Learning quantum many-body systems from a few copies

Cambyse Rouzé, Daniel Stilck França

Learning quantum many-body systems from a few copies

Cambyse Rouzé[1] and Daniel Stilck França[2,3]

[1]Department of Mathematics, Technische Universität München, 85748 Garching, Germany
[2]QMATH, Department of Mathematical Sciences, University of Copenhagen, Denmark
[3]Univ Lyon, ENS Lyon, UCBL, CNRS, Inria, LIP, F-69342, Lyon Cedex 07, France
11.03.2024

Estimating physical properties of quantum states from measurements is one of the most fundamental tasks in quantum science. In this work, we identify conditions on states under which it is possible to infer the expectation values of all quasi-local observables of a state from a number of copies that scale polylogarithmically with the system's size and polynomially on the locality of the target observables. We show that this constitutes a provable exponential improvement in the number of copies over state-of-the-art tomography protocols. We achieve our results by combining the maximum entropy method with tools from the emerging fields of classical shadows and quantum optimal transport. The latter allows us to fine-tune the error made in estimating the expectation value of an observable in terms of how local it is and how well we approximate the expectation value of a fixed set of few-body observables. We conjecture that our condition holds for all states exhibiting some form of decay of correlations and establish it for several subsets thereof. These include widely studied classes of states such as one-dimensional thermal and high-temperature Gibbs states of local commuting Hamiltonians on arbitrary hypergraphs or outputs of shallow circuits. Moreover, we show improvements of the maximum entropy method beyond the sample complexity that are of independent interest. These include identifying regimes in which it is possible to perform the postprocessing efficiently as well as novel bounds on the condition number of covariance matrices of many-body states.

1 Introduction

The subject of quantum tomography has as its goal devising methods for efficiently obtaining a classical description of a quantum system from access to experimental data. However, all tomographic methods for general quantum states inevitably require resources that scale exponentially in the size of the system [33, 55], be it in terms of the number of samples required or the post-processing needed to perform the task.

Cambyse Rouzé: rouzecambyse@gmail.com
Daniel Stilck França: daniel.stilck_franca@ens-lyon.fr

Fortunately, most of the physically relevant quantum systems can be described in terms of a (quasi)-local structure. These range from that of a local interaction Hamiltonian corresponding to a finite temperature Gibbs state to that of a shallow quantum circuit. Hence, locality is a physically motivated requirement that brings the number of parameters describing the system to a tractable number. Effective tomographic procedures should be able to incorporate this information. And, indeed, starting from physically motivated assumptions, many protocols in the literature achieve a good recovery guarantee in trace distance from a number of copies that scales *polynomially* with system size [5, 10, 23, 29, 64, 67].

Furthermore, in many cases, one is interested in learning only physical properties of the state on which tomography is being performed. These are mostly encoded into the expectation values of quasi-local observables that often only depend on reduced density matrices of subregions of the system. By Helstrom's theorem, obtaining a good recovery guarantee in trace distance is equivalent to demanding that the expectation value of *all bounded observables* are close for the two states, a much larger class of observables than quasi-local ones.

It is, in turn, desirable to design tomographic procedures that can take advantage of the fact that we wish to only approximate quasi-local observables, instead of demanding a recovery in trace distance. And some methods in the literature take advantage of that. For instance, the overlapping tomography or classical shadows methods of [22, 24, 40, 43] allow for approximately learning all k-local reduced density matrices of an n-qubit state with failure probability δ using $\mathcal{O}(e^{ck}k\log(n\delta^{-1})\epsilon^{-2})$ copies without imposing any assumptions on the underlying state. This constitutes an exponential improvement in the system size compared to the previously mentioned many-body setting at the expense of an undesirable exponential dependency in the locality of the observables.

In light of the previous discussion, it is natural to ask the guiding question of our work: *is it possible to devise a tomography protocol that has a sample complexity that is logarithmic in system size and polyno-*

mial in the locality of the observables we wish to estimate?

