varepsilon_{\text{tol}}/\nu \\ \|\Pi_{V^{(S)}}\vec{z} - \vec{v}\|_2 \leqslant \varepsilon_{\text{tol}} &\iff \|W^\dagger(\vec{t} - \vec{v}_{\overline{S}}/\nu)\|_2 \leqslant \varepsilon_{\text{tol}}/\nu \end{aligned}$$

Here, we set our tolerance parameter to be $\varepsilon_{\text{tol}}/\nu$, where $1 \leqslant \nu \leqslant B$, giving the desired running time. ♦

5.4 Showing correctness

Lemma 5.13. *For any vectors $\vec{u}, \vec{v}, \vec{a} \in \mathbb{C}^m$ and parameters $B > 1, \varepsilon > 0$, if $\|\vec{u}\|_2, \|\vec{v}\|_2, \|\vec{a}\|_2 \leqslant B$ and $\mathrm{d}_{\tan}(|\pi_{\vec{v}}\rangle, |\pi_{\vec{a}}\rangle) \leqslant B$ and $\|\vec{u} - \vec{v}\|_2 \leqslant \varepsilon$ where $\varepsilon \leqslant 1/(10mB)^6$ then*

$$|\mathrm{d}_{\tan}(|\pi_{\vec{a}}\rangle, |\pi_{\vec{v}}\rangle) - \mathrm{d}_{\tan}(|\pi_{\vec{a}}\rangle, |\pi_{\vec{u}}\rangle)| \leqslant \varepsilon(10mB)^6$$

Proof. Recall that

$$\mathrm{d}_{\tan}(|\pi_{\vec{a}}\rangle, |\pi_{\vec{v}}\rangle)^2 = \sum_{i=1}^{m}\left|\frac{v_i - a_i}{1 + a_i^* v_i}\right|^2.$$

The derivative of $\frac{v_i - a_i}{1 + a_i^* v_i}$ with respect to v_i is

$$\frac{1 + a_i^* v_i - a_i^*(v_i - a_i)}{(1 + a_i^* v_i)^2}.$$

Note that whenever $|1 + a_i^* v_i| \leqslant 1/2$ then $|v_i - a_i| \geqslant 0.1$. Thus, the magnitude of the derivative is at most $100B^2\left(1 + \left|\frac{v_i - a_i}{1 + a_i^* v_i}\right|^2\right)$. Thus, if we define the vectors

$$\vec{V} = \left\{\frac{v_i - a_i}{1 + a_i^* v_i}\right\}_{i\in[m]}, \quad \vec{U} = \left\{\frac{u_i - a_i}{1 + a_i^* u_i}\right\}_{i\in[m]}$$

then integrating the above and using the assumption that $\|\vec{u} - \vec{v}\|_2 \leqslant \varepsilon$ implies

$$\|\vec{V} - \vec{U}\|_2 \leqslant 200B^3 m\varepsilon.$$

From this, we immediately get the desired inequality. ♦

Corollary 5.14. *For two vectors $\vec{u}, \vec{v} \in \mathbb{C}^m$ such that $\|\vec{u} - \vec{v}\| < \text{poly}(1/B, 1/m, 1/b)$ and $\|\vec{u}\|, \|\vec{v}\| \leqslant B$, along with an arbitrary $\vec{a}$ and $\nu, b \in \mathbb{R}$, we have that*

$$\nu^2 - \|\vec{v}_{\overline{S}}\|_2^2 + \|\vec{v}_{\overline{S}} - \vec{a}_{\overline{S}}\|_2^2 \geqslant 1.5b^2 - \mathrm{d}_{\tan}(|\pi_{\vec{v}_S}\rangle, |\pi_{\vec{a}_S}\rangle)^2$$
$$\implies \nu^2 - \|\vec{u}_{\overline{S}}\|_2^2 + \|\vec{u}_{\overline{S}} - \vec{a}_{\overline{S}}\|_2^2 \geqslant 1.49b^2 - \mathrm{d}_{\tan}(|\pi_{\vec{u}_S}\rangle, |\pi_{\vec{a}_S}\rangle)^2.$$

Proof. Note that when $\mathrm{d}_{\tan}(|\pi_{\vec{v}_S}\rangle, |\pi_{\vec{a}_S}\rangle) \geqslant 2(b + B)$ then the inequalities are both trivially true. Also, if $\|\vec{v}_{\overline{S}} - \vec{a}_{\overline{S}}\|_2 \geqslant 2(B + b)$ then the inequalities are both trivially true.

Otherwise, we can just apply Lemma 5.13 to bound the difference between $\mathrm{d}_{\tan}(|\pi_{\vec{u}_S}\rangle, |\pi_{\vec{a}_S}\rangle)$ and $\mathrm{d}_{\tan}(|\pi_{\vec{v}_S}\rangle, |\pi_{\vec{a}_S}\rangle)$ and also directly bound the differences $\|\vec{v}_{\overline{S}}\|_2^2 - \|\vec{u}_{\overline{S}}\|_2^2$ and $\|\vec{v}_{\overline{S}} - \vec{a}_{\overline{S}}\|_2^2 - \|\vec{u}_{\overline{S}} - \vec{a}_{\overline{S}}\|_2^2$ to get the desired conclusion. ♦

Claim 5.15. There is a $\mathcal{N}_S$ which satisfies the following properties. First, $\mathcal{N}_S$ satisfies that, for every $(\vec{v}, \nu) \in \mathcal{N}_S$, the following conditions hold: $\vec{v} \in V_S$; $\|\vec{v}_S\|^2 + \nu^2 \leqslant B^2$; $\|\vec{v}\| \leqslant B$; and

$$\nu^2 - \|\vec{v}_{\overline{S}}\|_2^2 + \|\vec{v}_{\overline{S}} - \vec{a}_{\overline{S}}^{(s)}\|_2^2 \geqslant 1.49b^2 - \mathrm{d}_{\tan}(|\pi_{\vec{v}_S}\rangle, |\pi_{\vec{a}_S^{(s)}}\rangle)^2.$$

Second, for any $(\vec{v}, \nu) \in \mathcal{F}_S$, there is a $(\vec{v}', \nu')$ such that $\|\vec{v} - \vec{v}'\| \leqslant \varepsilon_{\text{tol}}$ and $|\nu - \nu'| \leqslant \varepsilon_{\text{tol}}$. This net has at most $(\frac{1}{c\varepsilon_{\text{tol}}})^{10(|S|+r+1)}$ elements and can be constructed in $(\frac{1}{c\varepsilon_{\text{tol}}})^{10(|S|+r+1)}$ time, for some $c = \text{poly}(1/B, 1/b, 1/m)$.

Proof. Let $\mathcal{A}_S$ be the set

$$\mathcal{A}_S = \left\{ \vec{v} \in V^{(S)},\ \nu \in [0, B] \;\middle|\; \|\vec{v}_S\|_2^2 + \nu^2 \leqslant B^2,\ \|\vec{v}\|_2 \leqslant B \right\}$$

i.e. it is $\mathcal{F}_S$ with the tangent distance constraint removed. Now we can construct a net $\mathcal{B}_S$ for the set $\mathcal{A}_S$ such that

- For all $(\vec{v}, \nu) \in \mathcal{F}_S$, there is $(\vec{v}', \nu') \in \mathcal{B}_S$ such that $\|\vec{v} - \vec{v}'\| \leqslant c\varepsilon_{\text{tol}}$ and $|\nu - \nu'| \leqslant c\varepsilon_{\text{tol}}$
- All elements of $\mathcal{B}_S$ are in $\mathcal{A}_S$
- The net $\mathcal{B}_S$ can be enumerated in $(\frac{1}{c\varepsilon_{\text{tol}}})^{10(|S|+r+1)}$ time

We can construct $\mathcal{B}_S$ by simply enumerating over a sufficiently fine grid for $\vec{v}$ and ν. Now given $\mathcal{B}_S$, we construct $\mathcal{N}_S$ by simply removing all points $(\vec{v}', \nu')$ such that

$$\nu'^2 - \|\vec{v'}_{\overline{S}}\|_2^2 + \|\vec{v'}_{\overline{S}} - \vec{a}_{\overline{S}}^{(s)}\|_2^2 < 1.49b^2 - \mathrm{d}_{\tan}(|\pi_{\vec{v}'_S}\rangle, |\pi_{\vec{a}_S^{(s)}}\rangle)^2.$$

By Corollary 5.14, this removal never removes any $(\vec{v}', \nu')$ such that $\|\vec{v} - \vec{v}'\| \leqslant \varepsilon_{\text{tol}}$ and $|\nu - \nu'| \leqslant \varepsilon_{\text{tol}}$ for some $(\vec{v}, \nu) \in \mathcal{F}_S$. Thus, $\mathcal{N}_S$ still covers all points in $\mathcal{F}_S$, as desired. ♦

Lemma 5.16 (Approximating tangent distance constraints)**.** *For $\vec{z}, \vec{a} \in \mathbb{C}^m$, $b > 0$, and $|S| \subseteq [m]$, if $\|\vec{z}_{\overline{S}}\|_\infty \leqslant \frac{1}{6}\min(\frac{1}{b}, b)$, then the following implications hold.*

$$\mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}}\rangle) \geqslant 1.5b \implies \mathrm{d}_{\tan}(|\pi_{\vec{z}_S}\rangle, |\pi_{\vec{a}_S}\rangle)^2 + \|\vec{z}_{\overline{S}} - \vec{a}_{\overline{S}}\|_2^2 \geqslant 1.5b^2$$
$$\mathrm{d}_{\tan}(|\pi_{\vec{z}_S}\rangle, |\pi_{\vec{a}_S}\rangle)^2 + \|\vec{z}_{\overline{S}} - \vec{a}_{\overline{S}}\|_2^2 \geqslant (1.4 - \|\vec{a}_{\overline{S}}\|_\infty)b^2 \implies \mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}}\rangle) \geqslant b$$

Proof. The main idea is that we can relate $\mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}}\rangle)^2$ to $\mathrm{d}_{\tan}(|\pi_{\vec{z}_S}\rangle, |\pi_{\vec{a}_S}\rangle)^2 + \|\vec{z}_{\overline{S}} - \vec{a}_{\overline{S}}\|_2^2$ up to some small constant multiplicative error. Let $\mu = \frac{1}{6}\min(\frac{1}{b}, b)$, so that $\|\vec{z}_{\overline{S}}\|_\infty \leqslant \mu$.

We consider two cases. First, suppose $|a_i| \leqslant \frac{1}{2\mu}$ for all $i \notin S$. For such constraints, we have the following bound on the difference between the two constraints:

$$\begin{aligned}
&\left|\mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}}\rangle)^2 - (\mathrm{d}_{\tan}(|\pi_{\vec{z}_S}\rangle, |\pi_{\vec{a}_S}\rangle)^2 + \|\vec{z}_{\overline{S}} - \vec{a}_{\overline{S}}\|_2^2)\right| \\
&= \left|\mathrm{d}_{\tan}(|\pi_{\vec{z}_{\overline{S}}}\rangle, |\pi_{\vec{a}_{\overline{S}}}\rangle)^2 - \|\vec{z}_{\overline{S}} - \vec{a}_{\overline{S}}\|_2^2\right| \\
&\leqslant \mathrm{d}_{\tan}(|\pi_{\vec{z}_{\overline{S}}}\rangle, |\pi_{\vec{a}_{\overline{S}}}\rangle)^2 (\max_{i \in \overline{S}} |z_i||a_i|)^2 \\
&\leqslant \frac{1}{4}\mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}}\rangle)^2
\end{aligned}$$

where the first inequality follows from Lemma 3.9, the second inequality uses that $|z_i| \leqslant \mu$ for all $i \notin S$ and by assumption $|a_i^{(s)}| \leqslant 1/(2\mu)$ for all $i \notin S$. So, we can conclude that

$$\mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}}\rangle)^2 \geqslant (1.5b)^2 \implies \mathrm{d}_{\tan}(|\pi_{\vec{z}_S}\rangle, |\pi_{\vec{a}_S}\rangle)^2 + \|\vec{z}_{\overline{S}} - \vec{a}_{\overline{S}}\|_2^2 \geqslant \frac{3}{4}(1.5b)^2 \geqslant 1.5b^2,$$

$$\mathrm{d}_{\tan}(|\pi_{\vec{z}_S}\rangle, |\pi_{\vec{a}_S}\rangle)^2 + \|\vec{z}_{\overline{S}} - \vec{a}_{\overline{S}}\|_2^2 \geqslant (1.4 - \|\vec{a}_{\overline{S}}\|_\infty)b^2 \implies \mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}}\rangle)^2 \geqslant \frac{4}{5}(1.4 - \|\vec{a}_{\overline{S}}\|_\infty)b^2,$$

where the last line gives the desired $\geqslant b^2$ bound by our case assumption that $\|\vec{a}_{\overline{S}}\|_\infty \leqslant \frac{1}{2\mu} \leqslant 0.1$.

For the other case, suppose $|a_i| > \frac{1}{2\mu}$ for some $i \notin S$. We then argue that all inequalities are true, so the implications hold trivially.

$$\mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}}\rangle) \geqslant \left|\frac{z_i - a_i}{1 + z_i^* a_i}\right| \geqslant \left|\frac{\mu - \frac{1}{2\mu}}{1 + \frac{1}{2}}\right| \geqslant \frac{1}{3}\left|\mu - \frac{1}{\mu}\right| \geqslant 1.5b.$$

Similarly, using the same bounds, we have that

$$\mathrm{d}_{\tan}(|\pi_{\vec{z}_S}\rangle, |\pi_{\vec{a}_S}\rangle)^2 + \|\vec{z}_{\overline{S}} - \vec{a}_{\overline{S}}\|_2^2 \geqslant |z_i - a_i|^2 \geqslant \left(\frac{1}{2\mu} - \mu\right)^2 \geqslant \frac{1}{4}\left(\frac{1}{\mu} - \mu\right)^2 \geqslant 1.5b^2.$$

Thus, all of the stated inequalities are true. ♦

Lemma 5.17. *Consider a vector $\vec{z} \in \mathbb{C}^m$ and set $S \subseteq [m]$ such that $\|\vec{z}_{\overline{S}}\|_\infty \leqslant \frac{\sqrt{\varepsilon_{\text{approx}}}}{\|\vec{z}\|_2}$. Further, for an $\varepsilon_{\text{approx}} > 0$, let $d \geqslant 8\|\vec{z}\|_2^2 + \log \frac{2}{\varepsilon_{\text{approx}}}$, and let ϱ_d satisfy the guarantees of the output of Algorithm 5.10: $\|\varrho_d - \Pi_{\leqslant d}\Pi_{\leqslant d}\|_{\text{op}} \leqslant \varepsilon_{\text{approx}}$ for a PSD matrix ϱ_d with trace at most 1 and supported only on the image of $\Pi_{\leqslant d}$. Then*

$$|\langle \pi_{\vec{z}}| \rho |\pi_{\vec{z}}\rangle - p_{\vec{z}_S}(\vec{z}_{\overline{S}})| \leqslant 3\varepsilon_{\text{approx}},$$

$$\text{where } p_{\vec{z}_S}(\vec{z}_{\overline{S}}) = \frac{e^{-\|\vec{z}_{\overline{S}}\|_2^2}}{\prod_{i \in S}(1 + |z_i|^2)} \sum_{x, x' \in \{0,1\}^m} \langle x| \varrho_d |x'\rangle (\vec{z}^*)^x (\vec{z})^{x'}.$$

Proof. Fix a vector $\vec{z}$. Recall that $\rho_d = \Pi_{\leqslant d}\rho\Pi_{\leqslant d}$ is the unknown state truncated to strings of Hamming weight at most d, and ϱ_d is our estimate of ρ_d. We first show the following:

$$|\langle \pi_{\vec{z}}| \rho |\pi_{\vec{z}}\rangle - \langle \pi_{\vec{z}}| \rho_d |\pi_{\vec{z}}\rangle| \leqslant \varepsilon_{\text{approx}}.$$

To see this, we apply Lemma 3.10 to $\vec{z}$ to get the following:

$$\begin{aligned}
|\langle \pi_{\vec{z}}| \rho |\pi_{\vec{z}}\rangle - \langle \pi_{\vec{z}}| \rho_d |\pi_{\vec{z}}\rangle| &= \left|\mathrm{tr}\Big(\rho(|\pi_{\vec{z}}\rangle\langle\pi_{\vec{z}}| - \Pi_{<d} |\pi_{\vec{z}}\rangle\langle\pi_{\vec{z}}| \Pi_{<d})\Big)\right| \\
&\leqslant \||\pi_{\vec{z}}\rangle\langle\pi_{\vec{z}}| - \Pi_{<d} |\pi_{\vec{z}}\rangle\langle\pi_{\vec{z}}| \Pi_{<d}\|_{\text{op}} \\
&\leqslant \||\pi_{\vec{z}}\rangle\langle\pi_{\vec{z}}| - \Pi_{<d} |\pi_{\vec{z}}\rangle\langle\pi_{\vec{z}}|\|_{\text{op}} + \|\Pi_{<d} |\pi_{\vec{z}}\rangle\langle\pi_{\vec{z}}| - \Pi_{<d} |\pi_{\vec{z}}\rangle\langle\pi_{\vec{z}}| \Pi_{<d}\|_{\text{op}} \\
&\leqslant 2\||\pi_{\vec{z}}\rangle - \Pi_{<d} |\pi_{\vec{z}}\rangle\|_2 \\
&= 2\|\Pi_{\geqslant d} |\pi_{\vec{z}}\rangle\|_2 \\
&\leqslant 2e^{-d(\log(d/\|\vec{z}\|_2^2) - 1)}
\end{aligned}$$

