Recent Breakthroughs in Quantum Hamiltonian Learning

Alvan Caleb Arulandu

20 December 2024

Abstract

We survey a series of works on the quantum Hamiltonian learning problem, culminating with the recent groundbreaking polytime Sum-of-Squares algorithm. We hope that our exposition clarifies the difficulties of extending techniques in prior work, motivates the algorithm's use of SoS relaxation, and frames the theoretical implications of the work with respect to quantum many-body systems and theoretical computer science.

1 Introduction

The classical *Ising model* has become the centerpiece of statistical mechanics, the simplest undirected graphical model to discuss phase transitions and mean-field theory. For a system of N spins $x_i \in \{\pm 1\}$ with *pairwise interactions* $J_{ij} \in \mathbb{R}$ and *external field* $h_i \in \mathbb{R}$, the associated *Gibbs distribution* over system configurations is given by

$$\mu(x) = \frac{1}{Z}\exp(-\beta H(x)), \quad H(x) \triangleq -\sum_{i,j} J_{ij}x_ix_j - \sum_i h_ix_i \tag{1}$$

where $Z \in \mathbb{R}$ is the *partition function* ensuring $\int \mu(x)dx = 1$, β is the inverse-temperature of the system, and H is the *Hamiltonian*, i.e. a description of the energy of the system for a given configuration x. A natural question, known as *structure learning*, asks to learn the entire Hamiltonian H to ℓ_∞ error ϵ given samples from the Gibbs distribution, $\mu(x)$. A series of works [Bre15; KM17; VMLC16] have given algorithms that achieve polytime structure learning for Ising models and other graphical models using optimal sample and time complexity with respect to N.

However, such classical Hamiltonians, known as Markov random fields, only make up an exponentially small subset of what are known as quantum Hamiltonians. The *quantum Ising model*, or transverse field Ising model, generalizes beyond discrete spins to quantum states $|x\rangle \in \mathbb{C}^2$ of spin-1/2 particles such that $|x\rangle = a\,|0\rangle + b\,|1\rangle$ for $|a|^2 + |b|^2 = 1$, where $|0\rangle$ and $|1\rangle$ are the spin "up" and "down" basis states respectively. The analog to (1) is the following *quantum Hamiltonian.*

$$H = -\beta J\left(\sum_{i,j} Z_iZ_j + g\sum_j X_j\right) \tag{2}$$

The algorithms for quantum Hamiltonian learning discussed in this paper rely on majorly classical insights, so it suffices for the classical reader to blackbox certain quantum tools. For the sake of brevity, quantum operations are represented as Hermitian matrices in the aforementioned standard basis, where the basis states of a multi-qubit quantum system are formed via tensor product. I, X, Y, Z, in particular, are known as the Pauli matrices, and we frequently restrict ourselves to working with Hamiltonians with terms that are Paulis or tensor products of Paulis, as any Hermitian matrix can be decomposed into a linear combination of such matrices. While we defer the full introduction of quantum information and computation to the standard text of Nielsen and Chuang [NC10], crucially, the quantum Ising model, and quantum Hamiltonians more broadly, are exponentially large as the state space of a n-qubit quantum system lives in $\mathbb{C}^N$ where $N = 2^n$ by the nature of the tensor product: $C^2 \otimes \cdots \otimes C^2$. For problems involving Hamiltonians parameterized by a poly(n) description, any efficient solution must be polynomial with respect to n, meaning that even writing down H in its matrix form is already infeasible.

Analogous to the Gibbs distribution $\mu(x)$, the *Gibbs state* for some quantum system with Hamiltonian H is given by the density matrix $\rho = \frac{e^{-\beta H}}{\mathrm{Tr}[e^{-\beta H}]}$. Naturally, the *quantum Hamiltonian learning* problem asks to use copies, i.e. "samples", of the Gibbs state to learn Hamiltonians of the form $H = \sum_{i=1}^{m} \lambda_i E_i$ where $\lambda_i \in [-1, 1]$ is the *interaction strength* of term E_i. While the corresponding structure learning problem involves learning E_i, quantum Hamiltonian learning refers to the setting where $\{E_i\}$ are distinct and known while we are asked to find accurate estimates for $\{\lambda_i\}$.

While the classical Ising model was conceived to study ferromagnetism, quantum Hamiltonian learning is fundamentally tied to quantum many-body systems of quantum particles that *locally* interact. Such systems can become highly entangled and understanding the effective interactions which fix system properties is highly relevant to our physical understanding of superconductors, superfluidity, and quantum material science.

Eleven days prior to the submission of this work, Google Quantum AI announced Willow [Nev24], a state-of-the-art quantum chip achieving quantum error correction below the surface code threshold [AAAA+24] and record performance for random circuit sampling (RCS). With growth in available quantum volume and multiple industry players, efficient classical verification of quantum advantage is an increasing concern. Recent cryptography research has yielded simple IP protocols for varying degrees of quantumness [BGKP+23; BCMV+21] and there is a general excitement in quantum cryptography regarding the explicit construction of pseudo-random unitaries [MH24]. While a quantum algorithm for learning Hamiltonians is not immediately relevant, general tools for natural quantum learning problems may prove useful in quantum hardness, the construction of benchmarks, and broader progress. In fact, prior to the Sum-of-Squares approach of interest [BLMT24], a prominent survey on quantum learning [AA24] hypothesized that low temperature Gibbs states may be pseudorandom!

This work frames the acclaimed algorithm of Bakshi et al. [BLMT24] against prior work on classical Hamiltonian learning, the first sample-efficient learning algorithm of Anshu et al. [AAKS21], and the improved time-efficient algorithm in the high-temperature regime of Haah, Kothari, and Tang [HKT22]. In the process, we aim to motivate the final construction and proof of the SoS system with respect to sufficient statistics, Taylor expansion, and other tools used by these preceding works.

2 Classical Hamiltonian Learning

We begin by considering parameter learning for classical Markov random fields. A Markov random field (MRF) refers to a hypergraph, $G = (V = [N] = \{1, \ldots, N\}, E)$ with vertices being particles of discrete spin $x_i \in \{\pm 1\}$ for $i \in [N]$ and $S \in E$ implying a $|S|$-way interaction between the respective vertices. Then, we have the following Gibbs distribution.

$$\mu(x) \propto \exp\left(-\beta \sum_{S \in E} \lambda_S x^S\right) \tag{3}$$

where $x^S = \prod_{i \in S} x_i$ and $\lambda_S \in [-1, 1]$. For notational convenience, we also define $x_S = (x_i)_{i \in S}$.

Given G and T samples from the MRF, we aim to estimate λ_S. Since the structure, E, is known, this corresponds to the specific quantum Hamiltonian learning problem where $z^S \mapsto \prod_{i \in S} Z_i$, a product of Paulis. We also define E_i to be the set of hyperedges containing i and N_i to be the neighborhood of i, excluding i. Typically, we are also given some locality guarantee $|S| \le \mathfrak{K}$ and an "average order" parameter $L \triangleq \frac{1}{\mathfrak{K}}(1 + \max_{i \in [N]} |N_i|)$.

Consolidating folklore learning results [Bre15; KM17; VMLC16], Haah, Kothari, and Tang [HKT22] explicitly give the following algorithm, with optimal sample and time complexity with respect to N.