At first, this might sound like a tall order: as we show in Section G by importing results of [27], even if we start from the assumption that the underlying state we wish to learn is a high-temperature product state with n qubits, the number of samples required to obtain an estimate that is ϵ close in trace distance from the target state scales like $\Omega(n\epsilon^{-2})$. Thus, to obtain a sample complexity that is logarithmic in system size we cannot quantify closeness in trace distance and need to resort to more physically motivated distinguishability measures. Moreover, we show in Section G that even for product states the classical shadows protocol will fail to produce a good estimate for k-local observables if the number of samples is not exponential in k. We conclude that protocols like shadow tomography on their own cannot achieve our goal of a sample complexity that is polynomial in the locality of the underlying observables and needs to be combined with other estimation methods in a nontrivial way.

Despite these challenges, we provide an affirmative answer for the guiding question above for a large class of physically motivated states. We achieve this by combining two insights. First, we observe that recently introduced *Wasserstein distances* [21, 26, 31, 45, 56, 59] are better suited than the trace distance to estimate by how much the expectation values of physically motivated observables can differ on two states. We introduce these distances and motivate this claim below. But in summary, the Wasserstein distance quantifies how well we can distinguish states through observables whose expectation value does not change much when we apply a unitary acting only on a few qubits. By focusing on the Wasserstein distance instead of the trace distance we can bypass the $\Omega(n\epsilon^{-2})$ lower bound we mentioned previously. Intuitively, this means that exponentially fewer samples are required to estimate all such local expectation values than arbitrary, global ones.

The second insight is to combine techniques from quantum optimal transport with the well-established maximum entropy method [42] and the classical shadows protocols in a novel way. In particular, we will demonstrate that so-called transportation cost inequalities [21, 26, 31, 45, 56, 59] allow us to control how well we approximate the expectation value of k-local observables by how well we approximate certain observables that only act on a constant number of qubits. Thus, we only use the shadows protocol to estimate the expectation of many observables that are highly local, the regime in which classical shadows excel, and bypass the exponential scaling of only using shadows for such an estimation task. This way we obtain a provable exponential improvement over known methods of many-body tomography [5, 10, 23, 29, 64, 67] that focus on the trace distance and recent shadow tomography or overlapping tomography techniques [22, 24, 40, 43], as summarized in Table 1.

Examples for which we obtain exponential improvements include thermal states of 1D systems and high-temperature thermal states of commuting Hamiltonians on arbitrary hypergraphs and outputs of shallow circuits. Furthermore, based on results by [34, 49], we conjecture that our results should hold for any high-temperature Gibbs state, even. More ambitiously, we conjecture that our results can be extended to states exhibiting exponential decay of correlations. This would allow us to extend our findings to classes of states that are not known to be tractable classically, such as ground states of gapped Hamiltonians in higher dimensional lattices [39].

The main ingredient to obtain our improvements are so-called transportation cost (TC) inequalities [61]. They allow us to bound the difference of expectation values of Lipschitz observables, a concept we will review shortly, on two states by their relative entropy. Such inequalities constitute a powerful tool from the theory of optimal transport [65] and are traditionally used to prove sharp concentration inequalities [58, Chapter 3]. Moreover, they have been recently extended to quantum states [26, 56, 59]. By combining such inequalities with the maximum entropy principle, we are able to easily control the relative entropy between the states and, thus, the difference of expectation values of Lipschitz observables.

Our revisit of the maximum entropy principle is further motivated by recent breakthroughs in Hamiltonian learning [5, 34], shadow tomography [40], the understanding of correlations and computational complexity of quantum Gibbs states [35, 46, 49, 51, 52] and quantum functional inequalities [18, 26] that shed new light on this seasoned technique.

Before we summarize our contributions in more detail, we first define and revise the main concepts required for our results, namely Lipschitz observables, transportation cost inequalities and the maximum entropy principle.

1.1 Lipschitz observables

In the classical setting, given a metric d on a sample space $\mathcal{S}$, the regularity of a function $f : \mathcal{S} \to \mathbb{R}$ can be quantified by its *Lipschitz constant* [58, Chapter 3]

$$\|f\|_{\mathrm{Lip}} = \sup_{x,y\in\mathcal{S}} |f(x) - f(y)|/d(x,y)\,. \tag{1}$$

For instance, if we consider functions on the n-dimensional hypercube $\{-1,1\}^n$ endowed with the Hamming distance, the Lipschitz constant quantifies by how much a function can change per flipped spin. It should then be clear that physical quantities like average magnetization have a small Lipschitz constant. Some recent works [56, 59] extended this notion to the noncommutative setting and we will focus on the