Since we take $d \geqslant 8\|\vec{z}\|_2^2 + \log \frac{2}{\varepsilon_{\text{approx}}}$, this makes the final quantity smaller than $\varepsilon_{\text{approx}}$. Next, because $\|\varrho_d - \rho_d\|_{\text{op}} \leqslant \varepsilon_{\text{approx}}$, by the definition of operator norm we have

$$|\langle \pi_{\vec{z}}| \rho_d |\pi_{\vec{z}}\rangle - \langle \pi_{\vec{z}}| \varrho_d |\pi_{\vec{z}}\rangle| \leqslant \varepsilon_{\text{approx}}\,.$$

We can relate this quantity to $p_{\vec{z}_S}(\vec{z}_{\overline{S}})$ as follows:

$$\begin{aligned} p_{\vec{z}_S}(\vec{z}_{\overline{S}}) &= \frac{e^{-\|\vec{z}_{\overline{S}}\|_2^2}}{\prod_{i\in S}(1+|z_i|^2)} \sum_{x,x'\in\{0,1\}^m} \langle x|\, \varrho_d\, |x'\rangle\, (\vec{z}^*)^x \vec{z}^{x'} \\ &= \frac{e^{-\|\vec{z}_{\overline{S}}\|_2^2} \prod_{i\in[m]}(1+|z_i|)^2}{\prod_{i\in S}(1+|z_i|^2)} \langle \pi_{\vec{z}}|\, \varrho_d\, |\pi_{\vec{z}}\rangle \\ &= \left(e^{-\|\vec{z}_{\overline{S}}\|_2^2} \prod_{i\notin S}(1+|z_i|)^2\right) \langle \pi_{\vec{z}}|\, \varrho_d\, |\pi_{\vec{z}}\rangle\,. \end{aligned}$$

By Lemma 3.7, we have:

$$e^{-\sum_{i\notin S}|z_i|^4} \leqslant e^{-\|\vec{z}_{\overline{S}}\|_2^2} \prod_{i\notin S}(1+|z_i|)^2 \leqslant 1\,,$$

where we also know that

$$\sum_{i\notin S}|z_i|^4 \leqslant \|\vec{z}_{\overline{S}}\|_2^2 \|\vec{z}_{\overline{S}}\|_\infty^2 \leqslant \varepsilon_{\text{approx}}.$$

Since $1 - x \leqslant e^{-x}$, we have the following multiplicative error bounds on $p_{\vec{z}_S}(\vec{z}_{\overline{S}})$:

$$(1-\varepsilon_{\text{approx}})\, \langle \pi_{\vec{z}}|\, \varrho_d\, |\pi_{\vec{z}}\rangle \leqslant p_{\vec{z}_S}(\vec{z}_{\overline{S}}) \leqslant \langle \pi_{\vec{z}}|\, \varrho_d\, |\pi_{\vec{z}}\rangle\,.$$

As ϱ_d is a sub-normalized quantum state, $\varepsilon_{\text{approx}} \langle \pi_{\vec{z}}|\, \varrho_d\, |\pi_{\vec{z}}\rangle \leqslant \varepsilon_{\text{approx}}$, so $|\langle \pi_{\vec{z}}|\, \varrho_d\, |\pi_{\vec{z}}\rangle - p_{\vec{z}_S}(\vec{z}_{\overline{S}})| \leqslant \varepsilon_{\text{approx}}$. Applying the triangle inequality, we get the desired bound:

$$\begin{aligned} |\langle \pi_{\vec{z}}|\, \rho\, |\pi_{\vec{z}}\rangle - p_{\vec{z}_S}(\vec{z}_{\overline{S}})| &\leqslant |\langle \pi_{\vec{z}}|\, \rho\, |\pi_{\vec{z}}\rangle - \langle \pi_{\vec{z}}|\, \rho_d\, |\pi_{\vec{z}}\rangle| \\ &\quad + |\langle \pi_{\vec{z}}|\, \rho_d\, |\pi_{\vec{z}}\rangle - \langle \pi_{\vec{z}}|\, \varrho_d\, |\pi_{\vec{z}}\rangle| + |\langle \pi_{\vec{z}}|\, \varrho_d\, |\pi_{\vec{z}}\rangle - p_{\vec{z}_S}(\vec{z}_{\overline{S}})| \leqslant 3\varepsilon_{\text{approx}} \quad \blacklozenge \end{aligned}$$

We now prove correctness. We split the desired guarantee into two parts: soundness and completeness. First, we prove soundness.

Claim 5.18 (Soundness)**.** The output of Algorithm 5.10 satisfies the desired correctness criteria: the output is either $\perp$ or a $\vec{z} \in \mathbb{C}^m$ such that

(a) $\langle \pi_{\vec{z}}|\, \rho\, |\pi_{\vec{z}}\rangle \geqslant \eta - \varepsilon$;

(b) For all $s \in [r]$, $\mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}^{(s)}}\rangle) \geqslant b$.

Proof. From inspecting Algorithm 5.10, we can observe that the output $\vec{u}$ satisfies the following guarantees, for some internal parameters S, $\vec{v} \in V^{(S)}$, and $\nu \in [0, B]$. First, it satisfies the guarantees from being in the domain $\mathcal{D}^{2\varepsilon_{\text{tol}}}$:

$$|\|\vec{u}_{\overline{S}}\|_2 - \nu| \leqslant 2\varepsilon_{\text{tol}} \qquad \|\Pi_{V^{(S)}}\vec{u} - \vec{v}\|_2 \leqslant 2\varepsilon_{\text{tol}} \qquad \|\vec{u}_{\overline{S}}\|_\infty \leqslant \mu + 2\varepsilon_{\text{tol}} \tag{12}$$

Since we have that $\varepsilon_{\text{tol}} \leqslant 0.01B$ and $\nu \leqslant B$, this implies that $|\|\vec{u}_{\overline{S}}\|_2^2 - \nu^2| \leqslant 3B\varepsilon_{\text{tol}}$. The output $\vec{u}$ also satisfies the tangent distance guarantees inherited from the ε_{tol}-net, the bound $\|\vec{u}_S\|_2^2 + \nu^2 \leqslant B^2$ also from the net, and the guarantee from the objective function:

$$\nu^2 - \|\vec{v}_{\overline{S}}\|_2^2 + \|\vec{v}_{\overline{S}} - \vec{a}_{\overline{S}}^{(s)}\|_2^2 \geqslant 1.49b^2 - \mathrm{d}_{\tan}(|\pi_{\vec{u}_S}\rangle, |\pi_{\vec{a}_S^{(s)}}\rangle)^2 \text{ for all } s \in [r] \tag{13}$$

$$p_{\vec{u}_S,\nu}(\vec{u}_{\overline{S}}) \geqslant \eta - \varepsilon/2 \tag{14}$$

From (14), we can conclude (a). First, we relate $p_{\vec{u}_S,\nu}(\vec{u}_{\overline{S}}) = e^{\nu^2 - \|\vec{u}_{\overline{S}}\|_2^2} p_{\vec{u}_S}(\vec{u}_{\overline{S}})$ to $\langle \pi_{\vec{u}} | \rho | \pi_{\vec{u}} \rangle$:

$$\begin{aligned}
|\langle \pi_{\vec{u}} | \rho | \pi_{\vec{u}} \rangle - p_{\vec{u}_S,\nu}(\vec{u}_{\overline{S}})| &\leqslant |\langle \pi_{\vec{u}} | \rho | \pi_{\vec{u}} \rangle - p_{\vec{u}_S}(\vec{u}_{\overline{S}})| + |p_{\vec{u}_S}(\vec{u}_{\overline{S}}) - p_{\vec{u}_S,\nu}(\vec{u}_{\overline{S}})| \\
&\leqslant 3\varepsilon_{\text{approx}} + |1 - e^{\nu^2 - \|\vec{u}_{\overline{S}}\|_2^2}||p_{\vec{u}_S}(\vec{u}_{\overline{S}})| \\
&\leqslant 3\varepsilon_{\text{approx}} + |1 - e^{\nu^2 - \|\vec{u}_{\overline{S}}\|_2^2}|(1 + 3\varepsilon_{\text{approx}}) \\
&\leqslant 3\varepsilon_{\text{approx}} + 0.1\varepsilon.
\end{aligned}$$

Above, we use that $\varepsilon_{\text{tol}} \leqslant 0.01\varepsilon/B$ and Lemma 5.17; we satisfy the assumptions of the lemma since $\|\vec{u}\|_2 \leqslant B + 2\varepsilon_{\text{tol}} \leqslant 1.1B$ and $\|\vec{u}_{\overline{S}}\|_\infty \leqslant 1.1\mu \leqslant \frac{\sqrt{\varepsilon_{\text{approx}}}}{2B} \leqslant \frac{\sqrt{\varepsilon_{\text{approx}}}}{\|\vec{u}\|_2}$.

$$\langle \pi_{\vec{u}} | \rho | \pi_{\vec{u}} \rangle \geqslant (\eta - \varepsilon/2) + (\langle \pi_{\vec{u}} | \rho | \pi_{\vec{u}} \rangle - p_{\vec{u}_S}(\vec{u}_{\overline{S}})) \geqslant \eta - \varepsilon/2 - 3\varepsilon_{\text{approx}} - 0.1\varepsilon \geqslant \eta - \varepsilon.$$

To get (b), we work with the tangent distance constraints (13). Along with the domain constraints (12), these imply the following:

$$\mathrm{d}_{\tan}(|\pi_{\vec{u}_S}\rangle, |\pi_{\vec{a}_S^{(s)}}\rangle)^2 + \|\vec{u}_{\overline{S}} - \vec{a}_{\overline{S}}^{(s)}\|_2^2 \geqslant 1.49b^2 - 20\varepsilon_{\text{tol}}(B + \|\vec{a}_{\overline{S}}^{(s)}\|_2) \text{ for all } s \in [r] \tag{15}$$

This follows from the below argument, where we use triangle inequality, $\varepsilon_{\text{tol}} \leqslant 0.01$, and the Pythagorean theorem. Since $V^{(S)}$ contains the subspace corresponding to S, there is a corresponding projector Π such that $\Pi\vec{x}_{\overline{S}} = (\Pi_{V^{(S)}}\vec{x})_{\overline{S}}$ (so, in particular, $\|\Pi\vec{u}_{\overline{S}} - \vec{v}_{\overline{S}}\|_2 = \|(\Pi_{V^{(S)}}\vec{u} - \vec{v})_{\overline{S}}\|_2 \leqslant 2\varepsilon_{\text{tol}}$ and $\Pi\vec{a}_{\overline{S}}^{(s)} = \vec{a}_{\overline{S}}^{(s)}$ for all $s \in [r]$).

$$\begin{aligned}
&|\nu^2 - \|\vec{v}_{\overline{S}}\|_2^2 + \|\vec{v}_{\overline{S}} - \vec{a}_{\overline{S}}^{(s)}\|_2^2 - \|\vec{u}_{\overline{S}} - \vec{a}_{\overline{S}}^{(s)}\|_2^2| \\
&\leqslant 3B\varepsilon_{\text{tol}} + |\|\vec{u}_{\overline{S}}\|_2^2 - \|\vec{v}_{\overline{S}}\|_2^2 + \|\vec{v}_{\overline{S}} - \vec{a}_{\overline{S}}^{(s)}\|_2^2 - \|\vec{u}_{\overline{S}} - \vec{a}_{\overline{S}}^{(s)}\|_2^2| \\
&= 3B\varepsilon_{\text{tol}} + |(\|\Pi\vec{u}_{\overline{S}}\|_2^2 - \|\vec{v}_{\overline{S}}\|_2^2) + (\|\vec{v}_{\overline{S}} - \vec{a}_{\overline{S}}^{(s)}\|_2^2 - \|\Pi(\vec{u}_{\overline{S}} - \vec{a}_{\overline{S}}^{(s)})\|_2^2)| \\
&\leqslant 3B\varepsilon_{\text{tol}} + \|\Pi\vec{u}_{\overline{S}} - \vec{v}_{\overline{S}}\|_2(\|\Pi\vec{u}_{\overline{S}}\|_2 + \|\vec{v}_{\overline{S}}\|_2 + \|\vec{v}_{\overline{S}} - \vec{a}_{\overline{S}}^{(s)}\|_2 + \|\Pi\vec{u}_{\overline{S}} - \vec{a}_{\overline{S}}^{(s)}\|_2) \\
&\leqslant 20\varepsilon_{\text{tol}}(B + \|\vec{a}_{\overline{S}}^{(s)}\|_2)
\end{aligned}$$

Notice that we have no bound on the size of $\vec{a}^{(S)}$, so we cannot remove this dependence. Next, we use that ε_{tol} is sufficiently small (specifically, smaller than $\frac{0.001}{(1+B)\sqrt{m}}b^2$) to conclude from (15) that

$$\mathrm{d}_{\tan}(|\pi_{\vec{u}_S}\rangle, |\pi_{\vec{a}_S^{(s)}}\rangle)^2 + \|\vec{u}_{\overline{S}} - \vec{a}_{\overline{S}}^{(s)}\|_2^2 \geqslant (1.4 - \|\vec{a}_{\overline{S}}\|_\infty)b^2 \text{ for all } s \in [r].$$

Then, we can appeal to Lemma 5.16, since $\|\vec{u}_{\overline{S}}\|_\infty \leqslant \mu + 2\varepsilon_{\text{tol}} \leqslant \frac{1}{6}\min(b, \frac{1}{b})$ for every $s \in [r]$, to get that (b) is satisfied:

$$\mathrm{d}_{\tan}(|\pi_{\vec{u}}\rangle, |\pi_{\vec{a}^{(s)}}\rangle) \geqslant b \text{ for all } s \in [r]. \qquad \blacklozenge$$

Claim 5.19 (Completeness)**.** If there is a product state $|\pi\rangle$ such that

(a′) $\langle \pi | \rho | \pi \rangle \geqslant \eta$;

(b′) For all $s \in [r]$, $\mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{a}^{(s)}}\rangle) \geqslant 1.5b$;

(c′) $\mathrm{d}_{\tan}(|\pi\rangle, |\varphi\rangle) \leqslant B$;

then the output of Algorithm 5.10 is guaranteed to not be $\perp$.

Proof. We work in the basis where $|\varphi\rangle$ is rotated to $|0^m\rangle$ via single-qubit unitaries. Let $|\pi_{\vec{u}}\rangle$ be a product state satisfying (a'), (b'), and (c'). We will show that Algorithm 5.10 will, in its search, find some $\vec{\tilde{u}}$ close to $\vec{u}$ and which achieves a similarly large objective value. Thus, the algorithm will not output $\perp$.