Theorem 2.1 (Learning Parameterized MRFs)**.** *Given $T = \exp(O(\beta)) \log(N/\delta)/(\beta^2 \epsilon^2)$ samples from a low-intersection MRF of the form* (3)*, there exists an algorithm to construct $\{\hat{\lambda}_S\}$ such that $|\hat{\lambda}_S - \lambda_S| \le \epsilon$ for all $S \in E$ with probability $\ge 1 - \delta$ in $O(TN)$ time.*

Proof. If an MRF is low-intersection, it follows that L and $\mathfrak{K}$ are $O(1)$. For vertex $v \in [N]$, define the following conditional expectation with respect to its neighbors via the Markov property, which holds for MRFs.

$$\mu(X_v = x_v | X_{N_v} = x_{N_v}) = \sigma\left(q_x^{(v)}\right), \quad q_x^{(v)} \triangleq 2\beta \sum_{S \in E_v} \lambda_S x^S$$

Now, suppose we have estimates $\{\hat{q}_x^{(v)}\}_v$ for $\{q_x^{(v)}\}_v$ and consider λ_S. Taking some $v \in S$, for all $T \in E_v$ such that $T \neq S$, if $T \not\subseteq S$, take $u \in T \setminus S$ and place $u \in N_{\text{out}}$, but if $T \subsetneq S$, choose $u \in S \setminus T$ and place $u \in N_{\text{in}}$. In the classical 2D-lattice Ising model, S is a connected portion of the lattice, $N_{\text{in}} = S$, and N_{out} is the boundary of S; in general, $N_{\text{in}}, N_{\text{out}}$ are disjoint since the former is $\subset S$ while the latter is $\subset S^c$. We then define $W = \{x \in \{\pm 1\}^N | x_i = 1 \forall i \notin N_{\text{in}} \cup N_{\text{out}}$, which is the set of configurations such that all spins outside the "closure" of S is 1. Then, consider the average $\frac{1}{2\beta}\mathbb{E}_{x\sim W}[x^{N_{\text{in}}} q_x^{(v)}]$. Expanding,

$$\frac{1}{2\beta}\mathbb{E}_{x\sim W}[x^{N_{\text{in}}} q_x^{(v)}] = \left[\mathbb{E}_{x_{N_{\text{in}}}} E_{x_{N_{\text{out}}}}\left[\sum_{T\in E_v} \lambda_T x^{N_{\text{in}}} x^T\right]\right]_{x=\vec{1}} = \left[\mathbb{E}_{x_{N_{\text{in}}}} \sum_{T\in E_v, T\cap N_{\text{out}}=\emptyset} \lambda_T x^{N_{\text{in}}} x^T\right]_{x=\vec{1}}$$

$$= \left[\mathbb{E}_{x_{N_{\text{in}}}} \sum_{T\in E_v, T\cap N_{\text{out}}=\emptyset, T^c\cap N_{\text{in}}=\emptyset} \lambda_T x^{T\setminus N_{\text{in}}}\right]_{x=\vec{1}} = \left[\lambda_S x^{S\setminus N_{in}}\right]_{x=\bar{1}} = \lambda_S$$

Here, the first equality comes from disjointness and standard rules of conditional expectation. The second equality uses the fact that any term with $T \cap N_{\text{out}} \neq \emptyset$ will average to 0. Similarly, for the third equality, any vertex $\notin T$ but $\in N_{\text{in}}$ will average to zero since it will only be present in one of the $x^{\cdots}$ terms of the summand. Finally, the fourth equality comes from the fact that all terms have $N_{\text{in}} \subset T$ and $T \subset N_{\text{out}}^c$, meaning $T = S$ must be true by nested inclusion, and the final step follows by evaluation. Then, if $\hat{q}_x^{(v)}$ is an estimate of $q_x^{(v)}$ to $2\beta\epsilon$ error, by triangle inequality,

$$\left|\frac{1}{2\beta}\mathbb{E}_{x\sim W}[x^{N_{\text{in}}} q_x^{(v)}] - \lambda_S\right| = \frac{1}{2\beta}\left|\mathbb{E}_{x\sim W}[x^{N_{\text{in}}}(\hat{q}_x^{(v)} - q_x^{(v)})]\right| \leq \frac{1}{2\beta}\cdot 2\beta\epsilon \cdot \mathbb{E}_{x\sim W}\left|x^{N_{\text{in}}}\right| \leq \epsilon$$

since $|x^{N_{\text{in}}}| = 1$. Since $|W| \leq 2^{|N_{\text{in}} + N_{\text{out}}|} \leq 2^d = O(1)$, this clever averaging trick does not impact the claimed runtime, and it suffices to estimate $q_x^{(v)}$ to $2\beta\epsilon$ error.

By Claim 4.2 of Klivans and Meka [KM17], $\min(1, |x - y|) \leq \exp(|x| + 3)|\sigma(x) - \sigma(y)|$ for all $x, y \in \mathbb{R}$. Since $|q_x^{(v)}| \leq 2\beta d$ always, with some technical manipulation, estimating the conditional probability $\sigma(q_x^{(v)})$ to $\exp(-2\beta d)\min(0.5, \beta\epsilon)$ error and inverting σ gives an estimate for $q_x^{(v)}$ to $2\beta\epsilon$ error.

Of course, by conditional expectation rules, $\sigma(q_x^{(v)})$ reduces to computing $\mu(X_S = x_S)$ for sets $S = \{v\}$ and $S = N_v \cup \{v\}$. By Lemma 2.1 of Bresler [Bre15], $\min_{b\in\{\pm 1\}} \geq \exp(-2\beta d)/2$ which is derived via the Markov property, Adam's Law, and Jensen's inequality. Using a Chernoff bound and some technical algebra, we find that $O(\exp(O(\beta))\log(N/\delta)/(\beta^2\epsilon^2)$ samples yield good estimates for all of these conditional probabilities. Since we pick a single $v \in S$ for each S and we have constant overhead assuming low-intersection, our time complexity is $O(TN)$. □

In essence, this sub-quadratic algorithm follows from estimating certain conditional marginals, constructing the necessary estimates via clever applications of Markov's property, and applying standard technical tools and Chernoff bounds to derive the runtime analysis. Curiously, if we instead desired estimation in the ℓ_2 norm of the parameters $\vec{\lambda} = (\lambda_S)_{S\in E}$ with full generality, our sample complexity is $\text{poly}(N)$ [AAKS20].

As we will see, the SoS algorithm of Bakshi et al. [BLMT24] will similarly begin by estimating certain marginals using standard quantum algorithms. While classical Gibbs states satisfy the Hammersley-Clifford theorem [CH71], commonly known as the Markov property, this argument fails to generalize as the Markov property does not even approximately hold in the low-temperature quantum Hamiltonian setting [KKB20]. This severely limits the number of clever conditional tricks we can play and forces us to either consider natively quantum algorithms or a wider set of marginals with classical optimization strategies. The algorithm of interest lies in the latter.

3 Sample-efficient Quantum Hamiltonian Learning

Since we desire theory for quantum many-body systems which permit local interactions giving way for global entanglement, we naturally restrict ourselves to what are known as *low-intersection* Hamiltonians.

Definition 1 (Low-intersection Hamiltonian)**.** *The support* $\mathrm{supp}(E_i)$ *of an operator is the subset of qubits on which* E_i *is non-trivial. A Hamiltonian* H *is* $\mathfrak{K}$*-local if* $|\mathrm{supp}(E_i)| \leq \mathfrak{K}$ *for all* i*. The dual interaction graph,* $\mathfrak{G} = (V, E)$*, is an undirected graph with vertices* $V = [m]$ *and edge* $(i, j) \in E$ *if and only if* $\mathrm{supp}(E_i) \cap \mathrm{supp}(E_j) \neq \emptyset$*. Let* $\mathfrak{d}$ *be the maximum degree of* $\mathfrak{G}$*. Then,* H *is low-intersection if both* $\mathfrak{K}$ *and* $\mathfrak{d}$ *are* $O(1)$*.*

A low-intersection Hamiltonian implies that interactions occur between a constant number of particles and each term only involves particles that participate in a constant number of terms. With this in mind, we begin our discussion of the quantum setting with the formal definition of the problem of interest.

Problem 1 (Quantum Hamiltonian Learning)**.** *Let* $H = \sum_{i=1}^{m} \lambda_i E_i \in \mathbb{C}^{N \times N}$ *be a low-intersection Hamiltonian on* n *qubits with known terms* $\{E_i\}_i$ *that are distinct, non-identity Pauli operators with coefficients* $\lambda_i \in \mathbb{R}$ *such that* $|\lambda_i| \leq 1$*. For a fixed* $\epsilon, \delta > 0$*, along with copies of the Gibbs state* $\rho = \frac{\exp(-\beta H)}{\mathrm{Tr}\exp(-\beta H)}$ *at known inverse temperature* $\beta > 0$*, find estimates* $\{\hat{\lambda}_i\}_i$ *such that* $(\hat{\lambda}_i - \lambda_i)^2 \leq \epsilon^2$ *for all* $a \in [m]$*, with proability* $\geq 1 - \delta$*.*

This first sample-efficient algorithm of Anshu et al. [AAKS20] considers the slightly modified setting of estimating $\vec{\lambda}$ in ℓ_2 norm, saturating the poly(N) sample complexity bound for classical Hamiltonians.