Structure	Assumptions on State	Assumptions on Observable	Samples
Many-body tomography	Many-body	none	$\mathrm{poly}(n, \epsilon^{-1})$ [5]
Classical Shadows	none	k-local	$\mathrm{poly}(e^{ck}, \log(n), \epsilon^{-1})$ [40]
This work	Many-body+transportation	k-local	$\mathrm{poly}(k, \log(n), \epsilon^{-1})$

Table 1: Summary of underlying assumptions and sample complexity of other approaches to perform tomography on quantum many-body states.

approach of [56] in the main text. This is justified by the fact that it is more intuitive and technically simpler. For the approach followed in [56], the Lipschitz constant of an observable on n qudits is defined as [1]

$$\frac{\|O\|_{\mathrm{Lip},\square}}{\sqrt{n}} := \max_{1\leq i\leq n} \max_{\substack{\rho,\sigma\in\mathcal{D}_{d^n} \\ \mathrm{tr}_i[\rho]=\mathrm{tr}_i[\sigma]}} \mathrm{tr}\left[O(\rho-\sigma)\right], \quad (2)$$

where $\mathcal{D}_{d^n}$ denotes the set of n-qudit states. That is, $\|O\|_{\mathrm{Lip},\square}$ quantifies the amount by which the expectation value of O changes for states that are equal when tracing out one site. It is clear that $\|O\|_{\mathrm{Lip}} \leq 2\sqrt{n}\|O\|_\infty$ always holds by Hölder's inequality, but it can be the case that $\|O\|_{\mathrm{Lip},\square} \ll \sqrt{n}\|O\|_\infty$. For instance, consider for some $k > 0$ the n-qubit observable

$$O = n^{-1}\sum_{i=1}^{n} \otimes_{j=i}^{i+k} Z_j, \quad (3)$$

where for each site j, Z_j denotes the Pauli observable Z acting on site j and we take addition modulo n. It is not difficult to see that $\|O\|_{\mathrm{Lip},\square} = 2kn^{-\frac{1}{2}}$, while $\|O\|_\infty = 1$. We refer to the discussion in Fig. 1 for another example.

Moreover, one can show that shallow local circuits or short-time local dynamics satisfying a Lieb-Robinson bound cannot substantially increase the Lipschitz constant of an observable when evolved in the Heisenberg picture. That is, if we have that Φ_t^* is the quantum channel that describes some local dynamics at time/depth t in the Heisenberg picture and it satisfies a Lieb-Robinson bound, then we have:

$$\|\Phi_t^*(O)\|_{\mathrm{Lip},\square} = \mathcal{O}(e^{vt}\|O\|_{\mathrm{Lip},\square}),$$

where v denotes the Lieb-Robinson velocity. This result is discussed in more detail in Section B.1 of the supplemental material. Thus, averages over local observables and short-time evolutions thereof all belong to the class of observables that have a small Lipschitz constant when compared to generic observables. These facts justify our claim that quasi-local observables are Lipschitz.

Once we are given a Lipschitz constant on observables or a set of quasi-local observables, we can define a Wasserstein-1 distance on states by duality [56, 59].

[1] the definition of [56] has a different normalization and does not have the $\sqrt{n}$ term. This normalization will be convenient to treat the constants of [59] and [56] on an equal footing.

The latter quantifies how well we can distinguish two states by their action on regular or local observables and is given by

$$W_1(\rho,\sigma) := \sup_{O:\|O\|_{\mathrm{Lip},\square}\leq 1} \mathrm{tr}\left[O(\rho-\sigma)\right]. \quad (4)$$

The definition (4) is in direct analogy with the variational definition of the trace distance, given by:

$$\|\rho-\sigma\|_{\mathrm{tr}} := \sup_{O:\|O\|_\infty\leq 1} \mathrm{tr}\left[O(\rho-\sigma)\right]. \quad (5)$$

Note, however, that the two quantities have different scalings. To illustrate this point, let us consider the observable in Eq. (3). If we measure the distance in trace distance, then we need that $\|\rho-\sigma\|_{\mathrm{tr}} \leq \epsilon$ to ensure that σ approximates the expectation of value of ρ up to ϵ on O. On the other hand, as $\|O\|_{\mathrm{Lip},\square} \leq 2kn^{-\frac{1}{2}}$, the bound $W_1(\rho,\sigma) \leq \frac{\epsilon\sqrt{n}}{2k}$ is sufficient to guarantee the same approximation. This difference in scaling is at the heart of our results, as we will see now.