First, let $S = \{i \mid |u_i| \geqslant \mu\}$. Using (c'), $\mathrm{d}_{\tan}(|\pi_{\vec{u}}\rangle, |0^m\rangle) = \|\vec{u}\|_2 \leqslant B$, so $\|u\|_2^2 \leqslant \sum_{i\in S} u_i^2/\mu^2 \leqslant B^2/\mu^2$. So, we can consider $\mathcal{F}_S$ for the corresponding choice of S. Our vector $\vec{u}$ has a corresponding point in the "feasible set", $(\Pi_{V^{(S)}}\vec{u}, \|\vec{u}_{\overline{S}}\|_2) \in \mathcal{F}_S$. The key constraint to check is the tangent distance constraint in the definition of $\mathcal{F}_S$ (10), which becomes

$$\begin{aligned}
&\|\vec{u}_{\overline{S}}\|_2^2 - \|(\Pi_{V^{(S)}}\vec{u})_{\overline{S}}\|_2^2 + \|(\Pi_{V^{(S)}}\vec{u})_{\overline{S}} - \vec{a}_{\overline{S}}^{(s)}\|_2^2 \geqslant 1.5b^2 - \mathrm{d}_{\tan}(|\pi_{\vec{u}_S}\rangle, |\pi_{\vec{a}_S^{(s)}}\rangle)^2 \\
\iff\ &\|\vec{u}_{\overline{S}}\|_2^2 - \|\Pi\vec{u}_{\overline{S}}\|_2^2 + \|\Pi(\vec{u}_{\overline{S}} - \vec{a}_{\overline{S}}^{(s)})\|_2^2 \geqslant 1.5b^2 - \mathrm{d}_{\tan}(|\pi_{\vec{u}_S}\rangle, |\pi_{\vec{a}_S^{(s)}}\rangle)^2 \\
\iff\ &\mathrm{d}_{\tan}(|\pi_{\vec{u}_S}\rangle, |\pi_{\vec{a}_S^{(s)}}\rangle)^2 + \|\vec{u}_{\overline{S}} - \vec{a}_{\overline{S}}^{(s)}\|_2^2 \geqslant 1.5b^2,
\end{aligned}$$

where as in Claim 5.18 we define Π to be the projector such that $\Pi\vec{z}_{\overline{S}} = (\Pi_{V^{(S)}}\vec{z})_{\overline{S}}$. By Lemma 5.16, (b') implies the above condition for all $s \in [r]$. So, by Claim 5.15, we can find a point $(\vec{v}, \nu) \in \mathcal{N}_S$ such that

$$\|\vec{v} - \Pi_{V^{(S)}}\vec{u}\|_2 \leqslant \varepsilon_{\mathrm{tol}}\,, \qquad |\|\vec{u}_{\overline{S}}\|_2 - \nu| \leqslant \varepsilon_{\mathrm{tol}}\,. \tag{16}$$

We now claim that the vector $\vec{u}_{\overline{S}} \in \mathcal{D}_{\varepsilon_{\mathrm{tol}}}$ as defined in (11). Thus, the vector $\vec{z} = (\vec{v}_S, \vec{u}_{\overline{S}})$ is a feasible solution to the optimization problem in (P0). To show this, observe that

$$\|\vec{z} - \vec{u}\|_2 = \|\vec{v}_S - \vec{u}_S\|_2 \leqslant \varepsilon_{\mathrm{tol}}$$

Here, we use the definition of the two norm squared as the sum over entries; that off of S, $\vec{z}$ and $\vec{u}$ are identical; and (16). From this, we can check that for the first constraint of $\mathcal{D}_{\varepsilon_{\mathrm{tol}}}$, we have that

$$|\|\vec{z}_{\overline{S}}\|_2 - \nu| \leqslant \varepsilon_{\mathrm{tol}}\,.$$

The second and third constraints of $\mathcal{D}_{\varepsilon_{\mathrm{tol}}}$ can also easily be verified:

$$\begin{aligned}
\|\Pi_{V^{(S)}}\vec{z} - \vec{v}\|_2 = \|\Pi_{V^{(S)}}(\vec{0}_S, (\vec{u} - \vec{v})_{\overline{S}})\|_2 \leqslant \|\Pi_{V^{(S)}}(\vec{u} - \vec{v})\|_2 \leqslant \varepsilon_{\mathrm{tol}} \\
\|\vec{z}_{\overline{S}}\|_\infty = \|\vec{u}_{\overline{S}}\|_\infty \leqslant \mu \leqslant \mu + \varepsilon_{\mathrm{tol}}.
\end{aligned}$$

In the first inequality in the first line line, we use that the subspace $V^{(S)}$ contains the subspace spanned by S, so the projector onto $\overline{S}$ commutes with $\Pi_{V^{(S)}}$. To finish the completeness proof, we now need to show that, for the above choice of $\vec{z}$, $p_{\vec{z}_S,\nu}(\vec{z}_{\overline{S}}) \geqslant \eta - \varepsilon/2 + \varepsilon_{\mathrm{approx}}$. Then, the best candidate solution $\vec{y}$ found by the algorithm must have an objective value of at least $\eta - \varepsilon/2$, and thus the algorithm will not output $\perp$. Because $\vec{z}$ and $\vec{u}$ satisfy $\|\vec{z} - \vec{u}\|_2 \leqslant \varepsilon_{\mathrm{tol}}$, we have the following bound:

$$\begin{aligned}
|\langle\pi_{\vec{z}}|\,\rho\,|\pi_{\vec{z}}\rangle - \langle\pi_{\vec{u}}|\,\rho\,|\pi_{\vec{u}}\rangle| &\leqslant \||\pi_{\vec{z}}\rangle\langle\pi_{\vec{z}}| - |\pi_{\vec{u}}\rangle\langle\pi_{\vec{u}}|\|_{\mathrm{op}} \\
&\leqslant \mathrm{d}_{\tan}(|\pi_{\vec{z}}\rangle, |\pi_{\vec{u}}\rangle) \\
&= \mathrm{d}_{\tan}(|\pi_{\vec{z}_{\overline{S}}}\rangle, |\pi_{\vec{u}_{\overline{S}}}\rangle) \\
&\leqslant \|\vec{z}_{\overline{S}} - \vec{u}_{\overline{S}}\|_2(1 + \max_{i\in\overline{S}}(|z_i||u_i|)) \\
&\leqslant \varepsilon_{\mathrm{tol}}(1 + \mu(\mu + \varepsilon_{\mathrm{tol}})) \\
&\leqslant 2\varepsilon_{\mathrm{tol}}
\end{aligned}$$

In the first line, we use the definition of the trace distance. Then we use Corollary 3.8, and then the definition of tangent distance to restrict to the coordinates where $\vec{z}$ and $\vec{u}$ are not equal. Then, we use Lemma 3.9 and then our bound $\|u_{\overline{S}}\|_\infty \leqslant \mu$. We now relate the fidelity to the objective function for $\vec{z}$ through Lemma 5.17, similarly to as in Claim 5.18. Recall that $p_{\vec{u}_S,\nu}(\vec{u}_{\overline{S}}) = e^{\nu^2 - \|\vec{u}_{\overline{S}}\|_2^2} p_{\vec{u}_S}(\vec{u}_{\overline{S}})$, so

$$\begin{aligned}
|\langle \pi_{\vec{u}}| \rho |\pi_{\vec{u}}\rangle - p_{\vec{u}_S,\nu}(\vec{u}_{\overline{S}})| &\leqslant |\langle \pi_{\vec{u}}| \rho |\pi_{\vec{u}}\rangle - p_{\vec{u}_S}(\vec{u}_{\overline{S}})| + |p_{\vec{u}_S}(\vec{u}_{\overline{S}}) - p_{\vec{u}_S,\nu}(\vec{u}_{\overline{S}})| \\
&\leqslant 3\varepsilon_{\text{approx}} + |1 - e^{\nu^2 - \|\vec{u}_{\overline{S}}\|_2^2}||p_{\vec{u}_S}(\vec{u}_{\overline{S}})| \\
&\leqslant 3\varepsilon_{\text{approx}} + |1 - e^{\nu^2 - \|\vec{u}_{\overline{S}}\|_2^2}|(1 + 3\varepsilon_{\text{approx}}) \\
&\leqslant 3\varepsilon_{\text{approx}} + 0.1\varepsilon.
\end{aligned}$$

Above, we use (16), $\varepsilon_{\text{tol}} \leqslant 0.01\varepsilon/B$, and Lemma 5.17; we satisfy the assumptions since $\|\vec{u}\|_2 \leqslant B + 2\varepsilon_{\text{tol}} \leqslant 1.1B$ and $\|\vec{u}_{\overline{S}}\|_\infty \leqslant \mu \leqslant \frac{\sqrt{\varepsilon_{\text{approx}}}}{2B} \leqslant \frac{\sqrt{\varepsilon_{\text{approx}}}}{\|\vec{u}\|_2}$. Altogether, we have that

$$\begin{aligned}
p_{\vec{z}_S,\nu}(\vec{z}_{\overline{S}}) &\geqslant \langle \pi_{\vec{z}}| \rho |\pi_{\vec{z}}\rangle - 3\varepsilon_{\text{approx}} - 0.1\varepsilon \\
&\geqslant \langle \pi_{\vec{u}}| \rho |\pi_{\vec{u}}\rangle - 3\varepsilon_{\text{approx}} - 0.1\varepsilon - 2\varepsilon_{\text{tol}} \\
&\geqslant \eta - \varepsilon/2 + \varepsilon_{\text{approx}}.
\end{aligned}$$

Thus, Algorithm 5.10 does not output $\perp$. ♦

6 Polynomial optimization

In this section, we provide an algorithm to solve the polynomial optimization problem subject to subspace constraints obtained in Section 5. We note that while optimizing worst-case polynomial systems is hard, we are working in the regime where the polynomial when viewed as a tensor has Frobenius norm bounded by 1. This regime is reminiscent of optimizing dense CSP's, where additive error approximations suffice. In contrast, we are optimizing the polynomial over the sphere, and do not require appealing to regularity-like statements. The key technical contribution in this section is to show that our polynomial optimization problem with subspace constraints admits a small ε-net.

Definition 6.1 (Polynomial notation)**.** Let $T^{(0)} \in \mathbb{C}, T^{(1)} \in (\mathbb{C}^n)^{\otimes 2}, \ldots, T^{(d)} \in (\mathbb{C}^n)^{\otimes 2d}$ be tensors. For $\vec{x} \in \mathbb{C}^n$, we denote

$$f_{T^{(0)},\ldots,T^{(d)}}(\vec{x}) = T^{(0)} + \langle T^{(1)}, \vec{x}^* \otimes \vec{x}\rangle + \cdots + \langle T^{(d)}, (\vec{x}^*)^{\otimes d} \otimes \vec{x}^{\otimes d}\rangle$$

Definition 6.2 (Domain with subspace and flatness constraints)**.** For a matrix $A \in \mathbb{C}^{r\times n}$ and vector $\vec{v} \in \mathbb{R}^r$, parameters ν, μ, and tolerance γ, we define the set $\mathcal{D}^\gamma_{\vec{v},A,\nu,\mu} = \{\vec{x} \in \mathbb{C}^n : |\|\vec{x}\|_2 - \nu| \leqslant \gamma, \|A\vec{x} - \vec{v}\|_2 \leqslant \gamma, \|\vec{x}\|_\infty \leqslant \mu + \gamma\}$.

The main theorem that we will prove is stated below.

Theorem 6.3 (Polynomial optimization over a subspace with flat vectors)**.** *Let $T^{(k)} \in (\mathbb{C}^n)^{\otimes 2k}$ be tensors for all $k = 0, 1, \ldots, d$, and assume $\sum_{k=0}^d \|T_k\|_F \leqslant 1$. Let $A \in \mathbb{C}^{r\times n}$ be a matrix with $\|A\|_{\text{op}} \leqslant 1$ and let $\vec{v} \in \mathbb{C}^r$ be a specified vector. Given positive ν, μ, ε such that $\nu \leqslant 1$ and $\varepsilon \leqslant 1$, let $\gamma = \text{poly}(1/n, 1/d, \varepsilon)$. Then, there is an algorithm that runs in time $(1/\gamma)^{\text{poly}(d,r,1/\varepsilon,1/\mu)}$ that outputs either an $x \in \mathcal{D}^{2\gamma}_{\vec{v},A,\nu,\mu}$ (or $\perp$ if $\mathcal{D}^{2\gamma}_{\vec{v},A,\nu,\mu}$ is empty). The output satisfies*

$$|f_{T^{(0)},\ldots,T^{(d)}}(\vec{x})| \geqslant \max_{\vec{y}\in\mathcal{D}^\gamma_{\vec{v},A,\nu,\mu}} |f_{T^{(0)},\ldots,T^{(d)}}(\vec{y})| - \varepsilon\,.$$

Now we describe the algorithm for solving the polynomial optimization problem.

Algorithm 6.4 (Polynomial optimization under product state cover constraints)**.**

Input: Tensors $T^{(0)} \in \mathbb{C}, \ldots, T^{(d)} \in (\mathbb{C}^n)^{\otimes 2d}$, matrix $A \in \mathbb{C}^{r \times n}$, accuracy parameter $0 < \varepsilon < 1$, bound $0 \leqslant \nu \leqslant 1$.

Output: A vector $\vec{x}$ such that

$$|f_{T^{(0)},\ldots,T^{(d)}}(\vec{x})| \geqslant \max_{\vec{y} \in \mathcal{D}^{\gamma}_{\vec{v},A,\nu,\mu}} |f_{T^{(0)},\ldots,T^{(d)}}(\vec{y})| - \varepsilon$$

where $\mathcal{D}^{\gamma}_{\vec{v},A,\nu,\mu}$ is defined in Definition 6.2 and $\gamma = \mathrm{poly}(1/n, 1/d, \varepsilon)$.

Operation:

1: **for** $j \in [d]$ **do**
2: **for** $i \in [2j]$ **do**
3: Let $M_{j,i}$ be the $n \times n^{2j-1}$ flattening of $T^{(j)}$ along the ith mode.
4: Compute the SVD of $M_{j,i}$ and let $W_{j,i}$ be the subspace corresponding to singular values that are at least $\varepsilon/(d+1)^2$.
5: Let W be the subspace corresponding to the combined span of $\{W_{j,i}, W^*_{j,i}\}_{j \in [d], i \in [2j]}$
6: Let W' be the combined span of W and the rows of A
7: Let $k = O(1/\mu^2)$.
8: **for** $S \in [\binom{n}{k}]$ **do**
9: Let Z_S be the subspace corresponding to the k coordinate vectors indexed by S.
10: Construct the net $\mathcal{N}_{W',Z_S,\gamma}$ by taking a γ-net of the ball of radius $1+\gamma$ in the subspace spanned by W' and Z_S and removing all elements that are not in $\mathcal{D}^{2\gamma}_{\vec{v},A,\nu,\mu}$
11: Construct a larger net by concatenating all the nets above, i.e. let $\mathcal{N}_\gamma = \cup_{S \in [\binom{n}{k}]} \mathcal{N}_{W',Z_S,\gamma}$.
12: Output

$$\vec{x} = \max_{\vec{y} \in \mathcal{N}_\gamma} |f_{T^{(0)},\ldots,T^{(d)}}(\vec{y})|$$

which can be computed by iterating over each $\vec{y} \in \mathcal{N}_\gamma$.

We begin by showing that the function $f_{T^{(0)},\ldots,T^{(d)}}(\vec{y})$ essentially only depends on the projection of $\vec{y}$ onto some constant-dimensional subspace W, up to additive error ε.

Lemma 6.5 (Effective dimension)**.** *Let $T^{(0)} \in \mathbb{C}, T^{(1)} \in (\mathbb{C}^n)^{\otimes 2}, \ldots, T^{(d)} \in (\mathbb{C}^n)^{\otimes 2d}$ be tensors and assume $\|T^{(0)}\|_F, \ldots, \|T^{(d)}\|_F \leqslant 1$. For any parameter $\varepsilon > 0$, the subspace W computed in line 5 of Algorithm 6.4 has dimension at most $8(d+1)^6/\varepsilon^2$ and satisfies that for any $\vec{y} \in \mathbb{C}^n$ with $\|\vec{y}\|_2 \leqslant 1$,*

$$|f_{T^{(0)},\ldots,T^{(d)}}(\vec{y}) - f_{T^{(0)},\ldots,T^{(d)}}(\Pi_W \vec{y})| \leqslant \varepsilon\,,$$

where Π_W is the orthogonal projection matrix for the subspace W.

Proof. Consider $j \in [d]$ and a fixed tensor, $T^{(j)} \in (\mathbb{C}^n)^{\otimes 2j}$ such that $\|T^{(j)}\|_F \leqslant 1$. Let $M_{j,i} \in \mathbb{C}^{n \times n^{d-1}}$ be the flattening of $T^{(j)}$ into a $n \times n^{d-1}$ matrix along the ith mode and let $\sigma =$

$(\sigma_{1,i}, \ldots, \sigma_{n,i})$ be the vector of singular values of $M_{j,i}$. Since $\|M_{j,i}\|_F^2 \leqslant 1$, it follows that $\|\sigma\|_2^2 \leqslant 1$ and therefore there are at most $4(d+1)^4/\varepsilon^2$ singular values that are at least $\varepsilon/(2(d+1)^2)$. Now $W_{j,i}$ is the subspace corresponding to the large singular values. We can bound

$$\begin{aligned}
&\langle T^{(j)}, (\vec{y}^*)^{\otimes j} \otimes \vec{y}^{\otimes j} \rangle - \langle T^{(j)}, (\Pi_W^* \vec{y}^*)^{\otimes j} \otimes \vec{y}^{\otimes j} \rangle \\
&= \sum_{i=1}^{j} \Big(\langle T^{(j)}, (\vec{y}^*)^{\otimes j-i+1} \otimes (\Pi_W^* \vec{y}^*)^{\otimes i-1} \otimes \vec{y}^{\otimes j} \rangle - \langle T^{(j)}, (\vec{y}^*)^{\otimes j-i} \otimes (\Pi_W^* \vec{y}^*)^{\otimes i} \otimes \vec{y}^{\otimes j} \rangle \Big) \\
&= \sum_{i=1}^{j} (\vec{y}^* - \Pi_W^* \vec{y}^*)^\top M_{j,i} \mathsf{vec}\Big((\vec{y}^*)^{\otimes j-i} \otimes (\Pi_W^* \vec{y}^*)^{\otimes i-1} \otimes \vec{y}^{\otimes j} \Big) \\
&\leqslant \frac{\varepsilon}{2(d+1)},
\end{aligned}$$

where the last inequality follows from observing that $(\vec{y}^* - \Pi_W^* \vec{y}^*)$ is orthogonal to the subspace W^* and since W^* contains $W_{j,i}$, this vector is also orthogonal to $W_{j,i}$. Similarly, we have

$$\langle T^{(j)}, (\Pi_W^* \vec{y}^*)^{\otimes j} \otimes \vec{y}^{\otimes j} \rangle - \langle T^{(j)}, (\Pi_W^* \vec{y}^*)^{\otimes j} \otimes (\Pi_W \vec{y})^{\otimes j} \rangle \leqslant \frac{\varepsilon}{2(d+1)}.$$

Now we can repeat the above argument for all of the tensors $T^{(1)}, \ldots, T^{(d)}$ and use triangle inequality to get the desired bound. Since W is the union of the spans of $W_{j,i}, W_{j,i}^*$, its dimension is at most $8(d+1)^6/\varepsilon^2$, as desired. ♦

Next, we show a structural statement that for sets of the form $\mathcal{D}^{\gamma}_{\vec{v},A,\nu,\mu}$, if they are nonempty, then they contain a feasible point that is a linear combination of the rows of A and a sparse vector.