Theorem 3.1 (Sample-efficient Hamiltonian Learning [AAKS20])**.** *Problem 1 for estimation in the* ℓ_2 *norm can be solved using*

$$T = \mathcal{O}\left(\frac{e^{\mathcal{O}(\beta^c)}}{\beta^{\tilde{c}}\epsilon^2} \cdot m^3 \cdot \log\left(\frac{m}{\delta}\right)\right)$$

copies of the Gibbs state where $c, \tilde{c} \geq 1$ *are constants depending the Hamiltonian geometry.*

Similar to the classical case, the algorithm proceeds by estimating certain local marginals, $\hat{e}_i \approx e_i = \mathrm{Tr}[\rho E_i]$, which are shown to be sufficient statistics: functions of the data such that the data conditioned on the statistic no longer depends on the parameters. We then consider the optimization problem over states constrained to these estimated statistics and seek to maximize the Shannon entropy; specifically, we consider its dual program.

$$\max_{\sigma} S(\sigma) \text{ s.t. } \mathrm{Tr}[\sigma E_i] = e_i \,\forall i \in [m],\ \sigma > 0,\ \mathrm{Tr}[\sigma] = 1. \tag{4}$$

$$\overset{\mathcal{D}}{\Rightarrow} \lambda^* = \mathrm{argmin}_{\lambda=(\lambda_1,\ldots,\lambda_m)} \log Z_\beta(\lambda) + \beta \sum_{i=1}^{m} \lambda_i e_i \tag{5}$$

$$\overset{\approx}{\Rightarrow} \hat{\lambda} = \mathrm{argmin}_{\lambda=(\lambda_1,\ldots,\lambda_m)} \log Z_\beta(\lambda) + \beta \sum_{i=1}^{m} \lambda_i \hat{e}_i \tag{6}$$

where $S(\sigma) = -\mathrm{Tr}[\sigma \log \sigma]$ is the *von Neumann entropy* of the state σ.

The solution to this optimization is the *maximum entropy estimator*, which is the natural least-biased estimate given the data samples [Jay57; Jay82]. Perhaps inspired by recent work on learning quantum Boltzmann machines via gradient descent [AARK+18], Anshu et al. [AAKS20] employ stochastic gradient descent (SGD) to solve the optimization problem in (6).

The proof of correctness and runtime analysis ultimately reduce to proving that the local marginals are sufficient statistics via Proposition 3.2, bounding the difference between the ideal optimization problem with exact statistics (5) versus realized optimization problem with estimated statistics (6), and analyzing the performance of SGD on the given problem. While we omit some of the technical details of the latter, we focus on the novelty of the former, which motivate the local marginals used in the SoS algorithm of later interest.

Proposition 3.2 (Sufficiency of Local Marginals [AAKS20])**.** *Consider the Gibbs state* $\rho(\lambda) = \frac{\exp(-\beta H(\lambda))}{\mathrm{Tr}\exp(-\beta H(\lambda))}$ *for the parameterized Hamiltonian* $H(\lambda) = \sum_i \lambda_i E_i$*. Let* λ *and* μ *be two sets of parameters such that the local marginals* $\mathrm{Tr}[\rho(\lambda)E_i] = \mathrm{Tr}[\rho(\mu)E_i]$ *agree for all* $i \in [m]$*. Then,* $\rho(\lambda) = \rho(\mu)$ *meaning* $\lambda_i = \mu_i$ *for all* $i \in [m]$*.*

Proof. Consider the relative entropy, also known as the Kullback-Leibler divergence, between the two Gibbs states. Expanding,

$$\begin{aligned} D_{KL}(\rho(\lambda)\|\rho(\mu)) = \mathrm{Tr}[\rho(\mu)(\log\rho(\mu) - \log\rho(\lambda))] &= -S(\rho(\mu)) + \beta\cdot\mathrm{Tr}\left[\rho(\mu)\sum_i \lambda_i E_i\right] + \log Z(\lambda) \\ &= -S(\rho(\mu)) + \beta\cdot\sum_i \lambda_i \mathrm{Tr}\left[\rho(\mu)E_i\right] + \log Z(\lambda) \\ &= -S(\rho(\mu)) + \beta\cdot\sum_i \lambda_i \mathrm{Tr}\left[\rho(\lambda)E_i\right] + \log Z(\lambda) \\ &= -S(\rho(\mu)) + (S(\rho(\lambda)) - \cancel{\log Z(\lambda)}) + \cancel{\log Z(\lambda)} \end{aligned}$$

By the positivity of KL-divergence, proved via Jensen's inequality, we then have $S(\rho(\lambda)) \geq S(\rho(\mu))$. Taking $D_{KL}(\rho(\mu)\|\rho(\lambda))$ with an analogous argument, we have $S(\rho(\mu)) \geq S(\rho(\lambda))$, meaning $S(\rho(\mu)) = S(\rho(\lambda))$. This implies that $D_{KL}(\rho(\mu)\|\rho(\lambda)) = 0$, meaning $\rho(\mu) = \rho(\lambda)$ by Jensen's equality condition, proving the claim. □

For the correctness of optimizing with estimated statistics as constraints, we leverage a key fact that the log-partition function is strongly convex.

Definition 2. *For a convex function $f : \mathbb{R}^m \to \mathbb{R}$ with gradient $\nabla f(x)$ and Hessian $\nabla^2 f(x)$, f is said to be α-strongly convex in its domain if it is differentiable and for all x, y,*

$$f(y) - f(x) \geq \nabla f(x)^\top (y - x) + \frac{1}{2}\alpha\|y - x\|_2^2 \Leftrightarrow \nabla^2 f(x) \succeq \alpha\cdot\mathbb{I}$$

Proposition 3.3. $\log Z(\lambda)$ *is α-strongly convex for $\alpha = e^{-\mathcal{O}(\beta^c)}\beta^{c'}/m$ on $\|\lambda\| \leq 1$ with $\frac{\partial}{\partial\lambda_i}\log Z(\lambda) = -\beta e_i$.*

Proving Proposition 3.3 is an arduous process of relating the Hessian to a variance which is lower bounded via a variety of tools including the quantum belief propagation operator [Has07]. While we defer this proof to Anshu et al. [AAKS20]'s Theorem 28, we are now able to relate (6) to the ideal optimization (5).

Proposition 3.4 (Marginal Error Propagation [AAKS20])**.** *Suppose we have marginal estimates up to error δ, i.e. $|e_i - \hat{e}_i| \leq \delta$ for all $i \in [m]$. Assume that $\log Z(\lambda)$ is α-strongly convex. Then, the error induced by* (6) *versus* (5) *is bounded by*

$$\|\lambda - \hat{\lambda}\|_2 \leq \frac{2\beta\sqrt{m}\delta}{\alpha}$$

Proof. By the nature of $\hat{\lambda}$ being the minimizer of the minand in (6),

$$\log Z(\hat{\lambda}) + \beta\sum_{i=1}^m \hat{\lambda}_i\hat{e}_i \leq \log Z(\lambda^*) + \beta\sum_{i=1}^m \lambda_i^*\hat{e}_i \Rightarrow \log Z(\hat{\lambda}) \leq \log Z(\lambda^*) + \beta\sum_{i=1}^m(\lambda_i^* - \hat{\lambda}_i)\hat{e}_i$$

Then, by Proposition 3.3 and Definition 2 for $y = \hat{\mu}$ and $x = \mu^*$,

$$\log Z(\hat{\lambda}) - \log Z(\lambda^*) \geq -\beta\sum_{i=1}^m(\hat{\lambda}_i' - \lambda_i^*)e_i + \frac{\alpha}{2}\|\hat{\lambda} - \lambda^*\|_2^2$$