1.2 Transportation cost inequalities

The paragraphs before motivated the idea that observables with a small Lipschitz constant capture quasilocal observables, and thus, that controlling the Wasserstein distance between two states gives rise to a more physically motivated distance measure than the trace distance. However, it is a priori not clear how to effectively control the Wasserstein distance between states, as it does not admit a closed formula in terms of eigenvalues like the trace distance.

In this work, we will achieve this by relating Wasserstein distances to the relative entropy between two states, $D(\rho\|\sigma) := \mathrm{tr}\left[\rho\left(\log(\rho)-\log(\sigma)\right)\right]$, for σ of full-rank. This can be achieved through the notion of a *transportation cost inequality*: an n-qudit state σ is said to satisfy a transportation cost inequality with parameter $\alpha > 0$ if the Wasserstein distance of σ to any other state ρ can be controlled by their relative entropy, i.e.

$$W_1(\rho,\sigma) \leq \sqrt{\frac{D(\rho\|\sigma)}{2\alpha}} \quad (6)$$

holds for all states $\rho \in \mathcal{D}_{d^n}$.

Such inequalities are particularly powerful whenever the constant α does not depend on the system size n or does so at most inverse polylogarithmically,

and can be thought of as a strengthening of Pinsker's inequality.

Transportation cost inequalities are closely related to the notion of Gaussian concentration [12, 56, 59], i.e. that Lipschitz functions strongly concentrate around their mean. Establishing analogs of such concentration inequalities for quantum many-body systems has been a fruitful line of research in the last years and they are related to fundamental questions in statistical physics, see e.g. [3, 15, 49, 50, 62]. Although we are certain that inequalities like Eq. (6) also shed new light on this matter, here we will focus on their application to learning a classical description of a state through maximum entropy methods. We refer to Table 2 for a summary of classes of states known to satisfy it, as discussed in more detail below. Unfortunately, for some important classes, the inequalities are only known for the more technically involved variations of the Wasserstein distance, and we refer the reader to the supplemental material, Section B.2 for precise definitions.

Recent works have established transportation cost inequalities with α either constant or logarithmic in system size for several classes of Gibbs states of commuting Hamiltonians [18, 26, 56]. In summary, they are known to hold for local Hamiltonians on arbitrary hypergraphs at high enough temperatures or in 1D. In this work we enlarge the class of examples by showing them for outputs of short-depth circuits in Sec. C.1. Note that Eq. (6) is trivial for pure states, as then the relative entropy between that state and any other is always $+\infty$. Thus, we first find an appropriate full-rank approximation of the pure state for which the inequality holds, as we will discuss below.

1.3 Maximum-entropy methods

Let us now show how transportation cost inequalities can be combined with maximum entropy methods. Such methods start from the assumption that we are given a set of self-adjoint, linearly independent Hermitian observables over an n-qudit system, $E_1, \ldots, E_m \in \mathcal{M}^{\mathrm{sa}}_{d^n}$, with $\|E_i\|_\infty \leq 1$, a maximal inverse temperature $\beta > 0$ and the promise that the state we wish to learn can be expressed as:

$$\sigma \equiv \sigma(\lambda) = \frac{\exp\left(-\beta \sum_{i=1}^{m} \lambda_i E_i\right)}{\mathcal{Z}(\lambda)} \quad , \tag{7}$$

where $\lambda \in \mathbb{R}^m$ with sup norm $\|\lambda\|_{\ell_\infty} \leq 1$ and

$$\mathcal{Z}(\lambda) = \mathrm{tr}\left[\exp\left(-\beta \sum_{i=1}^{m} \lambda_i E_i\right)\right]$$

the partition function. Denoting by $e(\lambda)$ the vector with components $e_i(\lambda) = \mathrm{tr}\left[E_i\, \sigma(\lambda)\right]$, the crux of the maximum entropy method is that λ is the unique optimizer of

$$\min_{\substack{\mu \in \mathbb{R}^m \\ \|\mu\|_{\ell_\infty} \leq 1}} \log(\mathcal{Z}(\mu)) + \beta \sum_{i=1}^{m} \mu_i e_i(\lambda)\,, \tag{8}$$

which gives us a convex variational principle to learn the state given $e(\lambda)$. We refer to Sec. A for a discussion of the maximum entropy principle and its properties.