Lemma 6.6 (Structure of the optimizer)**.** *Let $A \in \mathbb{C}^{r \times n}$ be a matrix. Also assume we are given parameters $0 < \nu \leqslant 1, \mu, \gamma > 0$ and $\vec{v} \in \mathbb{C}^r$. If $\mathcal{D}^{\gamma}_{\vec{v},A,\nu,\mu}$ is nonempty, then there exists some $\vec{x} \in \mathcal{D}^{\gamma}_{\vec{v},A,\nu,\mu}$ that can be written as a sum of two vectors $\vec{u} + \vec{s}$ where $\vec{u}$ is in the row subspace of A and $\vec{s}$ is $1/\mu^2 + 1$-sparse.*

Proof. We may assume $r + 1/\mu^2 < n$ as otherwise the statement is trivially true.

Consider the solution $\vec{x}^{(0)} \in \mathcal{D}^{\gamma}_{\vec{v},A,\nu,\mu}$ that is lexicographically maximal in coordinate magnitude, i.e. it lexicographically maximizes the sequence $|x_1^{(0)}|, \ldots, |x_n^{(0)}|$. Note that the set $\mathcal{D}^{\gamma}_{\vec{v},A,\nu,\mu}$ is closed so this $\vec{x}^{(0)}$ is well-defined.

Let $S \subseteq [n]$ be the set of coordinates j such that $|x_j^{(0)}| = \mu + \gamma$. Let V be the subspace spanned by the rows of A and $\{\vec{e}_j\}_{j \in S}$ (where $\vec{e}_j$ are the standard basis vectors).

Let $T \subseteq [n]$ be the set of coordinates j' such that $\vec{e}_{j'}$ is in V. If $|T| = n$, then we are trivially done since $|S| \leqslant 1/\mu^2$ and the rows of A and $\{\vec{e}_j\}_{j \in S}$ would span all of $\mathbb{C}^n$. Now assume that $|T| < n$ so there is some index that is not in T. Let j_0 be the lexicographically minimum index such that $j_0 \notin T$. We attempt to construct a vector $\vec{\Delta} \in \mathbb{C}^n$ that is orthogonal to V and $\vec{x}^{(0)}$ and has nonzero coordinate on j_0. If such a vector does not exist, then this implies that $\vec{e}_{j_0}$ is in the span of V and $\vec{x}^{(0)}$. Since by assumption, $\vec{e}_{j_0}$ is not in the span V, this also implies that $\vec{x}^{(0)}$ is in the span of V and $\vec{e}_{j_0}$ which then immediately implies the desired statement. Now it remains to consider the case when such a vector $\vec{\Delta}$ exists. Now consider replacing $\vec{x}^{(0)}$ with the vector

$$\vec{x}' = \Pi_V \vec{x}^{(0)} + \frac{\|(I - \Pi_V)\vec{x}^{(0)}\|_2}{\sqrt{\|(I - \Pi_V)\vec{x}^{(0)}\|_2^2 + |z|^2}} \cdot ((I - \Pi_V)\vec{x}^{(0)} + z\vec{\Delta})$$

for a complex number z. Note that $\|\vec{x}'\|_2 = \|\vec{x}^{(0)}\|_2$ since $\langle \vec{\Delta}, (I - \Pi_V)\vec{x}^{(0)} \rangle = \langle \Delta, \vec{x}^{(0)} \rangle = 0$ so

$$\|(I - \Pi_V)\vec{x}^{(0)} + z\vec{\Delta}\|_2 = \sqrt{\|(I - \Pi_V)\vec{x}^{(0)}\|_2^2 + |z|^2}\,.$$

Also, the projection onto the subspace V is unchanged so $A\vec{x}' = A\vec{x}^{(0)}$ and $\vec{x}'$ and $\vec{x}^{(0)}$ match on all coordinates in S. Thus, there is some positive δ such that the above vector is in $\mathcal{D}^{\gamma}_{\vec{v},A,\nu,\mu}$ for all complex numbers z with $|z| \leqslant \delta$. If $\|(I - \Pi_V)\vec{x}^{(0)}\|_2 > 0$, then there would be some choice of z that increases the magnitude of $x^{(0)}_{j_0}$ (without changing any of the coordinates with indices smaller than j_0) and this contradicts the maximality of $\vec{x}^{(0)}$. Thus, we must actually have $\Pi_V \vec{x}^{(0)} = \vec{x}^{(0)}$ meaning that $\vec{x}^{(0)}$ is in the span of the rows of A and $\{\vec{e}_j\}_{j \in S}$ which immediately gives the desired property. ♦

Combining the two lemmas above suffices to show that our optimization problem admits a small net.

Proof of Theorem 6.3. Using Lemma 6.5, we can ensure that the net $\mathcal{N}_\gamma$ described in Algorithm 6.4 is of size $(n/\gamma)^{\mathrm{poly}(d,r,1/\varepsilon,1/\mu)}$ and can be constructed efficiently via a greedy procedure. It remains to show that this net must contain a vector with $\vec{x}$ such that $|f_{T^{(0)},\ldots,T^{(d)}}(\vec{x})| \geqslant \max_{\vec{y} \in \mathcal{D}^{\gamma}_{\vec{v},A,\nu,\mu}} |f_{T^{(0)},\ldots,T^{(d)}}(\vec{y})| - \varepsilon$.

Note that for $\vec{x}, \vec{x}' \in \mathbb{C}^n$ with $\|\vec{x}\|_2 \leqslant 1 + 2\gamma$, $\|\vec{x} - \vec{x}'\|_2 \leqslant 2\gamma$, we have

$$|f_{T^{(0)},\ldots,T^{(d)}}(\vec{x}) - f_{T^{(0)},\ldots,T^{(d)}}(\vec{x}')| \leqslant \varepsilon\,.$$

By Lemma 6.5, it suffices to argue that when we project all points in $\mathcal{N}_\gamma$ onto W, the points form a 2γ-net of the projection of $\mathcal{D}^{\gamma}_{\vec{v},A,\nu,\mu}$ onto W.

Consider any point $\vec{y} \in \mathcal{D}^{\gamma}_{\vec{v},A,\nu,\mu}$. Let M be a matrix whose rows form an orthonormal basis for W and let B be the matrix obtained by stacking A and M and let $\vec{v}'$ be the vector obtained by stacking $\vec{v}$ and $M\vec{y}$. Then by construction, $\vec{y}$ is an element of the set $\mathcal{D}^{\gamma}_{\vec{v}',B,\nu,\mu}$ — in particular, $\mathcal{D}^{\gamma}_{\vec{v}',B,\nu,\mu}$ is nonempty. Lemma 6.6 then implies that there is some element of $\vec{y}' \in \mathcal{D}^{\gamma}_{\vec{v}',B,\nu,\mu}$ that can be written as the sum of a vector in the combined span of A and W and a $O(1/\mu^2)$-sparse vector. The construction of $\mathcal{N}_\gamma$ then implies that there is some $\vec{y}'' \in \mathcal{N}_\gamma$ such that $\|\vec{y}' - \vec{y}''\|_2 \leqslant \gamma$ — note this is because $\vec{y}' \in \mathcal{D}^{\gamma}_{\vec{v}',B,\nu,\mu} \subseteq \mathcal{D}^{\gamma}_{\vec{v},A,\nu,\mu}$ so the entire γ-radius ball around $\vec{y}'$ is contained in $\mathcal{D}^{2\gamma}_{\vec{v},A,\nu,\mu}$ and thus when constructing the net in line 10 of Algorithm 6.4, we never remove any relevant points. Then,

$$\|\Pi_W \vec{y}'' - \Pi_W \vec{y}\|_2 \leqslant \|\Pi_W \vec{y}'' - \Pi_W \vec{y}'\|_2 + \|\Pi_W \vec{y}' - \Pi_W \vec{y}\|_2 \leqslant 2\gamma$$

and this shows that $\mathcal{N}_\gamma$, when projected onto W, forms a 2γ-net of the projection of $\mathcal{D}^{\gamma}_{\vec{v},A,\nu,\mu}$ onto W. Thus $\mathcal{N}_\gamma$ must contain an ε-approximate maximizer and we are done.

♦

7 Hardness

Definition 7.1. The spectral norm of a tensor $T \in \mathbb{C}^{n \times n \times n \times n}$ is defined as follows.

$$\|T\|_{\mathrm{op}} = \max_{\vec{x},\vec{y},\vec{u},\vec{v} \in \mathbb{C}^n} \frac{|\langle T, \vec{x} \otimes \vec{y} \otimes \vec{u} \otimes \vec{v} \rangle|}{\|\vec{x}\|_2 \|\vec{y}\|_2 \|\vec{u}\|_2 \|\vec{v}\|_2}\,.$$

Note that we define spectral norm to be maximizing over vectors with complex entries.

Theorem 7.2. *It is NP-hard to approximate the spectral norm of an $n \times n \times n \times n$ tensor T to within additive error $\frac{\|T\|_F}{100n^4}$.*

Proof. This is essentially shown in [FL17, Theorem 8.6], but the theorem statement did not contain quantitative bounds. We will thus re-prove it here.

For an undirected graph $G = (V, E)$ on n vertices with at least one edge, define the tensor $A_G = \sum_{(s,t)\in E} A^{(st)}$ where $A^{(st)} \in \mathbb{C}^{n\times n\times n\times n}$ is the tensor where the (i, j, k, l)th entry is $1/2$ if and only if i, j, k, l is some permutation of two s's and two t's:

$$A^{(s,t)}_{ijkl} = \begin{cases} 1/2 & i = s,\ j = t,\ k = s,\ l = t \\ 1/2 & i = t,\ j = s,\ k = t,\ l = s \\ 1/2 & i = s,\ j = t,\ k = t,\ l = s \\ 1/2 & i = t,\ j = s,\ k = s,\ l = t \\ 0 & \text{otherwise.} \end{cases}$$

By [FL17, Theorem 8.4], $\|A_G\|_{\mathrm{op}} = \frac{\kappa(G)-1}{\kappa(G)}$, where $\kappa(G) \in [n]$ is the clique number of G. The clique number is NP-hard to compute [Kar72], and if we have an estimate ν such that $|\nu - \|A_G\|_{\mathrm{op}}|$ to $\frac{1}{100n^2}$ error, then $|\frac{1}{1-\nu} - \kappa(G)| = |\frac{1}{1-\nu} - \frac{1}{1-\|A_G\|_{\mathrm{op}}}| < \frac{1}{2}$, so we can determine $\kappa(G)$ by computing $\frac{1}{1-\nu}$ and rounding to the nearest integer. Thus, it is NP-hard to compute $\|A_G\|_{\mathrm{op}}$ to $\frac{1}{100n^2}$ error. To conclude, observe that $\|A_G\|_F \leqslant n^2$. ⧫

The main theorem that we will prove in this section is the following.

Theorem 7.3. *Given an algorithm for agnostically learning product states such that for any n qubit state and target error ε, the algorithm has sample complexity and running time $f(n, \varepsilon)$ for some function f and succeeds with probability 0.99, we can give a quantum algorithm for approximating the spectral norm of a $m \times m \times m \times m$ tensor T to additive error $\varepsilon\|T\|_F$ with running time $f(\mathrm{poly}(m/\varepsilon), \mathrm{poly}(\varepsilon))$ that succeeds with probability 0.9.*

Definition 7.4. Assume we are given a tensor $T \in \mathbb{C}^{n\times n\times n\times n}$ with $\|T\|_F = 1$. Then we construct a quantum state on $4n$ qubits as follows.

$$|\psi_T\rangle = \sum_{i,j,k,l} T'_{ijkl} \, |e_i e_j e_k e_l\rangle \,,$$

where for $i, j, k, l \in [n]$, $|e_i e_j e_k e_l\rangle = |e_i\rangle \otimes |e_j\rangle \otimes |e_k\rangle \otimes |e_l\rangle$, where we recall that $|e_i\rangle$ is the product state which is $|1\rangle$ on the ith qubit and $|0\rangle$ on the other $n - 1$ qubits.

One can verify that this is a valid quantum state, since $\||\psi_T\rangle\|_2 = \|T\|_F = 1$. We will consider the quantum state corresponding to $U^{\otimes 4}T$, where $U \in \mathbb{C}^{n\times m}$ is a Haar-random matrix with orthonormal columns and $T \in \mathbb{C}^{m\times m\times m\times m}$ for some $m \leqslant n$. Here, we are randomly embedding the tensor into a larger, n-dimensional space.

Claim 7.5. For $T \in \mathbb{C}^{m\times m\times m\times m}$ and a Haar-random $U \in \mathbb{C}^{n\times m}$, as long as $m \geqslant 10 \log n$, then with 0.99 probability over the randomness of U, $\max_{i,j,k,l\in[n]} |(U^{\otimes 4}T)_{ijkl}| \leqslant \frac{(10m)^2}{n^2}$.

Proof. Note that the columns of U determine a random m-dimensional subspace of $\mathbb{C}^n$. With 0.99 probability over the randomness of U, all unit vectors in the subspace spanned by the columns U have entries with magnitude at most $\sqrt{10m/n}$. If this happens, then since T' is in the subspace spanned by the columns of $U \otimes U \otimes U \otimes U$ and $\|T'\|_F = 1$, we must have $\max_{i,j,k,l\in[n]} |T'_{ijkl}| \leqslant \frac{(10m)^2}{n^2}$. ⧫

Lemma 7.6. *Given a tensor $T \in \mathbb{C}^{n\times n\times n\times n}$ with $\|T\|_F = 1$ and $M = \max_{i,j,k,l} n^2|T_{ijkl}|$. Let OPT_T be the spectral norm of T and let $\mathsf{OPT}_{|\psi_T\rangle}$ be defined as:*

$$\mathsf{OPT}_{|\psi_T\rangle} = \max_{\sigma=|\sigma_1\rangle\otimes\cdots\otimes|\sigma_n\rangle} |\langle\sigma|\psi_T\rangle|.$$

Then, for sufficiently large n,

$$e^{-2}\mathsf{OPT}_T - \frac{10M}{n^{0.2}} \leqslant \mathsf{OPT}_{|\psi_T\rangle} \leqslant e^{-2}\mathsf{OPT}_T + \frac{10}{n^{0.1}} + \frac{10M}{n^{0.2}}.$$

Proof. First we prove the lower bound. Let $\vec{x},\vec{y},\vec{u},\vec{v} \in \mathbb{C}^n$ be unit vectors such that $|\langle T, \vec{x}\otimes\vec{y}\otimes\vec{u}\otimes\vec{v}\rangle| = \mathsf{OPT}_T$. Let $\vec{x}',\vec{y}',\vec{u}'$, and $\vec{v}'$ be obtained by zeroing out all entries have magnitude larger than $1/n^{0.1}$.

$$\begin{aligned}|\mathsf{OPT}_T - \langle T, \vec{x}'\otimes\vec{y}'\otimes\vec{u}'\otimes\vec{v}'\rangle| &= |\langle T, \vec{x}\otimes\vec{y}\otimes\vec{u}\otimes\vec{v} - \vec{x}'\otimes\vec{y}'\otimes\vec{u}'\otimes\vec{v}'\rangle| \\ &\leqslant \frac{100M}{n^{0.4}}.\end{aligned}$$

The last inequality follows from observing that the right-hand side of the inner product has Frobenius norm bounded by 2 and is non-zero in at most $4n^{3.2}$ entries. Thus, we can apply Cauchy–Schwarz on the non-zero entries, and the Frobenius norm on T restricted to such entries is at most $\frac{M}{n^2}\cdot\sqrt{4n^{3.2}}$. Then the product state $|\pi_{\vec{x}',\vec{y}',\vec{u}',\vec{v}'}\rangle$ satisfies, for $\xi = (\prod_{a=1}^n(1+|x_i'|^2)(1+|y_i'|^2)(1+|u_i'|^2)(1+|v_i'|^2))^{-1/2}$,

$$|\langle\psi_T|\pi\rangle| = \xi\Big|\sum_{i,j,k,l\in[n]}(T_{ijkl})^* x_i' y_j' u_k' v_l'\Big| = \xi|\langle T, \vec{x}'\otimes\vec{y}'\otimes\vec{u}'\otimes\vec{v}'\rangle| \geqslant \xi\big(\mathsf{OPT}_T - \frac{10M}{n^{0.4}}\big).$$

Finally, by Lemma 3.7, $\xi \geqslant e^{-\frac{1}{2}(\|\vec{x}'\|_2^2+\|\vec{y}'\|_2^2+\|\vec{u}'\|_2^2+\|\vec{v}'\|_2^2)} \geqslant e^{-2}$.