Re-arranging and applying the Cauchy-Schwarz inequality,

$$\begin{aligned} \cancel{\log Z(\lambda^*)} - \beta\sum_{i=1}^m(\hat{\lambda}_i - \lambda_i^*)e_i + \frac{\alpha}{2}\|\hat{\lambda} - \lambda^*\|_2^2 \leq \log Z(\hat{\lambda}) &\leq \cancel{\log Z(\lambda^*)} + \beta\sum_{i=1}^m(\lambda_i^* - \hat{\lambda}_i)\hat{e}_i \\ \frac{\alpha}{2}\|\hat{\lambda} - \lambda^*\| &\leq \beta\cdot\sum_{i=1}^m(\hat{\lambda}_i - \lambda_i^*)(e_i - \hat{e}_i) \\ &\leq \beta\|\hat{\mu} - \mu\|_2\cdot\|\hat{e} - e\|_2 \end{aligned}$$

implying $\|\hat{\lambda} - \lambda\|_2 \leq \frac{2\beta\sqrt{m}\delta}{\alpha}$ since $\|\hat{e} - e\|_2 \leq \delta\sqrt{m}$ by triangle inequality, which proves the claim. □

From here, the sample complexity follows almost directly. Notice from Proposition 3.4 that it suffices to estimate e_i within $\delta \leq \frac{\alpha\epsilon}{2\beta\sqrt{m}}$ for $\|\lambda - \hat{\lambda}\|_2 \leq \epsilon$ to hold. This marginal estimation is carried out by modern techniques in the field of *quantum tomography.* One strategy is to group Hamiltonian terms E_i into mutually commuting observables which are simultaneously measured [CW20; BBO20]; broadly, work regarding measurement scheduling, partial tomography, and overlapping tomography are active areas of research. However, a recent breakthrough regarding shadow tomography [HKP20; Aar18] permits finding all marginals at once to accuracy δ with $O(2^{\mathcal{O}(\mathfrak{K})} \log(m)/\delta^2)$ state copies. Substituting α from Proposition 3.3, this gives a final sample complexity of

$$O\left(2^{\mathcal{O}(\mathfrak{K})}\log(m)\left(\frac{e^{-\mathcal{O}(\beta^c)}\beta^{c'}}{m}\cdot\frac{\epsilon}{2\beta\sqrt{m}}\right)^{-2}\right) = \mathcal{O}\left(\mathcal{O}\left(\frac{e^{\mathcal{O}(\beta^c)}}{\beta^{\tilde{c}}\epsilon^2}\cdot m^3\cdot\log\left(\frac{m}{\delta}\right)\right)\right)$$

For geometrically-local Hamiltonians, $m = \Theta(N)$, meaning our algorithm is $O_{\beta,\epsilon,\delta}(N^3 \log N)$ in sample complexity. By reducing to a quantum state discrimination problem via an ϵ-net type argument [AAKS20], we can prove the following sample lower bound.

Theorem 3.5 (Sample Complexity Lower Bound on Hamiltonian Learning)**.** *The number of copies T of the Gibbs state for Problem 1 with estimation in the ℓ_2 norm is lower bounded by*

$$T \geq \Omega\left(\frac{\sqrt{m} + \log(1-\delta)}{\beta\epsilon}\right)$$

While this shows tightness up to polynomial factors, prior to the work of Haah, Kothari, and Tang [HKT22] discussed in Section 4, tightness of this bound in ℓ_2 norm was still open. Note that there are two regimes of large sample complexity. As $\beta \to 0$, the Gibbs state approaches the maximally mixed state, independent of λ, increasing the sample complexity via the $1/\beta^{\tilde{c}}$ term. As $\beta \to \infty$, the Gibbs state is near the ground state space for various choices of λ, resulting in higher sample complexity via $e^{\mathcal{O}(\beta^c)}$.

Crucially, this algorithm is not computationally tractable; in fact, even evaluating the minand in (6) for an SGD iterate involves approximating the log-partition function, $\log Z(\lambda)$, which is NP-Hard [Mon15].

Still, our SoS algorithm of later interest draws strong inspiration from these ideas. Particularly, note that the only quantum portion of the algorithm regarded estimating the marginals e_i via shadow tomography. The breakthrough algorithm does the same; Bakshi et al. [BLMT24] solely leverage quantumness via the same shadow tomography result [HKP20] for a wider set of marginals. As in [AAKS20], the remainder of the algorithm is purely classical.

3.1 Commuting Hamiltonians

In a follow up note, [AAKS21] clarify that computationally efficient structure learning is tractable for commuting Hamiltonians on D-dimensional lattices, that is when the *commutator* $[E_i, E_j] = E_iE_j - E_jE_i = 0$. The key insight is to define an effective Hamiltonian $H_R = -\frac{1}{\beta}\log \operatorname{Tr}_{R^c}(\rho)$ for any region R of the lattice and prove that, beyond a critical temperature [KKB20], this effective Hamiltonian can be decomposed

$$H_R = \alpha_R I + h_R + \Phi$$

into commuting terms Φ, h_R supported on the boundary ∂R and on the region R respectively, followed by an identity term with $\alpha_R \in \mathbb{R}$ such that $\|\Phi\|_\infty \leq 2|\partial R|$. With this tool, for every term E_i of the Hamiltonian H, we take the smallest region $R_i \supset \operatorname{supp}(E_i)$ with $\partial R_i \cap \operatorname{supp}(E_i) = \emptyset$. Noting that $|R_i| \leq (3k)^D$, we divide the $\{R_i\}_i$ into at most $(kD)^D$ batches such that the regions in each batch are non overlapping and then perform tomography to obtain a classical description of the Gibbs state under Hamiltonian $h_{R_i} + \Phi_i$. We then use this classical description to construct an estimate for operator E_i, and with some techincal details, arrive at a sample complexity of $T = e^{\mathcal{O}(\beta)} \log(m/\delta)/\epsilon^2$ and time complexity of $O(mT)$, which is polynomial given $m = \operatorname{poly}(N)$.

4 High-Temperature Tractability

In Remark 4.5, Haah, Kothari, and Tang [HKT22] modify the final bounding step of Anshu et al. [AAKS21]'s proof of our Proposition 3.3 to yield a sample complexity of $O(2^{\mathrm{poly}(\beta)} N^2 \log(N)/(\beta^2\epsilon^2))$ for estimation in ℓ_∞ norm. Even still, compared to the classical setting, this has worse numerator and denominator dependence in β, with $N^2 \log(N)$ dependence compared to the logarithmic sample complexity of Theorem 2.1.

Haah, Kothari, and Tang [HKT22] give the first optimal learning algorithm in the high-temperature regime for both ℓ_2 and ℓ_∞ coefficient error.

Theorem 4.1 (High-Temperature Algorithm [HKT22])**.** *For low-intersection Hamiltonian H on N qubits, $\epsilon > 0$, and $\beta < \beta_c$, we can learn coefficients of H with ℓ_∞ error ϵ and failure probability δ using $T = O(\frac{1}{\beta^2\epsilon^2} \log \frac{N}{\delta})$ samples. For learning with ℓ_2 error ϵ, we require $O(\frac{N}{\beta^2\epsilon^2} \log \frac{N}{\delta})$ samples, and in both cases, the time complexity is $O(TN)$.*

Theorem 4.2 (High-Temperature Sample Lower Bound [HKT22])**.** *For any $\epsilon \in (0, 1/2]$, $\beta > 0$, and N, there exists a 2-local Hamiltonian on N qubits that requires sample complexity $\Omega\left(\frac{\exp(\beta)}{\beta^2\epsilon^2} \log \frac{N}{\delta}\right)$ and $\Omega\left(\frac{\exp(\beta)}{\beta^2\epsilon^2} N\right)$ for ℓ_∞ and ℓ_2 error ϵ respectively, with failure probability δ.*

Theorem 4.2 significantly improves the sample complexity lower bound of Theorem 3.5 via a simple 2-local Hamiltonian construction, a controlling of KL-divergence, and a powerful application of an information-theoretic result known as Fano's Lemma. Of course, this implies that the algorithm of Theorem 4.1 is tight, up to a $\log N$ factor in only the ℓ_2 error case.