Typical examples of observables E_i are e.g. all 2−local Pauli observables corresponding to edges of a given graph. This models the situation in which we are guaranteed that the state is a thermal state of a Hamiltonian with a known locality structure. More generally, for most of the examples discussed here the E_i are given as follows: we start from a hypergraph $G = (V, E)$ and assume that there is a maximum radius $r_0 \in \mathbb{N}$ such that, for any hyperedge $A \in E$, there exists a vertex $v \in V$ such that the ball $B(v, r_0)$ centered at v and of radius r_0 includes A. The E_i are then given by a basis of the traceless matrices on each hyperedge A. This definition captures the notion of a local Hamiltonian w.r.t. the hypergraph.

Our framework also encompasses pure states after making an appropriate approximation. For instance, in this article we will also consider the outputs of shallow quantum circuits and believe our framework extends to unique ground states of gapped Hamiltonians, which are pure. Indeed, although it might not be a-priori clear, we will show below how the outputs of constant depth constant circuits are contained in the class of ground states of gapped, commuting Hamiltonians. And these are well-approximated by Gibbs states at finite temperature. Although this connection between Gibbs states of commuting Hamiltonians and outputs of shallow circuits might not be obvious at first, it makes some intuitive sense that the outputs of a shallow circuit are characterized by the reduced density matrix of the lightcone of each qubit. And the same holds for Gibbs states: one of their defining properties is that they are uniquely determined by the marginals. And this is the property that links these two classes of states.

Let us specify further what it means to learn the output of a shallow circuit. Suppose that $|\psi\rangle = \mathcal{U}\,|0\rangle^{\otimes n}$ is the output of an unknown shallow circuit $\mathcal{U}$ of depth L with respect to a known interaction graph (V, E) on n vertices. That is,

$$\mathcal{U} = \prod_{\ell \in [L]} \bigotimes_{e \in \mathcal{E}_\ell} \mathcal{U}_{\ell,e}\,,$$

where each $\mathcal{E}_\ell \subset E$ is a subset of non-intersecting edges. Thus, in this setting the locality of the circuit is known, but the underlying unitary is not. As we then show in Theorem C.1, the state $|\psi\rangle$ is ϵ^2 close in Wasserstein distance to the Gibbs state σ corresponding to the Hamiltonian with local terms $\mathcal{U} Z_i \mathcal{U}^\dagger$ at the inverse temperature $\beta = \log(\epsilon^{-1})$. By a simple light-cone argument we can bound the support of

each $\mathcal{U}Z_i\mathcal{U}^\dagger$, since we know the underlying structure of the circuit. We then show in Thm. C.1 that it is indeed possible to efficiently learn the outputs of such circuits as long as the support of each time-evolved Z_i is at most logarithmic in system size.

We see from Eq. (8) that the expectation values of the E_i completely characterize the state $\sigma(\lambda)$. But it is possible to obtain a more quantitative version of this statement through the following identity, also observed in [5]:

$$\begin{aligned} D(\sigma(\mu)\|\sigma(\lambda)) + D(\sigma(\lambda)\|\sigma(\mu)) \\ = -\beta \langle \lambda - \mu | e(\lambda) - e(\mu) \rangle. \end{aligned} \tag{9}$$

In addition to showing that if $e(\lambda) = e(\mu)$ then $\sigma(\mu) = \sigma(\lambda)$, Eq. (9) implies that by controlling how well the local expectations values of one state approximate another, we can also easily control their relative entropies. In particular, if $m = \mathcal{O}(n)$ and $\|e(\mu) - e(\lambda)\|_{\ell_1} = \mathcal{O}(\epsilon n)$ for some $\epsilon > 0$, we obtain from an application of Hölder's inequality that:

$$\begin{aligned} D(\sigma(\mu)\|\sigma(\lambda)) &\leq D(\sigma(\mu)\|\sigma(\lambda)) + D(\sigma(\lambda)\|\sigma(\mu)) \\ &= \mathcal{O}(\beta\epsilon n)\,. \end{aligned} \tag{10}$$