Now we prove the upper bound. Let $\vec{x},\vec{y},\vec{u},\vec{v} \in \mathbb{C}^n$ be vectors attaining the optimum fidelity, $|\langle\psi_T|\pi_{\vec{x},\vec{y},\vec{u},\vec{v}}\rangle| = \mathsf{OPT}_{|\psi_T\rangle}$. Here, we interpret the $4n$-length product state as having 4 parameter vectors of length n. We have

$$\langle\psi_T|\pi_{\vec{x},\vec{y},\vec{u},\vec{v}}\rangle = \xi\Big(\sum_{i,j,k,l\in[n]}(T_{ijkl})^* x_i y_j u_k v_l\Big) = \xi\langle T, \vec{x}\otimes\vec{y}\otimes\vec{u}\otimes\vec{v}\rangle,$$

where $\xi = (\prod_{a=1}^n(1+|x_a|^2)(1+|y_a|^2)(1+|u_a|^2)(1+|v_a|^2))^{-1/2}$. Let $\vec{x}',\vec{y}',\vec{u}',\vec{v}'$ be obtained by zeroing out entries larger than $1/n^{0.1}$. Then by the same argument as above,

$$\begin{aligned}|\langle T, \vec{x}\otimes\vec{y}\otimes\vec{u}\otimes\vec{v}\rangle - \langle T, \vec{x}'\otimes\vec{y}'\otimes\vec{u}'\otimes\vec{v}'\rangle| &= |\langle T, \vec{x}\otimes\vec{y}\otimes\vec{u}\otimes\vec{v} - \vec{x}'\otimes\vec{y}'\otimes\vec{u}'\otimes\vec{v}'\rangle| \\ &\leqslant \frac{10M}{n^{0.4}}\|\vec{x}\|_2\|\vec{y}\|_2\|\vec{u}\|_2\|\vec{v}\|_2.\end{aligned}$$

Using this, we have

$$\begin{aligned}|\langle\psi_T|\pi_{\vec{x},\vec{y},\vec{u},\vec{v}}\rangle| &\leqslant \xi\Big(\langle T, \vec{x}'\otimes\vec{y}'\otimes\vec{u}'\otimes\vec{v}'\rangle + \frac{10M}{n^{0.4}}\|\vec{x}\|_2\|\vec{y}\|_2\|\vec{u}\|_2\|\vec{v}\|_2\Big) \\ &\leqslant \xi\|\vec{x}'\|_2\|\vec{y}'\|_2\|\vec{u}'\|_2\|\vec{v}'\|_2\mathsf{OPT}_T + \frac{10M}{n^{0.4}}\xi\|\vec{x}\|_2\|\vec{y}\|_2\|\vec{u}\|_2\|\vec{v}\|_2. \qquad (17)\end{aligned}$$

We split now into two cases. First, suppose that $\sum_{a=1}^n(\frac{|x_a|^2}{1+|x_a|^2}+\frac{|y_a|^2}{1+|y_a|^2}+\frac{|u_a|^2}{1+|u_a|^2}+\frac{|v_a|^2}{1+|v_a|^2}) \leqslant n^{0.1}$. This implies that

$$\begin{aligned}&\sum_{a=1}^n(|x_a'|^2+|y_a'|^2+|u_a'|^2+|v_a'|^2) \\ &\leqslant (1+1/n^{0.2})\sum_{a=1}^n(\frac{|x_a'|^2}{1+|x_a'|^2}+\frac{|y_a'|^2}{1+|y_a'|^2}+\frac{|u_a'|^2}{1+|u_a'|^2}+\frac{|v_a'|^2}{1+|v_a'|^2}) \\ &\leqslant 2n^{0.1}.\end{aligned}$$

In this case, using Lemma 3.7 and the norm bound and magnitude bound on $\vec{x}'$,

$$\begin{aligned}\xi &\leqslant \left(\prod_{a=1}^{n}(1+|x'_a|^2)(1+|y'_a|^2)(1+|\widetilde{u}_a|^2)(1+|\widetilde{v}_a|^2)\right)^{-1/2} \\ &\leqslant e^{-\frac{1}{2}(\|\vec{x}'\|_2^2+\|\vec{y}'\|_2^2+\|\vec{u}'\|_2^2+\|\vec{v}'\|_2^2-\sum_{a=1}^{n}(|x'_a|^4+|y'_a|^4+|u'_a|^4+|v'_a|^4))} \\ &\leqslant e^{-\frac{1}{2}(\|\vec{x}'\|_2^2+\|\vec{y}'\|_2^2+\|\vec{u}'\|_2^2+\|\vec{v}'\|_2^2-2n^{-0.1})}.\end{aligned}$$

So, plugging this into (17),

$$\begin{aligned}&|\langle\psi_T|\pi_{\vec{x},\vec{y},\vec{u},\vec{v}}\rangle| \\ &\leqslant e^{-\frac{1}{2}(\|\vec{x}'\|_2^2+\|\vec{y}'\|_2^2+\|\vec{u}'\|_2^2+\|\vec{v}'\|_2^2-n^{-0.1})}\|\vec{x}'\|_2\|\vec{y}'\|_2\|\vec{u}'\|_2\|\vec{v}'\|_2\mathsf{OPT}_T+\frac{10M}{n^{0.4}}\xi\|\vec{x}\|_2\|\vec{y}\|_2\|\vec{u}\|_2\|\vec{v}\|_2 \\ &\leqslant e^{-2}e^{n^{-0.1}}\mathsf{OPT}_T+\frac{10M}{n^{0.4}}\xi\|\vec{x}\|_2\|\vec{y}\|_2\|\vec{u}\|_2\|\vec{v}\|_2 \\ &\leqslant e^{-2}\left(1+\frac{10}{n^{0.1}}\right)\mathsf{OPT}_T+\frac{10M}{n^{0.2}} \\ &\leqslant e^{-2}\mathsf{OPT}_T+\frac{10}{n^{0.1}}+\frac{10M}{n^{0.2}},\end{aligned}$$

where we used that the maximum possible value of $e^{-\frac{1}{2}\theta^2}\theta$ is exactly $e^{-1/2}$. This gives the desired statement in the first case. In the second case, suppose that $\sum_{a=1}^{n}(\frac{|x_a|^2}{1+|x_a|^2}+\frac{|y_a|^2}{1+|y_a|^2}+\frac{|u_a|^2}{1+|u_a|^2}+\frac{|v_a|^2}{1+|v_a|^2})>n^{0.1}$. Then we can argue that the optimum value is very small: by Cauchy–Schwarz and Lemma 3.10, for a sufficiently large n,

$$\begin{aligned}|\langle\psi_T|\pi_{\vec{x},\vec{y},\vec{u},\vec{v}}\rangle| &= |\langle\psi_T|\,\Pi_{\leqslant 4}\,|\pi_{\vec{x},\vec{y},\vec{u},\vec{v}}\rangle| \\ &\leqslant \||\psi_T\rangle\|_2\|\Pi_{\leqslant 4}\,|\pi_{\vec{x},\vec{y},\vec{u},\vec{v}}\rangle\|_2 \\ &\leqslant e^{-(2n^{0.1}-4)\log(2-4/n^{0.1})+(n^{0.1}-4)} \\ &\leqslant e^{-0.1n^{0.1}}\leqslant\frac{10}{n^{0.2}}\leqslant\frac{10M}{n^{0.2}}.\end{aligned}$$

This completes the proof of the upper bound and we are done. ♦

Now we can complete the proof of Theorem 7.3.

Proof of Theorem 7.3. Without loss of generality, we may assume that the original $m\times m\times m\times m$ tensor T that we are given is symmetric: otherwise, we can just symmetrize it and this only decreases the Frobenius norm. We can further normalize such that $\|T\|_F=1$. Then consider $U^{\otimes 4}T$ for $U\in\mathbb{C}^{n\times m}$ a Haar-random isometry, and the corresponding quantum state $|\psi_{U^{\otimes 4}T}\rangle$ as defined in Definition 7.4. By Lemma 7.6,

$$e^{-2}\mathsf{OPT}_{U^{\otimes 4}T}-\frac{10M}{n^{0.2}}\leqslant\mathsf{OPT}_{|\psi_{U^{\otimes 4}T}\rangle}\leqslant e^{-2}\mathsf{OPT}_{U^{\otimes 4}T}+\frac{10}{n^{0.1}}+\frac{10M}{n^{0.2}}$$

By Claim 7.5, with 0.99 probability we can take $M=(10m)^2$, and because U is an isometry, $\mathsf{OPT}_{U^{\otimes 4}T}=\mathsf{OPT}_T$, so

$$e^{-2}\mathsf{OPT}_T-\frac{1000m^2}{n^{0.2}}\leqslant\mathsf{OPT}_{|\psi_{U^{\otimes 4}T}\rangle}\leqslant e^{-2}\mathsf{OPT}_T+\frac{10}{n^{0.1}}+\frac{1000m^2}{n^{0.2}}.$$

Taking $n=\Omega((m/\varepsilon)^{20})$ completes the proof: then, $e^{-2}(\mathsf{OPT}_T-\varepsilon)\leqslant\mathsf{OPT}_{|\psi_{U^{\otimes 4}T}\rangle}\leqslant e^{-2}(\mathsf{OPT}_T+\varepsilon)$. So, finding the optimal product state fidelity to accuracy $0.01\varepsilon^2$ gives the spectral norm of the tensor to ε additive error. ♦

Acknowledgments

We thank Sitan Chen for helpful discussions at the early stages of this work.

AB is supported by the NSF TRIPODS program (award DMS-2022448). JB is supported by Henry Yuen's AFOSR (award FA9550-21-1-036) and NSF CAREER (award CCF-2144219). WK acknowledges support from the U.S. Department of Energy, Office of Science, National Quantum Information Science Research Centers, Quantum Systems Accelerator. ZL is supported by the U.S. Department of Energy, Office of Science, National Quantum Information Science Research Centers, Quantum Systems Accelerator, and by NSF Grant CCF-2311733. AL is supported in part by an NSF GRFP and a Hertz Fellowship. RO is supported by ARO grant W911NF2110001 and by a gift from Google Quantum AI. ET is supported by the Miller Institute for Basic Research in Science, University of California, Berkeley.

References

[AA23] Anurag Anshu and Srinivasan Arunachalam. "A survey on the complexity of learning quantum states". In: *Nature Reviews Physics* 6.1 (Dec. 2023), pp. 59–69. ISSN: 2522-5820. DOI: 10.1038/s42254-023-00662-4. arXiv: 2305.20069 [quant-ph] (page 6).

[Aar20] Scott Aaronson. "Shadow tomography of quantum states". In: *SIAM Journal on Computing* 49.5 (Jan. 2020), STOC18-368-STOC18–394. ISSN: 1095-7111. DOI: 10.1137/18m120275x. arXiv: 1711.01053 [quant-ph] (page 2).

[AGT19] Vedat Levi Alev, Fernando Granha Jeronimo, and Madhur Tulsiani. "Approximating constraint satisfaction problems on high-dimensional expanders". In: *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, Nov. 2019, pp. 180–201. DOI: 10.1109/focs.2019.00021. arXiv: 1907.07833 [cs.DS] (pages 7, 13).

[BBHKSZ12] Boaz Barak, Fernando G.S.L. Brandao, Aram W. Harrow, Jonathan Kelner, David Steurer, and Yuan Zhou. "Hypercontractivity, sum-of-squares proofs, and their applications". In: *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*. STOC'12. ACM, May 2012, pp. 307–326. DOI: 10.1145/2213977.2214006. arXiv: 1205.4484 [cs.CC] (page 12).

[BDJKKV22] Ainesh Bakshi, Ilias Diakonikolas, He Jia, Daniel M. Kane, Pravesh K. Kothari, and Santosh S. Vempala. "Robustly learning mixtures of k arbitrary Gaussians". In: *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*. STOC '22. ACM, June 2022, pp. 1234–1247. DOI: 10.1145/3519935.3519953. arXiv: 2012.02119 [cs.DS] (page 7).

[BGGLT17] Vijay Bhattiprolu, Mrinalkanti Ghosh, Venkatesan Guruswami, Euiwoong Lee, and Madhur Tulsiani. "Weak decoupling, polynomial folds and approximate optimization over the sphere". In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, Oct. 2017, pp. 1008–1019. DOI: 10.1109/focs.2017.97. arXiv: 1611.05998 [cs.DS] (pages 7, 12).

[BH16] Fernando G. S. L. Brandão and Aram W. Harrow. "Product-state approximations to quantum states". In: *Communications in Mathematical Physics* 342.1 (Jan. 2016), pp. 47–80. ISSN: 1432-0916. DOI: 10.1007/s00220-016-2575-1. arXiv: 1310.0017 [quant-ph] (pages 1, 6).

[BK21] Ainesh Bakshi and Pravesh K. Kothari. "List-decodable subspace recovery: dimension independent error in polynomial time". In: *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*. Society for Industrial and Applied Mathematics, Jan. 2021, pp. 1279–1297. ISBN: 9781611976465. DOI: 10.1137/1.9781611976465.78. arXiv: 2002.05139 [cs.DS] (page 7).

[BKS15] Boaz Barak, Jonathan A. Kelner, and David Steurer. "Dictionary learning and tensor decomposition via the sum-of-squares method". In: *Proceedings of the forty-seventh annual ACM symposium on Theory of Computing*. STOC '15. ACM, June 2015, pp. 143–151. DOI: 10.1145/2746539.2746605. arXiv: 1407.1543 [cs.DS] (page 7).

[BKS17] Boaz Barak, Pravesh K Kothari, and David Steurer. "Quantum entanglement, sum of squares, and the log rank conjecture". In: *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. STOC '17. ACM, June 2017, pp. 975–988. DOI: 10.1145/3055399.3055488. arXiv: 1701.06321 [quant-ph] (page 7).

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, Feb. 2013. ISBN: 9780199535255. DOI: 10.1093/acprof:oso/9780199535255.001.0001 (page 19).

[BLMT24] Ainesh Bakshi, Allen Liu, Ankur Moitra, and Ewin Tang. "Learning quantum Hamiltonians at any temperature in polynomial time". In: *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*. STOC '24. ACM, June 2024, pp. 1470–1477. DOI: 10.1145/3618260.3649619. arXiv: 2310.02243 [quant-ph] (page 7).

[BO24] Costin Bădescu and Ryan O'Donnell. "Improved quantum data analysis". In: *TheoretiCS* Volume 3 (Mar. 2024). ISSN: 2751-4838. DOI: 10.46298/theoretics.24.7. arXiv: 2011.10908 [quant-ph] (page 6).

[BRS11] Boaz Barak, Prasad Raghavendra, and David Steurer. "Rounding semidefinite programming hierarchies via global correlation". In: *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*. IEEE, Oct. 2011, pp. 472–481. DOI: 10.1109/focs.2011.95. arXiv: 1104.4680 [cs.DS] (page 7).

[BS15] Mikhail Belkin and Kaushik Sinha. "Polynomial learning of distribution families". In: *SIAM Journal on Computing* 44.4 (Jan. 2015), pp. 889–911. ISSN: 1095-7111. DOI: 10.1137/13090818x. arXiv: 1004.4864 [cs.LG] (page 7).

[Can20] Clément L. Canonne. *A survey on distribution testing: Your data is big. But is it blue?* Graduate Surveys 9. Theory of Computing Library, 2020, pp. 1–100. DOI: 10.4086/toc.gs.2020.009 (page 3).

[CGYZ24] Sitan Chen, Weiyuan Gong, Qi Ye, and Zhihan Zhang. "Stabilizer bootstrapping: a recipe for efficient agnostic tomography and magic estimation". Aug. 13, 2024. arXiv: 2408.06967 [quant-ph] (pages 5, 6).

[Cha24] Garnet Kin-Lic Chan. "Quantum chemistry, classical heuristics, and quantum advantage". In: (July 15, 2024). DOI: 10.48550/ARXIV.2407.11235. arXiv: 2407.11235 [quant-ph] (page 1).