The foundational insight of Theorem 4.1 is that the Taylor series in β of the expectation $\mathrm{Tr}(E_i\rho)$ converges in the high-temperature regime, as β being the inverse-temperature is near zero. Following cluster expansion techniques from [KS20], Haah, Kothari, and Tang [HKT22] rigorously show that the expectation converges and give an explicit algorithm for computing the Taylor expansion when $\{E_i\}$ are Pauli operators. Via the quantum shadow tomography methods used in Anshu et al. [AAKS20], we then estimate $\hat{e}_i \approx e_i = \mathrm{Tr}(E_i\rho)$ while also expanding this expectation as a polynomial in $\{\lambda_i\}$ by truncating the respective Taylor series. This polynomial system is then classically solved via the Newton-Raphson root-finding method which can be showed to converge in $O(\log(1/(\beta\epsilon)))$ iterations, which is negligible relative to the problem input.

We focus on the sample complexity bound which prioritizes themes of Taylor expansion, embodied in the following Lemma.

Lemma 4.3 (High-Temperature Sample Complexity [HKT22])**.** *Suppose $\{E_i\}$ are traceless and orthonormal with respect to the Hilbert-Schmidt inner product. Then, for any β such that $100e^6(\mathfrak{d}+1)^8\beta) \leq 1$, we can find $x \in [-1,1]^m$ such that $\|x - \lambda\|_\infty \leq \epsilon$ with failure probability δ using*

$$O\left(\frac{\mathfrak{d}}{\beta^2\epsilon^2} \log\left(\frac{m}{\delta}\right)\right)$$

copies of the Gibbs state.

Proof. As in Theorem 3.1, computing $\hat{e}_i$ to error $\beta\epsilon$ of e_i for all i, with failure probability δ, requires $T = O\left(\frac{\mathfrak{d}}{\beta^2\epsilon^2} \log\left(\frac{m}{\delta}\right)\right)$ copies. Let $\mathcal{F}_i : [-1,1]^m \to \mathbb{R}^m$ be the Taylor series expansion of $\mathrm{Tr}(E_i\rho)$ with respect to β, truncated at $\mathfrak{m}$ terms and shifted by $\hat{e}_i$ such that

$$\mathcal{F}_i(x) \triangleq -\hat{e}_i + \sum_{k=1}^{\mathfrak{m}} \beta^k p_k^{(a)}(x)$$

Let $\mathcal{F}(x) = (\mathcal{F}_0(x), \ldots, \mathcal{F}_m(x))^\top$ and consider any x such that $\|\mathcal{F}(x)\|_\infty = O(\beta\epsilon)$. Our algorithm will be to compute such an x via Newton-Raphson and prove its time complexity and correctness, but since we only consider sample complexity at the moment, we take $\mathfrak{m} \to \infty$ and it suffices to show that for all such x, x must be close to λ.

Of course, we know that $x = \lambda$ satisfies this inequality, so we aim to use an intermediate value style argument. Let $J = d\mathcal{F}$ be the Jacobian of $\mathcal{F}$ such that $J_{ij} = \partial_j \mathcal{F}_i$. Then, for each $i \in [m]$, by multivariate mean value theorem, we have some $\nu(i) \in (-1,1)^m$ such that

$$\mathcal{F}_i(x) - \mathcal{F}_i(\lambda) = (J|_{\nu(i)}(x-\lambda))_i \Rightarrow |x_i - \lambda_i| = \left|\sum_j (J|_{\nu(i)}^{-1})_{ij}(\mathcal{F}_j(x) - \mathcal{F}_j(x))\right|$$
$$\leq \|J|_{\nu(i)}^{-1}\|_{\infty\to\infty}(\|\mathcal{F}(x)\|_\infty + \|\mathcal{F}(\lambda)\|_\infty) \leq 2\beta^{-1}(2\beta\epsilon) = 4\epsilon$$

where we apply triangle inequality with the fact that $\|J(x)^{-1}\|_{\infty\to\infty} \leq 2\beta^{-1}$ for $\mathfrak{m} \geq 1$ via Lemma 4.3 of Haah, Kothari, and Tang [HKT22], proven via the band-diagonal property of J. This proves the claim. □

As we waived SGD earlier, we also defer the proof of time complexity and convergence of the Newton-Raphson method to Haah, Kothari, and Tang [HKT22]'s Section 4. The main tool here is a strong convexity bound similar to Proposition 3.3 for $\alpha = \beta^2/2$ in the high-temperature regime. This removal of a factor of m^{-1} in α plays a major role in achieving the optimal time and sample complexity.

4.1 Learning from Real-Time Evolution

As an aside, Haah, Kothari, and Tang [HKT22] also show time-optimal Hamiltonian learning from real-time dynamics.

Theorem 4.4 (Learning Hamiltonians from Real-Time Dynamics [HKT22])**.** *Give a low-intersection Hamiltonian H on N qubits and a blackbox unitary $U = e^{-itH}$ for $t < t_c$, we can learn coefficients of H to ℓ_∞ error ϵ with failure probability δ using $T = O(\frac{1}{t\epsilon^2}\log\frac{N}{\delta})$ query complexity, of U, and $O(NT)$ time.*

Here, e^{-itH} is the real-time evolution operator which evolves a quantum state according to Hamiltonian H for time t according to Schrödinger's equation. Similar to Theorem 4.1, proof of Theorem 4.4 analyzes the matrix-valued polynomial expansion of $UPU^\dagger$ with respect to t.

Overall, this idea of building a polynomial system from estimated and approximate expansions of marginals is exactly what is used in the SoS algorithm of later interest, though in the low-temperature regime, such a Taylor expansion fails to converge as β is large. Regardless, the tools like cluster-related arguments which are used to explicitly construct the Taylor expansion in Haah, Kothari, and Tang [HKT22] are used to bound relevant terms in some technical details of Bakshi et al. [BLMT24].

5 Efficient Learning at Any Temperature

As we've seen, polytime Hamiltonian learning at low temperatures has remained elusive, even though this is the most important problem regime. Most quantum many-body systems operate at high β, yielding macroscopic phenomena [BLMT24], and we typically seek quantum advantage in the low temperature regime as the high temperature setting is often classically simulable. Yet, partition function computation for Theorem 3.1 is computationally intractable, Taylor expansion for Theorem 4.1 does not converge for high β, and traditional cluster expansion arguments for Theorem 4.1 also fails.

Theorem 5.1 (Efficient Learning at Low Temperature [BLMT24])**.** *Given $\epsilon > 0, \beta \geq \beta_c$, and $\mathfrak{K}$-local Hamiltonian with dual interaction graph of max degree $\mathfrak{d}$, there exists an algorithm to estimate $\{\hat{\lambda}_i\}_i$ such that with failure probability δ, $(\hat{\lambda}_i - \lambda_i)^2 \leq \epsilon^2$ for all $i \in [m]$ using*

$$O\left(\left((m^6/\epsilon^{e^{f(\mathfrak{K},\mathfrak{d})\beta}}) + (f(\mathfrak{K},\mathfrak{d})/(\beta^2\epsilon^2))\right)\log(m/\delta)\right)$$

copies of the Gibbs state with running time

$$\mathrm{poly}(m, \log(1/\delta))(1/\epsilon)^{e^{f(\mathfrak{K},\mathfrak{d})\beta}} + f(\mathfrak{K},\mathfrak{d})(m/(\beta^2\epsilon^2))\log(m/\delta)$$

where $f(\mathfrak{K},\mathfrak{d})$ is a positive function depending only on $\mathfrak{K}$ and $\mathfrak{d}$.

Here, β_c is the same critical temperature as Theorem 4.1 such that for $\beta < \beta_c$ we simply run the optimal high-temperature algorithm [HKT22].

Of course, nearly three years after the sample-efficient algorithm [AAKS20], Theorem 5.1 was a delightful shock to the quantum learning community. Given recent hardness hypotheses regarding the low temperature regime [AA24], not only was Theorem 5.1 suprising in itself, but the particular method, Sum-of-Squares relaxation, was not expected to yield success.