We refer to Section A for more details. Thus, if we can find a state that approximates the expectation value of each E_i up to ϵ, we are guaranteed to have a $\mathcal{O}(\beta\epsilon)$ relative entropy density. This observation is vital to ensure that the maximum entropy principle still yields a good estimate of the state even under some statistical noise in the vector of expectation values $e(\lambda)$. Indeed, the variational principle of Eq. (8) would allow us to recover the state exactly if we had access to the exact values of $e(\lambda)$. However, it turns out that solving Eq. (8) with some estimate $\hat{e}(\lambda)$ such that $\|\hat{e}(\lambda) - e(\lambda)\|_{\ell_\infty} \leq \epsilon$ still yields a Gibbs state $\sigma(\mu)$ satisfying Eq. (10). The maximum entropy problem is a strictly convex optimization problem. Thus, it can be solved efficiently with access to the gradient of the target function. The gradient turns out to be proportional to $e(\lambda) - e(\mu)$, where μ is the current guess for the optimum. Although we will discuss the details of solving the problem later in Sec. A, in a nutshell, the maximum entropy problem can be solved efficiently if it is possible to efficiently compute expectation values of the observables E_i on the family of Gibbs states under consideration.

1.4 Combining TC with the maximum entropy principle

Suppose now that we have that each of the E_i acts on at most l_0 qudits. Then, by using e.g. the method of classical shadows, we can estimate the expectation values of all E_i up to ϵ with failure probability at most $\delta > 0$ with $\mathcal{O}(4^{l_0}\epsilon^{-2}\log(m\delta^{-1}))$ samples. From our discussion above, we see that this is enough to obtain a state $\sigma(\mu)$ satisfying Eq. (10). Further assuming that we have a TC with some constant $\alpha > 0$ for $\sigma(\lambda)$ we conclude that:

$$\begin{aligned} |\mathrm{tr}\left[O(\sigma(\lambda) - \sigma(\mu))\right]| &\leq \|O\|_{\mathrm{Lip}} W_1(\sigma(\lambda), \sigma(\mu)) \\ &\leq \|O\|_{\mathrm{Lip}} \sqrt{\frac{D(\sigma(\mu)\|\sigma(\lambda))}{2\alpha}} \\ &= \mathcal{O}(\sqrt{\beta\epsilon n}\,\|O\|_{\mathrm{Lip}}). \end{aligned}$$

Finally, recall that for sums of k local operators on a 2D lattice like in Fig. 1, where we have $k = L^2$, the Lipschitz constant satisfies $\|O\|_{\mathrm{Lip}} = \mathcal{O}(\sqrt{n})$ and we require a precision of $\mathcal{O}(\epsilon n/k)$ to obtain a relative error of ϵ. Putting all these elements together, we conclude that by setting $\epsilon = \beta\,\tilde{\epsilon}^2/(k^2)$ for some $\tilde{\epsilon} > 0$ we arrive at

$$|\mathrm{tr}\left[O(\sigma(\lambda) - \sigma(\mu))\right]| = \mathcal{O}(\tilde{\epsilon}\,k^{-1}n), \tag{11}$$

which constitutes a relative error for the expectation value. In particular, we see that the sample complexity required to obtain this error was

$$\mathcal{O}(4^{l_0}k^4\beta^2\,\tilde{\epsilon}^{-4}\log(m)). \tag{12}$$

We then obtain:

Theorem 1.1 (Learning Gibbs states)**.** *Let $\sigma(\lambda)$ be a Gibbs state as defined in Eq. (7) and such that each E_i acts on at most l_0 qudits. Moreover, suppose that $\sigma(\lambda)$ satisfies a* TC *inequality with α depending at most inverse logarithmically with system size. Then with probability of success $1-\delta$ we can obtain a state $\sigma(\mu)$ such that for all observables $O \in \mathcal{M}_{d^n}$*

$$|\mathrm{tr}\left[O(\sigma(\lambda) - \sigma(\mu))\right]| \leq \mathcal{O}(\epsilon\sqrt{n}\,\|O\|_{\mathrm{Lip}}) \tag{13}$$

from $\mathcal{O}(4^{l_0}\beta^2\ \mathrm{poly}(\epsilon^{-1}, \log(m\delta^{-1})))$ samples of $\sigma(\lambda)$. Moreover, if it is possible to compute the expectation values of the E_i on $\sigma(\tau)$ for $\|\tau\|_{\ell_\infty} \leq 1$, then the postprocessing can also be in polynomial time.