[Cra+10] Marcus Cramer, Martin B. Plenio, Steven T. Flammia, Rolando Somma, David Gross, Stephen D. Bartlett, Olivier Landon-Cardinal, David Poulin, and Yi-Kai Liu. "Efficient quantum state tomography". In: *Nature Communications* 1.1 (Dec. 2010). DOI: 10.1038/ncomms1147. arXiv: 1101.4366 [quant-ph] (pages 5, 67, 69).

[CSV17] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. "Learning from untrusted data". In: *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. STOC '17. ACM, June 2017. DOI: 10.1145/3055399.3055491. arXiv: 1611.02315 [cs.LG] (page 7).

[DKKLMS19] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. "Robust estimators in high-dimensions without the computational intractability". In: *SIAM Journal on Computing* 48.2 (Jan. 2019), pp. 742–864. ISSN: 1095-7111. DOI: 10.1137/17m1126680. aRxIV: 1604.06443 (cs.DS) (page 7).

[DS18] Gabriele De Chiara and Anna Sanpera. "Genuine quantum correlations in quantum many-body systems: a review of recent progress". In: *Reports on Progress in Physics* 81.7 (June 2018), p. 074002. ISSN: 1361-6633. DOI: 10.1088/1361-6633/aabf61. arXiv: 1711.07824 [quant-ph] (page 6).

[FL17] Shmuel Friedland and Lek-Heng Lim. "Nuclear norm of higher-order tensors". In: *Mathematics of Computation* 87.311 (Sept. 2017), pp. 1255–1281. ISSN: 1088-6842. DOI: 10.1090/mcom/3239. arXiv: 1410.6072 [cs.CC] (pages 4, 56).

[FO24] Steven T. Flammia and Ryan O'Donnell. "Quantum chi-squared tomography and mutual information testing". In: *Quantum* 8 (June 2024), p. 1381. ISSN: 2521-327X. DOI: 10.22331/q-2024-06-20-1381. arXiv: 2305.18519 [quant-ph] (pages 43, 68).

[GIKL24] Sabee Grewal, Vishnu Iyer, William Kretschmer, and Daniel Liang. "Agnostic tomography of stabilizer product states". Apr. 4, 2024. arXiv: 2404.03813 [quant-ph] (pages 2, 5, 6).

[HKMN23] Samuel B. Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. "Robustness implies privacy in statistical estimation". In: *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*. STOC '23. ACM, June 2023, pp. 497–506. DOI: 10.1145/3564246.3585115. arXiv: 2212.05015 [cs.DS] (page 7).

[HKOT23] Jeongwan Haah, Robin Kothari, Ryan O'Donnell, and Ewin Tang. "Query-optimal estimation of unitary channels in diamond distance". In: *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, Nov. 2023. DOI: 10.1109/focs57990.2023.00028. arXiv: 2302.14066 [quant-ph] (page 29).

[HKP20] Hsin-Yuan Huang, Richard Kueng, and John Preskill. "Predicting many properties of a quantum system from very few measurements". In: *Nature Physics* 16.10 (June 2020), pp. 1050–1057. DOI: 10.1038/s41567-020-0932-7. arXiv: 2002.08953 [quant-ph] (pages 2, 3, 6, 27–29, 67).

[HM13] Aram W. Harrow and Ashley Montanaro. "Testing product states, quantum Merlin-Arthur games and tensor optimization". In: *Journal of the ACM* 60.1 (Feb. 2013), pp. 1–43. ISSN: 1557-735X. DOI: 10.1145/2432622.2432625. arXiv: 1001.0017 [quant-ph] (pages 2, 3, 6).

[HSS15] Samuel B. Hopkins, Jonathan Shi, and David Steurer. "Tensor principal component analysis via sum-of-square proofs". In: *Proceedings of The 28th Conference on Learning Theory*. Ed. by Peter Grünwald, Elad Hazan, and Satyen Kale. Vol. 40. Proceedings of Machine Learning Research. Paris, France: PMLR, 2015, pp. 956–1006. arXiv: 1507.03269 [cs.LG] (page 7).

[ICKHC16] Raban Iten, Roger Colbeck, Ivan Kukuljan, Jonathan Home, and Matthias Christandl. "Quantum circuits for isometries". In: *Physical Review A* 93.3 (Mar. 2016), p. 032318. ISSN: 2469-9934. DOI: 10.1103/physreva.93.032318. arXiv: 1501.06911 [quant-ph] (page 68).

[JV14] Richard Jozsa and Marrten Van Den Nest. "Classical simulation complexity of extended Clifford circuits". In: *Quantum Info. Comput.* 14.7 & 8 (May 2014), pp. 633–648. ISSN: 1533-7146. DOI: 10.26421/QIC14.7-8-7. arXiv: 1305.6190 [quant-ph] (page 3).

[Kar72] Richard M. Karp. "Reducibility among combinatorial problems". In: *Complexity of Computer Computations*. Springer US, 1972, pp. 85–103. ISBN: 9781468420012. DOI: 10.1007/978-1-4684-2001-2_9 (page 56).

[KKK19] Sushrut Karmalkar, Adam Klivans, and Pravesh Kothari. "List-decodable linear regression". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019. arXiv: 1905.05679 [cs.DS] (page 7).

[KRT17] Richard Kueng, Holger Rauhut, and Ulrich Terstiege. "Low rank matrix recovery from rank one measurements". In: *Applied and Computational Harmonic Analysis* 42.1 (Jan. 2017), pp. 88–116. ISSN: 1063-5203. DOI: 10.1016/j.acha.2015.07.007. arXiv: 1410.6913 [cs.IT] (page 68).

[KSS18] Pravesh K. Kothari, Jacob Steinhardt, and David Steurer. "Robust moment estimation and improved clustering via sum of squares". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. STOC '18. ACM, June 2018, pp. 1035–1046. DOI: 10.1145/3188745.3188970 (page 7).

[Lee+23] Seunghoon Lee, Joonho Lee, Huanchen Zhai, Yu Tong, Alexander M. Dalzell, Ashutosh Kumar, Phillip Helms, Johnnie Gray, Zhi-Hao Cui, Wenyuan Liu, Michael Kastoryano, Ryan Babbush, John Preskill, David R. Reichman, Earl T. Campbell, Edward F. Valeev, Lin Lin, and Garnet Kin-Lic Chan. "Evaluating the evidence for exponential quantum advantage in ground-state quantum chemistry". In: *Nature Communications* 14.1 (Apr. 2023). ISSN: 2041-1723. DOI: 10.1038/s41467-023-37587-6. arXiv: 2208.02199 [physics.chem-ph] (page 7).

[LM21] Allen Liu and Ankur Moitra. "Settling the robust learnability of mixtures of Gaussians". In: *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. STOC '21. ACM, June 2021, pp. 518–531. DOI: 10.1145/3406325.3451084. arXiv: 2011.03622 [cs.DS] (page 7).

[Low21] Angus Lowe. "Learning quantum states without entangled measurements". MA thesis. University of Waterloo, 2021 (page 43).

[LRV16] Kevin A. Lai, Anup B. Rao, and Santosh Vempala. "Agnostic estimation of mean and covariance". In: *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, Oct. 2016, pp. 665–674. DOI: 10.1109/focs.2016.76. arXiv: 1604.06968 [cs.DS] (page 7).

[MBSBN18] Jarrod R. McClean, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven. "Barren plateaus in quantum neural network training landscapes". In: *Nature Communications* 9.1 (Nov. 2018). ISSN: 2041-1723. DOI: 10.1038/s41467-018-07090-4 (page 8).

[MKB05] Florian Mintert, Marek Kuś, and Andreas Buchleitner. "Concurrence of mixed multipartite quantum states". In: *Physical Review Letters* 95.26 (Dec. 2005), p. 260502. ISSN: 1079-7114. DOI: 10.1103/physrevlett.95.260502. arXiv: quant-ph/0411127 (page 6).

[MR17] Pasin Manurangsi and Prasad Raghavendra. "A birthday repetition theorem and complexity of approximating dense CSPs". en. In: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2017. DOI: 10.4230/LIPIcs.ICALP.2017.78. arXiv: 1607.02986 [cs.CC] (pages 7, 13).

[MS08] Claire Mathieu and Warren Schudy. "Yet another algorithm for dense max cut: Go greedy". In: *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '08. San Francisco, California: Society for Industrial and Applied Mathematics, 2008, pp. 176–182 (page 13).

[MV10] Ankur Moitra and Gregory Valiant. "Settling the polynomial learnability of mixtures of Gaussians". In: *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE, Oct. 2010, pp. 93–102. DOI: 10.1109/focs.2010.15. arXiv: 1004.4223 [cs.LG] (page 7).

[Nar24] Shyam Narayanan. "Improved algorithms for learning quantum hamiltonians, via flat polynomials". 2024. arXiv: 2407.04540 [quant-ph] (page 7).

[ODV08] Román Orús, Sébastien Dusuel, and Julien Vidal. "Equivalence of critical scaling laws for many-body entanglement in the Lipkin-Meshkov-Glick model". In: *Physical Review Letters* 101.2 (July 2008), p. 025701. ISSN: 1079-7114. DOI: 10.1103/physrevlett.101.025701. arXiv: 0803.3151 [cond-mat.other] (page 6).

[OW10] Román Orús and Tzu-Chieh Wei. "Visualizing elusive phase transitions with geometric entanglement". In: *Physical Review B* 82.15 (Oct. 2010), p. 155120. ISSN: 1550-235X. DOI: 10.1103/physrevb.82.155120. arXiv: 0910.2488 [cond-mat.str-el] (page 6).

[PVWC07] D. Perez-Garcia, F. Verstraete, M.M. Wolf, and J.I. Cirac. "Matrix product state representations". In: *Quantum Information and Computation* 7.5 & 6 (July 2007), pp. 401–430. ISSN: 1533-7146. DOI: 10.26421/qic7.5-6-1. arXiv: quant-ph/0608197 (page 5).

[RRS17] Prasad Raghavendra, Satish Rao, and Tselil Schramm. "Strongly refuting random CSPs below the spectral threshold". In: *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. STOC '17. ACM, June 2017, pp. 121–131. DOI: 10.1145/3055399.3055417. arXiv: 1605.00058 [cs.DS] (page 7).

[RT12] Prasad Raghavendra and Ning Tan. "Approximating CSPs with global cardinality constraints using SDP hierarchies". In: *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, Jan. 2012, pp. 373–387. DOI: 10.1137/1.9781611973099.33. arXiv: 1110.1064 [cs.DS] (page 7).

[RY20] Prasad Raghavendra and Morris Yau. "List decodable learning via sum of squares". In: *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, Jan. 2020, pp. 161–180. ISBN: 9781611975994. DOI: 10.1137/1.9781611975994.10. arXiv: 1905.04660 [cs.DS] (page 7).

[Sch11] Ulrich Schollwöck. "The density-matrix renormalization group in the age of matrix product states". In: *Annals of Physics* 326.1 (2011), pp. 96–192. DOI: 10.1016/j.aop.2010.09.012 (page 5).

[SSY23] Nadish de Silva, Wilfred Salmon, and Ming Yin. "Fast algorithms for classical specifications of stabiliser states and Clifford gates". 2023. arXiv: 2311.10357 [quant-ph] (page 29).

[Sub20] Eliran Subag. "Following the ground states of full-RSB spherical spin glasses". In: *Communications on Pure and Applied Mathematics* 74.5 (June 2020), pp. 1021–1044. ISSN: 1097-0312. DOI: 10.1002/cpa.21922. arXiv: 1812.04588 [math.PR] (page 7).

[SW22] Mehdi Soleimanifar and John Wright. "Testing matrix product states". In: *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. Society for Industrial and Applied Mathematics, Jan. 2022, pp. 1679–1701. ISBN: 9781611977073. DOI: 10.1137/1.9781611977073.68. arXiv: 2201.01824 [quant-ph] (page 3).

[Van21] Ewout Van Den Berg. "A simple method for sampling random Clifford operators". In: *2021 IEEE International Conference on Quantum Computing and Engineering (QCE)*. IEEE, Oct. 2021, pp. 54–59. DOI: 10.1109/QCE52317.2021.00021. arXiv: 2008.06011 [quant-ph] (page 29).

[VPRK97] V. Vedral, M. B. Plenio, M. A. Rippin, and P. L. Knight. "Quantifying entanglement". In: *Physical Review Letters* 78.12 (Mar. 1997), pp. 2275–2279. ISSN: 1079-7114. DOI: 10.1103/physrevlett.78.2275. arXiv: quant-ph/9702027 (page 6).

[WDMVG05] Tzu-Chieh Wei, Dyutiman Das, Swagatam Mukhopadyay, Smitha Vishveshwara, and Paul M. Goldbart. "Global entanglement and quantum criticality in spin chains". In: *Physical Review A* 71.6 (June 2005), p. 060305. ISSN: 1094-1622. DOI: 10.1103/physreva.71.060305. arXiv: quant-ph/0405162 (page 6).

[WG03] Tzu-Chieh Wei and Paul M. Goldbart. "Geometric measure of entanglement and applications to bipartite and multipartite quantum states". In: *Physical Review A* 68.4 (Oct. 2003), p. 042307. ISSN: 1094-1622. DOI: 10.1103/physreva.68.042307. arXiv: quant-ph/0307219 (pages 1, 6).

[Yar14] Grigory Yaroslavtsev. "Going for speed: sublinear algorithms for dense r-CSPs". 2014. arXiv: 1407.7887 [cs.DS] (page 13).

A Agnostic learning of a discrete class of product states

Theorem A.1 (Agnostic learning of a discrete class of product states)**.** *Suppose we are given copies of an n-qudit state ρ, along with classical descriptions of sets of single-qudit pure states $\{\mathcal{A}_k\}_{k\in[n]}$ such that $|\mathcal{A}_k| \leqslant s$ and, for some $\gamma \geqslant 1/e$, $|\langle\phi|\phi'\rangle|^2 \leqslant \gamma$ for all distinct $|\phi\rangle, |\phi'\rangle \in \mathcal{A}_k$. These implicitly*

define a class of product states $\mathcal{P}$,

$$\mathcal{P} = \mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_n = \{|\pi_1\rangle \otimes \cdots \otimes |\pi_n\rangle \mid |\pi_k\rangle \in \mathcal{A}_k\}.$$

Let $\mathcal{P}_\eta$ be the states in the class with fidelity at least η with π:

$$\mathcal{P}_\eta = \{|\pi\rangle \in \mathcal{P} \mid \langle\pi| \rho |\pi\rangle \geqslant \eta\}.$$

Then given parameters $\eta, \varepsilon, \delta \in (0,1)$ with $\varepsilon \leqslant \eta/2$, we can output a set $\mathcal{S}$ such that $\mathcal{P}_\eta \subseteq \mathcal{S} \subseteq \mathcal{P}_{\eta-\varepsilon}$ with probability $\geqslant 1-\delta$ using $O((10ns)^{\log \frac{25}{\eta} / \log \frac{1}{\gamma}} \frac{1}{\varepsilon^2} \log \frac{1}{\delta})$ copies of ρ.

The classical part of the algorithm is linear-time, costing $O(nC)$, where C is the number of copies of ρ used. The quantum gate complexity is also $O(nC)$, where we consider a gate to be an operation on a qudit; the only circuits we will need are single layers of single-qudit gates, followed by measurement, to estimate quantities like $\langle\pi| \rho |\pi\rangle$ for $|\pi\rangle \in \mathcal{P}$.

Claim A.2. In the set-up of Theorem A.1, $|\mathcal{P}_\eta| \leqslant (10ns)^{\log \frac{2}{\eta} / \log \frac{1}{\gamma}}$.

Proof. The basic idea is that $\mathcal{P}_\eta$ consists of a small number of balls, where a ball is the set of elements of $\mathcal{P}$ close to a particular $|\pi\rangle \in \mathcal{P}$. The number of balls is small because ρ must place mass in the direction of every ball, and the mass of ρ is bounded.

We formalize this argument now. For $\ell = \lfloor \log \frac{2}{\eta} / \log \frac{1}{\gamma} \rfloor$, construct a net $\mathcal{N} \subseteq \mathcal{P}_\eta$ with the following properties.

1. For any $|\pi\rangle \in \mathcal{P}_\eta$, there exists a $|\omega\rangle \in \mathcal{N}$ such that $|\pi\rangle$ and $|\omega\rangle$ differ in at most ℓ qudits;
2. Any distinct $|\pi\rangle, |\omega\rangle \in \mathcal{N}$ differ in at least $\ell + 1$ qudits.

Such a net can be constructed greedily: start with an empty net, and while there is a violation of condition 1, add the corresponding $|\pi\rangle$ to the net. Let M be the matrix whose columns are the elements of the net $\mathcal{N} = \{|\pi^{(i)}\rangle\}_i$. Then following the same argument as Claim 5.4,

$$\|M^\dagger M\|_{\text{op}} \leqslant 1 + |\mathcal{N}|\gamma^{\ell+1} \leqslant 1 + |\mathcal{N}|(\eta/2)$$
$$\|MM^\dagger\|_{\text{op}} \geqslant \text{tr}(MM^\dagger\rho) = \sum_i \langle\pi^{(i)}|\rho|\pi^{(i)}\rangle \geqslant |\mathcal{N}|\eta$$

Together, we can conclude that $|\mathcal{N}| \leqslant \frac{2}{\eta}$.