However, in the context of prior work, Sum-of-Squares seems at least some what motivated. Particularly, the key insight of Bakshi et al. [BLMT24] is not just to employ Sum-of-Squares but to widen the set of statistics beyond simply $\hat{e}_i \approx e_i = \mathrm{Tr}[E_i\rho]$ as both Anshu et al. [AAKS20] and Haah, Kothari, and Tang [HKT22] had done. Specifically, consider the wider constrain system:

$$\begin{cases} \forall i \in [m] & -1 \leq \lambda_i' \leq 1 \\ & H' = \sum_{i\in[m]} \lambda_i' E_i, \\ \forall P, Q \in \mathcal{P}_{\text{local}}, & \mathrm{Tr}(Qe^{-\beta H'}Pe^{\beta H'}\rho) = \mathrm{Tr}(PQ\rho) \end{cases} \tag{7}$$

where $\mathcal{P}_{\text{local}}$ is the set of Pauli matrices with K-local support. Notice that the true parameters $\lambda' = \lambda$ satisfy the final constraint by the cyclic property of the trace

$$\mathrm{Tr}(Qe^{-\beta H'}Pe^{\beta H'}\rho) = \mathrm{Tr}\left(Qe^{-\beta H}Pe^{\beta H} \cdot \frac{e^{-\beta H}}{\mathrm{Tr}[e^{-\beta H}]}\right) = \mathrm{Tr}(Q\rho P) = \mathrm{Tr}(PQ\rho)$$

The third constraint of (7) widens the typical polynomial system involving expansions of $\mathrm{Tr}[E_i\rho]$. For tractability, we must consider approximate low degree polynomial constraints to replace the trace term involving matrix exponentials $e^{\pm\beta H'}$; the intuition of adding constraints is to permit these approximations without failing correctness. We then take this polynomial system in $\{\lambda'\}$ and apply a convex relaxation technique known as Sum-of-Squares. Proving correctness then reduces to constructing these polynomial approximations and presenting relevant "SoS proofs" for extracting the intended solution out of the relaxed system.

5.1 Constructing a Polynomial System

Desiring to work with polynomials over $\mathbb{C}$, we begin by relating third constraint of (7) to what is known as a *nested commutator polynomial.*

Definition 3 (Nested Commutator Polynomial)**.** *For matrices $A, B \in \mathbb{C}^{n\times n}$, the nested commutator is defined as $[A,B]_k \triangleq [A,[A,B]_{k-1}]$ for $k \geq 1$ and $[A,B]_0 = B$. For polynomial $p(x) = \sum_{k=0}^d a_k x^k$, $p(X|A) = \sum_{k=0}^d a_k [X,A]_k$ is the respective nested commutator polynomial.*

The nested commutator relates to the Hadamard product, or element-wise product. Note that in the basis where A is diagonal such that $A_{ii} = \alpha_i$,

$$[A,B] = AB - BA = B \circ \{\alpha_i\}_{ij} - B \circ \{\alpha_j\}_{ij} = B \circ \{(\alpha_i - \alpha_j)\}_{ij}$$

By induction, it follows that $[A,B]_k = B \circ \{(\alpha_i - \alpha_j)^k\}_{ij}$ meaning $p(H'|P) = P \circ \{p(\sigma_i - \sigma_j)\}_{ij}$ where σ_i are the eigenvalues of H'. As it turns out, nested commutators relate quite naturally to the central P conjugated by $e^{\beta H'}$ via the Hadamard formula:

$$e^{-\beta H'}Pe^{\beta H'} = \sum_{\ell=0}^{\infty} \frac{\beta^\ell [H',P]_\ell}{\ell!} = P \circ \{e^{\beta(\sigma_i - \sigma_j)}\}_{ij} = (e^{\beta x})(H'|P) \tag{8}$$

Thus, via nested commutators, we have shown that reducing our third constraint to a polynomial in $\{\lambda_i'\}$ hinges on a clever approximation of $e^{\beta x}$. Specifically, via the band-diagonal property of small support operations in the basis of H' [AKL16], which was also used in the technical detail of Lemma 4.3, it holds that in the eigenbasis $\{v_i\}$ of H',

$$|P_{ij}| = |v_i^\top P v_j| \leq e^{-\Omega(|\sigma_i - \sigma_j|)}$$

Then,

$$p(H'|P) - (e^{\beta x})(H'|P) = P \circ \{p(\sigma_i - \sigma_j) - \exp(-\beta(\sigma_i - \sigma_j))\}_{ij}$$

is the error of our approximation, where each term is weighted inverse exponentially in $\sigma_i - \sigma_j$. Thus, we desire a *flat* approximation that is strong near 0 and may slowly diverge outside this neighborhood.

Definition 4 (Flat Exponential Approximation)**.** *Given $\epsilon, \eta \in (0,1)$ and $\kappa \geq 1$, we say a polynomial $p(x)$ is a (κ, η, ϵ)-flat exponential approximation if $|p(x) - e^x| \leq \epsilon$ for $x \in [-\kappa, \kappa]$, meaning it is controlled around 0, and $|p(x)| \leq \max(1, e^x)e^{\eta|x|}$, meaning that it may diverge at most exponentially.*

Constructing such polynomials are quite hard and standard Taylor or Chebyshev truncations fail on the negative tail, which exponentially blows up. Inspired by an idea known as "iterative peeling" used in Lieb-Robinson bounds [LR72; Has10], a bound used in the technical details of [AAKS20], Bakshi et al. [BLMT24] product multiple Taylor expansions truncated at varying degrees such that the error tails do not constructively interfere.

Theorem 5.2 (Construction of a Flat Exponential Approximation)**.** *Let $s_\ell(x) = \sum_{k=0}^{\ell} \frac{x^k}{k!}$ be the degree-ℓ Taylor truncation of e^x. Let $p_{k,\ell}(x) = \prod_{j=1}^{k} s_{2^j\ell}(x/k)$ and $q_{k,\ell}(x) = 1 + \int_0^x p_{k,\ell}(y)dy$ such that $p_{k,\ell}, q_{k,\ell}$ have degree $(2^{k+1}-1)\ell$ and $(2^{k+1}-1)\ell+1$ respectively. Then, $p_{k,\ell}, q_{k,\ell}$ are (κ, η, ϵ)-flat exponential approximations for $\kappa \geq \max(1, 5/\eta)$ and $\ell \geq 100(\kappa + \log \kappa/\epsilon)$.*

The proof of Theorem 5.2 is an arduous exercise in approximation theory, but with it, we are able to show that $Qp(H|p)\rho \approx Qe^{-\beta H}Pe^{\beta H}\rho$ when p is a flat approximation of small η. Intuitively, near 0, the ϵ-control of p counters the large $\exp(-\beta(\sigma_i - \sigma_j))$ weight while outside the neighborhood, the exponentially small weight controls the exponential divergence of p. With this, we loosen the third equality constraint in (7) to the following polynomial constraint.

$$\begin{cases} \forall i \in [m] & -1 \leq \lambda_i' \leq 1 \\ & H' = \sum_{i \in [m]} \lambda_i' E_i, \\ \forall P, Q \in \mathcal{P}_{\text{local}}, & |\widetilde{\text{Tr}}(Qp(H'|P)\rho) - \widetilde{\text{Tr}}(PQ\rho)|^2 \leq \epsilon^2 \end{cases} \tag{9}$$

where $\widetilde{\text{Tr}}$ is the estimated trace via shadow tomography-like procedures [HKP20] which incurs polynomial sample and time costs.

5.2 Sum-of-Squares Relaxation

We now give a correct algorithm for solving a close variant of (9) via a classical polynomial relaxation technique known as Sum-of-Squares (SoS).

Definition 5 (Pseudo-Distributions and Expectations)**.** *Recall that a degree-ℓ pseudo-distribution is a finitely-support function $D : \mathbb{R}^m \to \mathbb{R}$ such that $\sum_x D(x) = 1$ and $\sum_x D(x)p(x)^2 \geq 0$ for all polynomials p of degree $\leq \ell/2$, where we sum over the support of D. The pseudo-expectation with respect to D, $\hat{\mathbb{E}}_D$, takes any polynomial of degree $\leq \ell$ and outputs $\hat{\mathbb{E}}_D[f(x)] = \sum_x D(x)f(x)$.*

Definition 6 (Constrained Pseudo-distributions)**.** *Given a system $\mathcal{C} = \{p_1 \geq 0, \ldots, p_c \geq 0\}$ of polynomial inequality constraints of degree $\leq d$ in m variables and a degree-ℓ pseudo-distribution, D, over $\mathbb{R}^m$, we say D satisfies $\mathcal{A}$ at degree $\ell \geq 1$ if for every $S \subset [c]$ and sum-of-squares polynomial q with $\deg(q) + \sum_{i \in S} \deg(p_i) \leq \ell$, we have that $\hat{\mathbb{E}}_D[q \prod_{i \in S} p_i] \geq 0$. Moreover, D approximately satisfies $\mathcal{A}$ is $\hat{\mathbb{E}}_D[q \prod_{i \in S} p_i] \geq -2^{-n^\ell}\|q\| \prod_{i \in S} \|p_i\|$ where $\| \cdot \|$ is the ℓ_2 norm of the polynomial coefficients in the standard basis.*

SoS effectively relaxes the notion of a distribution over solutions; for certain polynomial systems, a pseudo-distribution of interest can be efficiently found via semidefinite programming.

Theorem 5.3 (Efficient Computation of Pseudo-distributions)**.** *Given a satisfiable system $\mathcal{C}$ of r constraints in m variables such that at least one is of the form $\|x\|^2 \leq 1$, there exists a $(m + r)^{O(\ell)}$-time algorithm to output a degree-ℓ pseudo-distribution that approximately satisfies $\mathcal{A}$.*

Given a satisfying pseudo-distribution, SoS algorithms extract the final problem solution from the moment tensor $\hat{\mathbb{E}}[(1, x_1, \dots, x_m)^{\otimes \ell}]$. SoS proofs deduce pseudo-distribution properties from constraints via standard inference rules of addition, multiplication, substitution, and transitivity. The SoS proof system is known to be sound and complete.

Algorithm 5.1 Learning a Hamiltonian from Gibbs states [BLMT24]

```
Input: $T = (m^6/\epsilon^{e^{f(\mathfrak{K},\mathfrak{d})\beta}}) \log(m/\delta)$ copies of the Gibbs state $\rho = \frac{e^{-\beta H}}{\mathrm{Tr}(e^{-\beta H})}$ for an unknown low-intersection Hamiltonian $H = \sum_a \lambda_a E_a$ with known terms $\{E_a\}$.
Operation:
  1. Set $\epsilon_0 = \frac{\epsilon^{10^{C_{\mathfrak{K},\mathfrak{d}}}}}{m^3}$ where $C_{\mathfrak{K},\beta}$ is a sufficiently large constant depending only on $\mathfrak{K}, \mathfrak{d}$. Set $\epsilon = 2^{C_{\mathfrak{K},\beta}} \log(1/\epsilon)$, $\ell_1 = 4\mathfrak{K}$. Define $\mathcal{A} = \mathcal{P}_{4^{C_{\mathfrak{K},\mathfrak{d}}\beta}\ell_0}$, $\mathcal{B} = \mathcal{P}_{\ell_1}$.
  2. For all $A_1, A_2, A_3 \in \mathcal{A}$, compute estimates $\widetilde{\mathrm{Tr}}(A_1 A_2 A_3 \rho)$ of $\mathrm{Tr}(A_1 A_2 A_3 \rho)$ to $\epsilon_0$ error via [HKP20].
  3. Consider the following constraint system:
     $$\mathcal{C}_{\lambda'} = \begin{cases} -1 \le \lambda'_i \le 1, & \forall i \in [m], \\ H' = \sum_{i \in [m]} \lambda'_i E_i, & \\ \left| \widetilde{\mathrm{Tr}}\big(A_1 A_2 (H'\rho - \rho H')\big) \right|^2 \le \epsilon_0^2, & \forall A_1, A_2 \in \mathcal{A}, \\ \left| \widetilde{\mathrm{Tr}}\big(B_2 q_{C_{\mathfrak{K},\mathfrak{d}}\beta, \ell_0}(-\beta H' | B_1)\rho\big) - \widetilde{\mathrm{Tr}}(B_1 B_2 \rho) \right|^2 \le \epsilon^2, & \forall B_1, B_2 \in \mathcal{B}. \end{cases}$$
  4. Compute a degree-$\mathcal{O}(2^{C_{\mathfrak{K},\mathfrak{d}}\beta} \ell_0)$ pseudo-distribution $\hat{\mathbb{E}}$ consistent with $\mathcal{C}_{\lambda'}$;
Output: $\hat{\lambda} = \hat{\mathbb{E}}[\lambda']$.
```

As pictured in Algorithm 5.1 5.1, we set up a polynomial constraint system $\mathcal{C}_{\lambda'}$ very similar to (9). With polynomials of degree at most $\mathcal{O}(2^{O(\beta)}\ell_0)$, we then compute a valid degree-$\mathcal{O}(2^{O(\beta)}\ell_0)$ pseudo-distribution $\hat{\mathbb{E}}$ and compute $\hat{\mathbb{E}}[\lambda']$ via the moment tensor to obtain our estimates. By the soundness of SoS, it suffices to give a degree-$\mathcal{O}(2^{O(\beta)}\ell_0)$ SoS proof that $\mathcal{C}_{\lambda'}$ implies correctness.

Theorem 5.4 (SoS Proof of Identifiability [BLMT24])**.** *For any $i \in [m]$, given $\mathcal{C}_{\lambda'}$, there is a degree-$\mathcal{O}(2^{O(\beta)}\ell_0)$ SoS proof that $\{(\lambda_i - \lambda')^2 \le 2^{C_{\mathfrak{K},\mathfrak{d}}\beta}\epsilon\}$.*

This proof of Theorem 5.4 is highly non-trivial. Broadly, the proof requires frequently translating between polynomials and nested commutators. Bakshi et al. [BLMT24] begin by proving various smaller tools regarding bivariate nested commutators, continuing to a proof that $\mathrm{Tr}([H, H'](H'\rho - \rho H'))$ is small, implying that $[H, H']$ is small, which is used to argue that $(\lambda_i - \lambda')^2$ is small. Fundamentally, the particular constraints of $\mathcal{C}$ with respect to the general structure of (9) are constructed in order to aid the SoS proofs and not vice versa. The first trace inequality constraint in particularly is entirely a proof tool. With these changes in the constraint system, a proof of feasibility is also required, though external to SoS.

Theorem 5.5 (Proof of Feasibility [BLMT24])**.** *$\mathcal{C}_{\lambda'}$ is satisfied when $\lambda' = \lambda$, provided $|\widetilde{\mathrm{Tr}}(A_1 A_2 A_3 \rho) - \widetilde{\mathrm{Tr}}(A_1 A_2 A_3 \rho)| \le \epsilon_0$ for all $A_1, A_2, A_3 \in \mathcal{A}$.*

Proof. Since $\lambda' = \lambda$, we have that $H' = H$. Since H commutes with its Gibbs state ρ, $\mathrm{Tr}(A_1 A_2 (H\rho - \rho H)) = 0$ meaning the first trace inequality constraint of $\mathcal{C}_{\lambda'}$ is satisfied. For the remaining trace constraint, in the eigenbasis of H,

$$\begin{aligned} \mathrm{Tr}((B_2 q(-\beta H | B_1) - B_1 B_2)\rho) &= \mathrm{Tr}((B_2 (B_1 \circ \{q(-\beta(\sigma_i - \sigma_j))\}_{ij} - B_1 B_2)\rho) \\ &\approx \mathrm{Tr}((B_2 (B_1 \circ \{\exp(-\beta(\sigma_i - \sigma_j))\}_{ij} - B_1 B_2)\rho) \\ &= \mathrm{Tr}((B_2 \rho B_1 \rho^{-1} - B_1 B_2)\rho) = \mathrm{Tr}(B_2 \rho B_1 - B_1 B_2 \rho) = 0 \end{aligned}$$

where we use the fact that q is a strong flat exponential approximation and $B_1 \circ \{\exp(-\beta(\sigma_i - \sigma_j))\} = \rho B_1 \rho^{-1}$ by matrix multiplication since ρ is the Gibbs state of H. We defer the specifics of propagating these polynomial approximation errors to Bakshi et al. [BLMT24]. □

5.3 Reducing SoS Search Space

For the target runtime of Theorem 5.1, we leverage a technique known as linearization [Mar21] to reduce the degree of our SoS system. This derives from the key observation that only a specific family of monomials, generated via cluster expansion ideas [KS20; HKT22], are involved in the SoS proof of identifiability. From intermediary results in the SoS proof, Bakshi et al. [BLMT24] derive that the number of relevant monomials is at most $m \cdot (1/\epsilon)^{10^{C_{\mathfrak{K},\mathfrak{d}}\beta}}$. This lends itself well to a recent SoS optimization result.