Finally, by property 2 of the net, $\mathcal{P}_\eta$ is contained in the set of product states in $\mathcal{P}$ which differ from an element of $\mathcal{N}$ in at most ℓ qudits. Consider some $|\pi\rangle \in \mathcal{P}_\eta$. There are $\sum_{i=0}^{\ell} \binom{n}{i} s^i \leqslant \sum_{i=0}^{\ell} (ns)^i \leqslant 2(ns)^\ell$ elements of $\mathcal{P}$ which differ from $|\pi\rangle$ in at most ℓ qudits. So, $|\mathcal{P}_\eta| \leqslant |\mathcal{N}|2(ns)^\ell \leqslant \frac{4}{\eta}(ns)^\ell$, which gives the desired bound. ♦

Now, we present the proof of Theorem A.1. The algorithm is given below: we start from qudit 1 and iteratively add a qubit, maintaining at iteration m a set $\mathcal{S}_m$ of product states which have good fidelity with ρ.

Algorithm A.3 (Agnostic learning for discrete product state classes)**.**

Input: Copies of an n-qubit state ρ, a class of n-qubit product states $\mathcal{P} = \mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_n$, with parameters $\eta, \gamma, \varepsilon, \delta$ as in Theorem A.1;

Output: A set of product states $\mathcal{S}$ such that $\mathcal{P}_\eta \subseteq \mathcal{S} \subseteq \mathcal{P}_{\eta-\varepsilon}$ with probability $\geqslant 1 - \delta$;

Procedure:

Let $\mathcal{S}_0 = \varnothing$;
for m from 1 to n **do**
    Initialize $\mathcal{S}_m = \varnothing$;
    for all $|\pi\rangle \in \mathcal{S}_{m-1} \otimes \mathcal{A}_m$ **do**
        Estimate $\langle\pi| \rho_{[m]} |\pi\rangle$ to $\varepsilon/2$ error with success probability $\geqslant 1 - \delta/(10ns)^{\log \frac{20}{\eta} / \log \frac{1}{\gamma}}$;
        if the estimate is at least $\eta - \varepsilon/2$ **then**
            Add $|\pi\rangle$ to $\mathcal{S}_m$;
return $\mathcal{S}_n$

Claim A.4 (Correctness)**.** With probability $\geqslant 1 - \delta$, at the completion of Algorithm A.3, for every $m \in \{0, 1, \ldots, n\}$,

$$\mathcal{S}_m \supseteq \{|\pi\rangle \in \mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_m \mid \langle\pi| \rho_{[m]} |\pi\rangle \geqslant \eta\}$$
$$\mathcal{S}_m \subseteq \{|\pi\rangle \in \mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_m \mid \langle\pi| \rho_{[m]} |\pi\rangle \geqslant \eta - \varepsilon\}$$

Proof. First, we consider the algorithm under the event that the algortihm never fails. We prove by induction on m. The base case $m = 0$ is true trivially. For the inductive step, consider some $m > 0$, and consider some $|\pi\rangle \in \mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_m$ such that $\langle\pi| \rho_{[m]} |\pi\rangle \geqslant \eta$. Let $|\pi'\rangle$ be the product state $|\pi\rangle$ with the mth qubit traced out. Then

$$\langle\pi'| \rho_{[m-1]} |\pi'\rangle \geqslant \langle\pi| \rho_{[m]} |\pi\rangle \geqslant \eta,$$

so $|\pi'\rangle \in \mathcal{S}_{m-1}$ by the inductive hypothesis, and $|\pi\rangle \in \mathcal{S}_{m-1} \otimes \mathcal{A}_m$. Because we are assuming that the estimation procedure succeeds, this means that $|\pi\rangle$ will be added to $\mathcal{S}_m$, proving the first equation of the claim. The second equation holds because, again assuming that the estimation procedure always succeeds, all elements added to $\mathcal{S}_m$ have fidelity at least $\eta - \varepsilon/2 - \varepsilon/2$ with $\rho_{[m]}$.

To conclude, we account for failure. Supposing that the algorithm never fails, the second equation in the claim along with Claim A.2 (applied on $\rho_{[m]}$, the state on the m-qudit subsystem) implies that $|\mathcal{S}_m| \leqslant (10ms)^{\log \frac{2}{\eta-\varepsilon} / \log \frac{1}{\gamma}}$. So, on a successful run, the estimation procedure is run at most $ns \cdot (10ns)^{\log \frac{2}{\eta-\varepsilon} / \log \frac{1}{\gamma}}$ times, and so the failure probability is chosen such that the probability a failure occurs is at most δ. ♦

The above claim implies that the output of the algorithm satisfies $\mathcal{P}_\eta \subseteq \mathcal{S} \subseteq \mathcal{P}_{\eta-\varepsilon}$, which is the desired correctness condition. What remains is to analyze the complexity. The dominating cost is the estimation step, where the fidelity of (a subsystem of) ρ with a product state is estimated to $\varepsilon/2$ error with failure probability $\delta/(10ns)^{\log \frac{20}{\eta} / \log \frac{1}{\gamma}}$. This step is run at most $s \cdot \sum_{m=1}^{n} |\mathcal{S}_m| \leqslant ns \cdot (10ns)^{\log \frac{2}{\eta-\varepsilon} / \log \frac{1}{\gamma}} = O((10ns)^{\log \frac{20}{\eta} / \log \frac{1}{\gamma}})$ times, by Claim A.4 and Claim A.2.

Each estimation protocol costs $O(\frac{1}{\varepsilon^2}\log((10ns)^{\log\frac{20}{\eta}/\log\frac{1}{\gamma}}/\delta))$ copies of ρ, where we use the naive algorithm of measuring in the appropriate basis and estimating the corresponding probability.

Remark A.5. The estimation procedure could also be done using randomized Clifford measurements as in [HKP20], which reduces the sample complexity to poly-logarithmic in n. However, computing the resulting estimators takes exponential time, making the resulting algorithm computationally inefficient, except for limited settings, such as in the case of stabilizer product states.

B Agnostic improper learning of matrix product states

The task of product state learning motivates a more general question: What ensembles of "low-entanglement" states can we perform computationally-efficient agnostic learning for? One physically-motivated ensemble is the class of low bond-dimension matrix product states, for which the (non-agnostic) learning task was studied in [Cra+10].

Definition B.1 (Matrix product state with bond dimension r, open boundary condition)**.** A matrix product state (MPS) is a state over n total d-dimensional qudits that can be written as follows

$$|\psi\rangle = \sum_{s_1,\ldots,s_n\in[d]^n}\sum_{\alpha_1,\ldots,\alpha_n\in[r]^n}\left(A_1^{(s_1)}\right)_{1,\alpha_1}\left(A_2^{(s_2)}\right)_{\alpha_1,\alpha_2}\cdots\left(A_{n-1}^{(s_{n-1})}\right)_{\alpha_{n-2},\alpha_{n-1}}\left(A_n^{(s_n)}\right)_{\alpha_{n-1},\alpha_n}|s_1,\ldots,s_n\rangle\ ,$$

where for all $i = 2,\ldots,n-1$ and all s_i, we have that $A_i^{(s_i)}$ are $r\times r$ complex matrices, and for $i\in\{1,n\}$ and all s_i. The dimension r is known as the *bond dimension* of the MPS. We let $\mathrm{MPS}_{n,d,r}$ denote the set of all such states.[6]

A MPS has a natural representation in the form of a tensor network as in Figure 1, where here each A_i represents the concatenation of all of the $A_i^{(s_i)}$ into a 3-tensor.

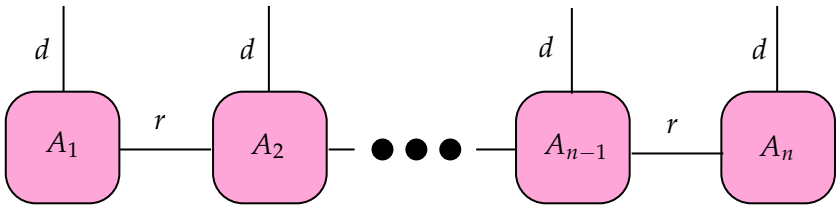

Figure 1: Tensor network representation of a matrix product state with bond dimension r.

In this section we present an agnostic learning algorithm for matrix product states very closely based on the algorithm from [Cra+10]. In fact, it is nearly identical to the algorithm presented there, but with different parameters. For completeness, here we will present a more or less self-contained description of the algorithm, and a proof of its correctness.

Unlike in the product state setting, given a state that is close to a matrix product state, we will not learn a matrix product state with the same bond dimension, but one whose bond dimension is a polynomial-factor higher, hence "agnostic improper learning". At a very high level, the algorithm of [Cra+10] demonstrates that one can successively learn unitary rotations that act on few qubits at a time that allow you to disentangle qubits successively. This is because

[6] Matrix product states with periodic boundary conditions are defined by taking the trace of the product of the $A_i^{(s_i)}$, or equivalently enforcing that the first index of $A_1^{(s_1)}$ is equal to α_n, but we do not consider these in this paper.

when ρ is a MPS (or very close to one), then it has low Schmidt rank along every cut in the tensor network, and consequently essentially all of the information about ρ is contained in a r-dimensional subspace along every cut. While this is obviously not true when ρ is arbitrary, we show that we can find $\mathrm{poly}(n, d, r, 1/\varepsilon)$-dimensional subspaces which do preserve all of the correlation structure between ρ and any matrix product state.

Theorem B.2 (Agnostic improper learning of matrix product states)**.** *Suppose we are given copies of an n-qudit state ρ. Then given parameters $\varepsilon, \delta \in (0,1)$ and a bond dimension parameter r, we can output a description of a matrix product state $|\hat{\phi}\rangle$ with bond dimension $dn^2 \cdot \mathrm{poly}(r, 1/\varepsilon)$ such that, with probability $\geqslant 1-\delta$,*

$$\langle\hat{\phi}|\rho|\hat{\phi}\rangle \geqslant \max_{|\phi\rangle \in MPS_r} \langle\phi|\rho|\phi\rangle - \varepsilon. \tag{18}$$

This algorithm uses $N = \mathrm{poly}(n, d, r, 1/\varepsilon, \log\frac{1}{\delta})$ copies of ρ, along with $\mathrm{poly}(N)$ quantum gates and classical overhead.

We will use the following notion of Schmidt rank and decomposition.

Definition B.3. Let $|\phi\rangle$ be an n-qudit state, i.e. a vector in $\left(\mathbb{C}^d\right)^{\otimes n}$. For $i = 1, \ldots, n$ we say that the *Schmidt decomposition of $|\phi\rangle$ at position i* is the Schmidt decomposition of $|\phi\rangle$ when viewed as an element of the bipartite system $A \otimes B$ where $A = \left(\mathbb{C}^d\right)^{\otimes i}$ and $B = \left(\mathbb{C}^d\right)^{\otimes(n-i)}$. We correspondingly say that the Schmidt rank of $|\phi\rangle$ at position i is the rank of this Schmidt decomposition.

Note that a matrix product state with bond dimension r has Schmidt rank r across all bipartitions. We will also need the following two operations:

Definition B.4 (Disentangling unitary)**.** Let W be a subspace of $(\mathbb{C}^d)^{\otimes\kappa}$ of dimension $d^{\kappa-1}$. We say a unitary matrix $U : (\mathbb{C}^d)^{\otimes\kappa} \to (\mathbb{C}^d)^{\otimes\kappa}$ is a disentangling unitary for W if for all $|\phi\rangle \in W$, we have that $U|\phi\rangle = |0\rangle \otimes |\psi\rangle$ for some state $|\psi\rangle$ on the remaining $\kappa - 1$ qudits, and for all $|\phi\rangle$ orthogonal to W, we have $\mathrm{tr}\left[(|0\rangle\langle 0| \otimes I)U|\phi\rangle\langle\phi|U^\dagger\right] = 0$.

Lemma B.5 ([ICKHC16])**.** *Given a classical description of a $d^{\kappa-1}$-dimensional subspace W, there is an implementation of the disentangling unitary in time $\mathrm{poly}(d^\kappa)$, when d is a power of 2.*

We also need a result concerning the tomography of sub-normalized states in trace distance, inspired by the sub-normalized fidelity-squared tomography algorithm of [FO24].

Lemma B.6 (Sub-normalized tomography)**.** *Let ρ be a quantum state on r registers of dimension d. Let $\Pi = |0^i\rangle\langle 0^i| \otimes I^{\otimes r-i}$ and $\mu = \mathrm{tr}[\Pi\rho]$, and $\delta, \varepsilon > 0$. Then there exists an algorithm that performs tomography of $\sigma = \Pi\rho\Pi$ to within trace distance error ε with failure probability δ using $O(\mu \cdot \frac{d^{3(r-i)}}{\varepsilon^2}\log(1/\delta))$ copies of ρ and in time $\mathrm{poly}(d^{r-i}, 1/\varepsilon, \log(1/\delta))$.*

Proof. Similar to [FO24], the algorithm first takes $m = O(\mu \cdot \frac{d^{3(r-i)}}{\varepsilon^2}\log(1/\delta))$ copies of ρ, and measures the PVM $\{\Pi, I-\Pi\}$ on all of them. It keeps all of the copies that had measurement outcome Π, and is thus left with $(\rho|_\Pi)^{\otimes m'}$, where $\rho|_\Pi = \Pi\rho\Pi/\mu$ and $m' \sim \mathsf{Binomial}(m, \mu)$. Then the algorithm then traces out the first i registers (which are in the state $|0^i\rangle\langle 0^i|$), and performs tomography on the remaining m' states with error $\varepsilon/2\mu$ and failure probability $\delta/2$. to produce an estimate $\hat{\rho}|_\Pi$. Since we want a (runtime) efficient algorithm, we use the random Clifford tomography technique described by [KRT17], which uses sample-complexity $O((d')^3/(\varepsilon')^2\log(1/\delta'))$ (where d' is the dimension of the input to this subroutine, ε' is the error parameter given to this subroutine and δ' is the failure probability of this subroutine), and runs

in polynomial time in all of those parameters. Substituting $d' = d^{r-i}$ and $\varepsilon' = \varepsilon/\mu$, this step requires $m' = O(\mu^2 (d^{r-i})^3/\varepsilon^2)$ copies of the post-selected state. Finally, the algorithm outputs $\hat{\sigma} = (m'/m)\hat{\rho}|_\Pi$.

In order to show that the algorithm produces an estimate with the correct trace norm bound, we can write the following:

$$\begin{aligned} \|\sigma - \hat{\sigma}\|_1 &= \|\mu\rho|_\Pi - (m'/m)\hat{\rho}|_\Pi\|_1 \\ &= \mu\|\rho|_\Pi - (m'/\mu m)\hat{\rho}_\Pi\|_1 . \end{aligned}$$

We can apply the triangle inequality and linearity of expectation to get the following:

$$\|\rho|_\Pi - (m'/\mu m)\hat{\rho}_\Pi\|_1 = \|\rho|_\Pi - \hat{\rho}|_\Pi\|_1 + \|(1 - (m'/(\mu m)))\hat{\rho}_\Pi\|_1 .$$

We want to bound the probability that this value is greater than ε. From the guarantee of the tomography algorithm, the first term is bounded by $\varepsilon/2\mu$ except with probability $\delta/2$. From Hoeffding's inequality, as long as $m \geqslant 2\log(2/\delta)\mu/\varepsilon$, the second term is also bounded by $\varepsilon/(2\mu)$ with except with probability $\delta/2$. Plugging these back into the difference in norm of σ and applying a union bound, we get that the trace-norm error is at most ε except with probability δ, as desired. Since we needed $O(\mu^2 \frac{d^{3(r-i)}}{\varepsilon^2}\log(1/\delta))$ copies of the state to perform the tomography, and in expectation μm copies survive the measurement of Π, we need to start with $O(\mu \cdot \frac{d^3}{\varepsilon^2}\log(1/\delta))$ copies of ρ, as desired. ♦

Note that in the remainder of this section, $r - i$ will be roughly $\log_d(\text{poly})$, where poly is some polynomial in all parameters. Thus, the scaling in d^{r-i} will be polynomial in the inputs.

Our overall algorithm proceeds now as follows, quite similarly to [Cra+10]. Let ρ be the overall, unknown state. Let $\tau = \frac{\varepsilon^2}{9n^2r^4}$, and let $\kappa = \lceil \log_d(1/\tau) \rceil + 1$. We will produce a sequence of disentangling unitaries $U_0, U_1, \ldots, U_{n-\kappa+1}$, where the j-th unitary will act only on the (j)-th through $(j+\kappa)$-th sites.[7] We will also maintain a sequence of intermediate unnormalized states $\rho'_0, \ldots, \rho'_{n-\kappa+1}$ which we can efficiently prepare given ρ and the U_i. Each state ρ'_i will act on the last $n - i$ qudits. Initially set $U_0 = I$ and $\rho'_0 = \rho_0 = \rho$.