Theorem 5.6 (Degree Reduction via Linearization [Mar21])**.** *Let $p : \mathbb{R}^n \to \mathbb{R}$ be a multivariate polynomial of degree $\leq t$, and let $\mathcal{C} = \{q_1 \geq 0, \ldots, q_c \geq 0\}$ be a system of polynomial equalities such that $\mathcal{C} \vdash \{p(x) \geq 0\}$ and assume the proof can be written as $\sum_i r_i(x)^2 \prod_{j \in S_i} q_j(x)$ where there are at most M distinct sets $S_i \subseteq [c]$ and at most N distinct monomials in $p(x)$. Then, we can write a system $\mathcal{C}'$ in x and some auxiliary variables such that $\mathcal{C}'$ is feasible when $\mathcal{C}$ is feasible, $\mathcal{C}' \vdash \{p(x) \geq 0\}$, and we can compute a pseudo-expectation for $\mathcal{A}'$ in $O(c + M + (tN)^3)$ time.*

Applying Theorem 5.6, Bakshi et al. [BLMT24] arrive at the complexity bounds of Theorem 5.1.

6 Discussion

While it pains us to withhold proof detail from our discussion of the SoS algorithm, we hope it comes at the benefit of a well-motivated framing of quantum learning and its relevant tools. From our understanding, strong sample and time complexity lower bounds are unknown in the low-temperature regime, so we are excited to see greater interest in the setting. While SoS is classical learning theory technique, the meta-algorithm has been used to study best state separation [BKS17] and fermionic Hamiltonians [HO22; Has23] in recent years.

Generally, quantum learning is an emergent field, and questions in state tomography and many-body systems are quite exciting. In fact, some recent work considers the relation between PAC-learning with quantum algorithms and quantum circuit lower bounds [AGGO+22] in complexity theory which is quite curious. Regardless, we highly recommend that motivated readers explore the exposition and proofs presented by Bakshi et al. [BLMT24].

7 Acknowledgment

We would like to thank Professor Sitan Chen, TF Marvin Li, and TF Weiyuan Gong for their support. While this work began as a final project for Harvard's CS 224, it has evolved into a rewarding intellectual exploration, and we are looking forward to continued study of quantum information and avid attendance of CS colloquium talks.

References

[AA24] Anurag Anshu and Srinivasan Arunachalam. "A survey on the complexity of learning quantum states". In: *Nature Reviews Physics* 6.1 (2024), pp. 59–69.

[AAAA+24] Rajeev Acharya et al. "Quantum error correction below the surface code threshold". In: *Nature* (Dec. 2024). ISSN: 1476-4687. DOI: 10.1038/s41586-024-08449-y. URL: https://doi.org/10.1038/s41586-024-08449-y.

[AAKS20] Anurag Anshu et al. "Sample-efficient learning of quantum many-body systems". In: *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2020, pp. 685–691.

[AAKS21] Anurag Anshu et al. "Efficient learning of commuting hamiltonians on lattices". In: *Unpublished notes avaible at Anurag Anshu's website,(link to note)* (2021).

[Aar18] Scott Aaronson. "Shadow tomography of quantum states". In: *Proceedings of the 50th annual ACM SIGACT symposium on theory of computing*. 2018, pp. 325–338.

[AARK+18] Mohammad H Amin et al. "Quantum boltzmann machine". In: *Physical Review X* 8.2 (2018), p. 021050.

[AGGO+22] Srinivasan Arunachalam et al. "Quantum learning algorithms imply circuit lower bounds". In: *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2022, pp. 562–573.

[AKL16] Itai Arad, Tomotaka Kuwahara, and Zeph Landau. "Connecting global and local energy distributions in quantum spin models on a lattice". In: *Journal of Statistical Mechanics: Theory and Experiment* 2016.3 (2016), p. 033301.

[BBO20] Xavier Bonet-Monroig, Ryan Babbush, and Thomas E O'Brien. "Nearly optimal measurement scheduling for partial tomography of quantum states". In: *Physical Review X* 10.3 (2020), p. 031064.

[BCMV+21] Zvika Brakerski et al. "A cryptographic test of quantumness and certifiable randomness from a single quantum device". In: *Journal of the ACM (JACM)* 68.5 (2021), pp. 1–47.

[BGKP+23] Zvika Brakerski et al. "Simple tests of quantumness also certify qubits". In: *Annual International Cryptology Conference*. Springer. 2023, pp. 162–191.

[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*. 2017, pp. 975–988.

[BLMT24] Ainesh Bakshi et al. "Learning quantum Hamiltonians at any temperature in polynomial time". In: *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*. 2024, pp. 1470–1477.

[Bre15] Guy Bresler. "Efficiently learning Ising models on arbitrary graphs". In: *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. 2015, pp. 771–782.

[CH71] P Clifford and JM Hammersley. "Markov fields on finite graphs and lattices". In: (1971).

[CW20] Jordan Cotler and Frank Wilczek. "Quantum overlapping tomography". In: *Physical review letters* 124.10 (2020), p. 100401.

[Has07] Matthew B Hastings. "Quantum belief propagation: An algorithm for thermal quantum systems". In: *Physical Review B—Condensed Matter and Materials Physics* 76.20 (2007), p. 201102.

[Has10] Matthew B Hastings. "Locality in quantum systems". In: *Quantum Theory from Small to Large Scales* 95 (2010), pp. 171–212.

[Has23] Matthew B Hastings. "Field Theory and The Sum-of-Squares for Quantum Systems". In: *arXiv preprint arXiv:2302.14006* (2023).

[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 (2020), pp. 1050–1057.

[HKT22] Jeongwan Haah, Robin Kothari, and Ewin Tang. "Optimal learning of quantum Hamiltonians from high-temperature Gibbs states". In: *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2022, pp. 135–146.

[HO22] Matthew B Hastings and Ryan O'Donnell. "Optimizing strongly interacting fermionic Hamiltonians". In: *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*. 2022, pp. 776–789.

[Jay57] Edwin T Jaynes. "Information theory and statistical mechanics". In: *Physical review* 106.4 (1957), p. 620.

[Jay82] Edwin T Jaynes. “On the rationale of maximum-entropy methods”. In: *Proceedings of the IEEE* 70.9 (1982), pp. 939–952.

[KKB20] Tomotaka Kuwahara, Kohtaro Kato, and Fernando GSL Brandão. “Clustering of conditional mutual information for quantum Gibbs states above a threshold temperature”. In: *Physical review letters* 124.22 (2020), p. 220601.

[KM17] Adam Klivans and Raghu Meka. “Learning graphical models using multiplicative weights”. In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2017, pp. 343–354.

[KS20] Tomotaka Kuwahara and Keiji Saito. “Gaussian concentration bound and ensemble equivalence in generic quantum many-body systems including long-range interactions”. In: *Annals of Physics* 421 (2020), p. 168278.

[LR72] Elliott H Lieb and Derek W Robinson. “The finite group velocity of quantum spin systems”. In: *Communications in mathematical physics* 28.3 (1972), pp. 251–257.

[Mar21] Dániel Marx. *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 2021.

[MH24] Fermi Ma and Hsin-Yuan Huang. “How to construct random unitaries”. In: *arXiv preprint arXiv:2410.10116* (2024).

[Mon15] Andrea Montanari. “Computational implications of reducing data to sufficient statistics”. In: (2015).

[NC10] Michael A Nielsen and Isaac L Chuang. *Quantum computation and quantum information*. Cambridge university press, 2010.

[Nev24] Hartmut Neven. *Meet Willow, our state-of-the-art quantum chip*. `https://blog.google/technology/research/google-willow-quantum-chip`. Dec. 2024.

[VMLC16] Marc Vuffray et al. “Interaction screening: Efficient and sample-optimal learning of Ising models”. In: *Advances in neural information processing systems* 29 (2016).