Then, for all $i \geqslant 1$, given ρ'_{i-1}, form $\sigma_i = \mathrm{tr}_{\geqslant\kappa}(\rho'_{i-1})$, and perform state tomography to obtain the classical description of a state $\widehat{\sigma}_i$ satisfying

$$\|\widehat{\sigma}_i - \sigma_i\|_1 \leqslant \tau \tag{19}$$

with probability $1 - \delta/n$. Let W_i denote the subspace spanned by the singular values of $\widehat{\sigma}_i$ that exceed τ. Note that since $\widehat{\sigma}$ has trace 1, the subspace W has dimension at most $1/\tau < d^{\kappa-1}$. Extend this subspace arbitrarily to have dimension $d^{\kappa-1}$, and let U_i be a disentangling unitary for this subspace. Finally, let ρ'_i be the result of projecting the first qudit of $U_i\rho'_{i-1}U_i^\dagger$ onto $|0\rangle\langle 0|$.

After we have produced these disentangling unitaries $U_1, \ldots, U_{n-\kappa+1}$, form $\rho'_{n-\kappa+1}$ and $\widehat{\sigma}_{n-\kappa+1}$ as before, and let $|\psi\rangle$ denote the top eigenvector of $\widehat{\sigma}_{n-\kappa+1}$. Our final estimate of ρ is the state $|\hat{\phi}\rangle$ given by

$$|\hat{\phi}\rangle = (U_1 \ldots U_{n-\kappa+1})\left(|0^{n-\kappa}\rangle \otimes |\psi\rangle\right) .$$

The formal pseudocode for this algorithm is given in Algorithm B.8.

Lemma B.7. *$|\hat{\phi}\rangle$ is a matrix product state with bond dimension $dn^2\,\mathrm{poly}(r, 1/\varepsilon)$.*

[7] Here we slightly abuse notation and also let U_i denote the extension of the disentangling unitary to the entire space which acts on the identity outside of the aforementioned sites.

Proof. We can proceed by writing out the entries of the state $|\hat{\phi}\rangle$. Letting κ be the number of qudits each unitary (and $|\psi\rangle$ acts on), we can write $|\psi\rangle = \sum_{t_{n-\kappa+1},\ldots,t_{n-1},s_n} \alpha_{t_{n-\kappa+1},\ldots,t_{n-1},s_n} |t_{n-\kappa+1},\ldots,t_{n-1},s_n\rangle$. We can also write the unitary $U_{n-\kappa+1}$ as follows

$$U_{n-\kappa+1} = \sum_{\substack{s_{n-\kappa},\ldots s_{n-1} \\ t_{n-\kappa},\ldots,t_{n-1}}} (U_{n-\kappa+1})^{(s_{n-\kappa},\ldots,s_{n-1})}_{(t_{n-\kappa},\ldots,t_{n-1})} |s_{n-\kappa},\ldots,s_{n-1}\rangle\langle t_{n-\kappa},\ldots,t_{n-1}| \ .$$

Writing the product of these, we have the following state after applying a single unitary

$$U_{n-\kappa+1}(|0\rangle\,|\psi\rangle) = \sum_{s_{n-\kappa},\ldots,s_n} \sum_{t_{n-\kappa-1},\ldots,t_{n-1}} \left((U_{n-\kappa+1})^{(s_{n-\kappa},\ldots,s_{n-1})}_{(0,t_{n-\kappa+1},\ldots,t_{n-1})} \alpha_{t_{n-\kappa+1},\ldots,t_{n-1},s_n} \right) |s_{n-\kappa},\ldots,s_n\rangle$$

Then, writing down the 3-tensor with entries

$$\begin{aligned} \left(A_1^{(s_n)}\right)_{1,(t_{n-\kappa},\ldots,t_{n-1})} &= \alpha_{t_{n-\kappa+1},\ldots,t_{n-1},s_n} \\ \left(A_2^{(s_{n-\kappa},\ldots,s_{n-1})}\right)_{(t_{n-\kappa+1},\ldots,t_{n-1})} &= (U_{n-\kappa+1})^{(s_{n-\kappa},\ldots,s_{n-1})}_{(0,t_{n-\kappa+1},\ldots,t_{n-1})}, \end{aligned}$$

we find that this state is exactly a matrix product state, where the bond dimension is equal to the number of entries in the inner sum, or $d^{\kappa-1} \leqslant \frac{9dr^4n^2}{\varepsilon^2}$ (note that taking the ceiling in the definition of κ causes us to incur an additional factor of d here). Also note that the two tensors are indexed by disjoint registers, as desired. Applying this idea recursively to get the entries of every tensor A_i, we find that we can express every tensor for, $i \geqslant 2$, as follows

$$\left(A_i^{(s_{n+1-i})}\right)^{(a_{n-\kappa+2-i},\ldots,a_{n-i})}_{(b_{n-\kappa+3-i},\ldots,b_{n+1-i})} = (U_{n-\kappa+3-i})^{(a_{n-\kappa+2-i},\ldots,a_{n-i},s_{n+1-i})}_{(0,b_{n-\kappa+3-i},\ldots,b_{n+1-i})}.$$

Writing out the matrix product state that results from these tensors, we will get $|\hat{\phi}\rangle$, and this is a matrix product state with open boundary condition, with bond dimension $dn^2\,\mathrm{poly}(r, 1/\varepsilon)$ since the dimensions of both the a's and b's is $d^{\kappa-1}$. Note that we can reverse the order of the s_i to write this in the canonical form from Definition B.1. This completes the proof that we have produced a matrix product state with the desired bond dimension. ♦

Algorithm B.8 (Agnostic learning of matrix product states)**.**

Input: Copies of an unknown quantum state $\rho \in \mathsf{R}_1 \otimes \ldots \otimes \mathsf{R}_n$ where each R_i is dimension d, such that there exists matrix product state with bond dimension at most r with fidelity η with ρ, and error parameter ε and failure probability δ.

Output: A description of a MPS

Procedure:

1: Let $\tau = \frac{\varepsilon^2}{9n^2r^4}$;
2: Let $\kappa = \lceil \log_d(1/\tau) \rceil + 1$;
3: Let $\rho'_0 = \rho$;
4: **for** i from 1 to $n - \kappa$ **do**
5: Let $\sigma_i = \mathrm{tr}_{\geqslant \kappa}\left(\rho'_{i-1}\right)$;
6: Let $\hat{\sigma}$ be the output of Tomography with error τ and failure probability δ/n on $O\left(\frac{d^3}{\tau^5}\log(n/\delta)\right)$ copies of σ_i;
7: Let U_i be the disentangling unitary for the extension of the $\geqslant \tau$ subspace of $\hat{\sigma}_i$;
8: Apply U_i to all copies of ρ'_{i-1} and project onto $|0\rangle\langle 0|$ to get copies of

$$\rho'_i = \mathrm{tr}_{\mathsf{R}_{i-1}}\Big((|0\rangle\langle 0| \otimes I)U_i\rho'_{i-1}U_i^\dagger(|0\rangle\langle 0| \otimes I)\Big);$$

9: Let ρ^* be the remaining state on the final κ qubits after applying all of the disentangling unitaries and projecting onto $|0^{n-\kappa}\rangle$:

$$\rho^* = \mathrm{tr}_{\mathsf{R}_{<n-\kappa}}((|0^{n-\kappa}\rangle\langle 0^{n-\kappa}| \otimes I)U_{n-\kappa}\ldots U_1\rho U_1^\dagger \ldots U_{n-\kappa}^\dagger(|0^{n-\kappa}\rangle\langle 0^{n-\kappa}| \otimes I));$$

10: Let $\hat{\rho}^*$ be the output of Tomography on with error parameter τ and failure probability δ/n on $O(\frac{d^3}{\tau^5}\log(n/\delta))$ copies of ρ^*.
11: Let $|\psi\rangle$ be the top eigenvalue of $\hat{\rho}^*$;
12: **return** the MPS $U_1^\dagger \ldots U_{n-\kappa}^\dagger(|0^{n-\kappa}\rangle \otimes |\psi\rangle)$.

We now proceed with the proof of correctness for this algorithm.

Lemma B.9. *Let ρ, σ be unnormalized mixed states satisfying $\|\rho - \sigma\|_1 \leqslant \eta$. Let W be the span of all eigenvectors of σ with eigenvalues exceeding η. Then, letting Π_W denote orthogonal projection onto W, we have that $\|(I - \Pi_W)^\dagger\rho(I - \Pi_W)\|_\infty \leqslant 2\eta$.*

Proof. Suppose not, i.e. there exists some state $|\phi\rangle$ so that $|\phi\rangle$ is orthogonal to W and so that $\langle\phi|\rho|\phi\rangle > 2\eta$. But then by duality, we have

$$\|\rho - \sigma\|_1 \geqslant \mathrm{tr}((\rho - \sigma)\,|\phi\rangle\langle\phi|) > \eta\ ,$$

which is a contradiction. ♦

We will prove the correctness of this algorithm inductively. For all $i = 1, \ldots, n - \kappa + 1$, define the matrices

$$E_i = U_i \cdot \ldots \cdot U_1\ , \text{and}\ \ \rho_i = E_i^\dagger\Big(|0^i\rangle\langle 0^i| \otimes \rho'_i\Big)E_i\ . \tag{20}$$

Note that by construction, U_i only touches qudits i through $i + \kappa$. As an immediate consequence of this, E_i acts nontrivially only on the first $i + \kappa$ qudits.

We will say that a call to Tomography *succeeds* if it outputs a $\widehat{\sigma}_i$ satisfying Equation (19). By a union bound, since we do at most n calls to Tomography, all calls succeed simultaneously with probability at least $1 - \delta$. For the rest of the section, condition on the event that this occurs. We first observe the following:

Lemma B.10. *For all $i = 1, \ldots, n - \kappa + 1$, we have that $\rho_i \preceq \rho_{i-1}$.*

Proof. By properties of post-selection, we have that $|0\rangle\langle 0| \otimes \rho'_i \preceq U_i \rho'_{i-1} U_i^\dagger$. The claim then follows from unraveling the definitions. ♦

Our main claim is the following:

Lemma B.11. *Fix $i \in \{1, \ldots, n - \kappa + 1\}$. Let E_i and ρ_i be defined as above, and suppose that every step of* Tomography *succeeds. Suppose that $|\phi\rangle$ has Schmidt rank at most r at position $i + \kappa$. Then,*

$$|\langle\phi|\rho_{i-1}|\phi\rangle - \langle\phi|\rho_i|\phi\rangle| \leqslant \frac{\varepsilon}{2n} .$$

Proof. Let $|\phi'\rangle = E_{i-1}|\phi\rangle$. Since E_{i-1} only acts non-trivially on the first $i + \kappa - 1$ qudits, $|\phi'\rangle$ has the same Schmidt rank as $|\phi\rangle$ at position $i + \kappa$, so in particular it has Schmidt rank at most r at position $i + \kappa$. Write $|\phi'\rangle = |0^{k-1}\rangle|\psi\rangle + |\perp\rangle$, where $|\perp\rangle$ is orthogonal to all states beginning with $|0^{k-1}\rangle$. By definition, $|\psi\rangle$ has Schmidt rank at most r at position κ. By definition, we have that

$$\langle\phi|\rho_{i-1}|\phi\rangle = \langle\psi|\rho'_{i-1}|\psi\rangle ,$$

and additionally, we have that

$$\begin{aligned}
\langle\phi|\rho_i|\phi\rangle &= \langle\phi'| U_i^\dagger \left(|0^i\rangle\langle 0^i| \otimes \rho'_i\right) U_i |\phi'\rangle \\
&= \langle\psi| U_i^\dagger \left(|0\rangle\langle 0| \otimes \rho'_i\right) U_i |\psi\rangle \\
&= \langle\psi| (\Pi_W \otimes I)^\dagger \rho'_{i-1} (\Pi_W \otimes I) |\psi\rangle ,
\end{aligned}$$

where in the second line we use the fact that the extension of U_i acts as the identity outside of qudits i through $i + \kappa$, the third line follows because ρ'_i is obtained by postselecting on outcome $|0\rangle$, and the last line follows since Tomography succeeds, and Lemmas B.9 and B.12, the latter of which is proven below. ♦

Lemma B.12. *Let A, B be two Hilbert spaces, and let ρ be a density matrix over $A \otimes B$. Let $|\phi\rangle$ be a pure state in $A \otimes B$ with Schmidt rank at most r. Let $W \subset A$ be a subspace, with Π_W denoting the projection onto W, such that $\|(I - \Pi_W)^\dagger \operatorname{tr}_B(\rho)(I - \Pi_W)\|_\infty \leqslant \eta$. Then*

$$\left|\langle\phi| (\Pi_W \otimes I)^\dagger \rho (\Pi_W \otimes I) |\phi\rangle - \langle\phi|\rho|\phi\rangle\right| \leqslant 2r\sqrt{\eta} .$$

Proof. First, assume that $|\phi\rangle$ has Schmidt rank 1, i.e. $|\phi\rangle = |a\rangle|b\rangle$. Let $|c\rangle = |a - \Pi_W a\rangle$. Then

$$\begin{aligned}
\left|\langle\phi| (\Pi_W \otimes I)^\dagger \rho (\Pi_W \otimes I) |\phi\rangle - \langle\phi|\rho|\phi\rangle\right| &= |\langle b| \langle \Pi_W a| \rho |\Pi_W a\rangle |b\rangle - \langle b| \langle a| \rho |a\rangle |b\rangle| \\
&\leqslant |\langle b| \langle c| \rho |\Pi_W a\rangle |b\rangle| + |\langle b| \langle a| \rho |c\rangle |b\rangle| \\
&\leqslant 2\|\rho |c\rangle |b\rangle\| ,
\end{aligned}$$

where the last line follows from Cauchy-Schwarz. To finish, we observe that

$$\begin{aligned}
\|\rho |c\rangle |b\rangle\|^2 &= \operatorname{tr}(\rho^2 |c\rangle\langle c| \otimes |b\rangle\langle b|) \\
&\leqslant \operatorname{tr}(\rho |c\rangle\langle c| \otimes |b\rangle\langle b|) \\
&\leqslant \langle c| \operatorname{tr}_B(\rho) |c\rangle \\
&\leqslant \eta ,
\end{aligned}$$

since c is orthogonal to W. The case of general Schmidt rank then follows from linearity. ♦

Proof of Theorem B.2. We first concern ourselves with the sample complexity and runtime of Algorithm B.8. From Lemma B.6, taking the dimension to be $d^\kappa \leqslant d/\tau$, error parameter to be τ and failure probability δ/n, we can perform tomography on the sub-normalized state σ_i with sample complexity $O(\frac{d^3}{\tau^5}\log(n/\delta)) = O\left(\frac{d^3 n^{10} r^{10}}{\varepsilon^{10}}\log(n/\delta)\right) = \mathrm{poly}(n, d, r, 1/\varepsilon, \log(1/\delta))$ (note that the algorithm takes samples of a unitary applied to ρ, the original state), and runtime that is polynomial in the same parameters. The overall algorithm performs this $n - \kappa$ times. In addition to tomography, the algorithm also applies disentangling unitaries to the state. From Lemma B.5, given a description of a subspace W of dimension $d^{\kappa-1}$, there is an implementation of the disentangling unitary in time $\mathrm{poly}(d^\kappa) = \mathrm{poly}(d, r, n, \varepsilon)$, when d is a power of 2. Thus, the whole algorithm run in polynomial time and uses a polynomial number of samples of ρ.

We now turn our attention to correctness. Let $|\phi\rangle$ be a matrix product state with bond dimension r. Since such a MPS has Schmidt rank at most r at every position, by iteratively applying Lemma B.11 with our chosen parameters, we obtain that $|\langle\phi|\rho|\phi\rangle - \langle\phi|\rho_{i-\kappa+1}|\phi\rangle| \leqslant \varepsilon/2$. Together with the fact that the last call to Tomography succeeds, this implies that the vector $|\psi\rangle$ satisfies $\langle\psi|\rho'_{n-\kappa+1}|\psi\rangle \geqslant \langle\phi|\rho|\phi\rangle - \varepsilon$. To conclude, we apply Lemma B.10 and the previous bound to obtain that

$$\begin{aligned}\langle\hat{\phi}|\rho|\hat{\phi}\rangle &= \langle\hat{\phi}|\rho_0|\hat{\phi}\rangle \\ &\geqslant \langle\hat{\phi}|\rho_{n-\kappa+1}|\hat{\phi}\rangle \\ &\geqslant \langle\phi|\rho|\phi\rangle - \varepsilon.\end{aligned}$$

The result follows by taking a supremum over all matrix product states. ♦