We once again stress that the recovery guarantee in Eq. (13) suffices to give good relative approximations for the expectation value of quasilocal observables. Furthermore, if we did not resort to the Wasserstein distance but rather the trace distance, as in known results for the tomography of many-body states [5, 10, 23, 53, 67], the sample complexity would be exponentially worse, as we prove in Sec. G. More precisely, any algorithm that estimates Gibbs states on a lattice at inverse temperatures $\beta = \Omega(n^{-\frac{1}{2}})$ up to trace distance ϵ requires $\Omega(n\epsilon^{-2})$ samples. Thus, even for states whose inverse temperature goes to 0 as the system size increases, a focus on the trace distance instead of the Wasserstein distance implies an exponentially worse sample complexity.

Theorem 1.1 also provides an exponential improvement over shadow techniques in the locality of the observables, as we argue in Sec. G. However, unlike our methods, shadow techniques do not need to make any

Structure	Samples	Postprocessing
Lightcone l_0 circuit	$\epsilon^{-4}2^{12l_0}$	$\exp(3n)$
Commuting Gibbs for $\beta < \beta_c$	ϵ^{-2}	$\mathrm{poly}(n)$
1D commuting	$\min\{\epsilon^{-2}e^{\mathcal{O}(\beta)}, \epsilon^{-4}\}$	$\mathrm{poly}(n)$

Table 2: Performance of the algorithm under various assumptions of the underlying state to obtain a state that is $\epsilon\sqrt{n}$ close in Wasserstein distance. All the estimates are up to $\mathrm{polylog}(n)$ factors. We refer to Sec. B.3 for proofs of the TC used and C for how to combine them with maximum entropy methods. In Section D we explain how to obtain the sample complexity by combining Thm. A.1 with strong convexity bounds. For the postprocessing we refer to section E. The case of shallow circuits is discussed in more detail in Sec. C.1. By lightcone l_0 we mean the size of the largest lightcone of each qubit in the circuit.

assumptions on the underlying states. Thus, we see that Theorem 1.1 opens up the possibility of highly efficient characterization of quantum states and provably exponentially better sample complexities when compared with recovery in trace distance.

We also remark that it is possible to improve the scaling in accuracy in Eq. (12) from ϵ^{-4} to the expected ϵ^{-2}. To do that, it is important to bound the condition number of the Hessian of the log-partition function, as we explain in the methods.

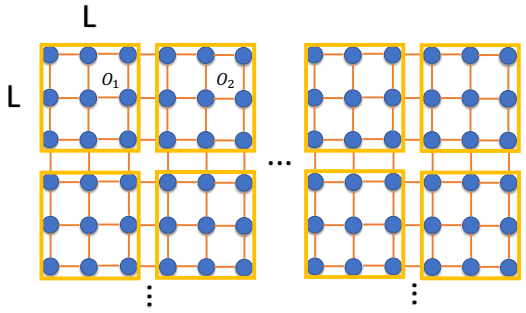

Figure 1: example of observable $O = \sum_i O_i$ for $2D$ lattice system of size n. Each O_i is supported on a $L \times L$ square ($L = 3$ in the figure). We have $\|O\|_{\mathrm{Lip}} = \mathcal{O}(\sqrt{n})$ and $\|O\| = n/L^2$. Thus our methods require $\mathrm{poly}(L, \log(n), \epsilon^{-1})$ samples to estimate the expectation value of all such observables. Shadow-like methods require $\mathrm{poly}(e^{cL^2}, \log(n), \epsilon^{-1})$ samples, an exponentially worse dependency in L. Even for moderate values of L, say $L = 5$, this can lead to 10^7 factor savings sample complexity and gives an exponential speedup for $L = \mathrm{poly}(\log(n))$. Other many-body methods have a $\mathrm{poly}(L, n, \epsilon^{-1})$ [5, 10, 23, 53, 67] scaling, which in turn is exponentially worse in the system size.

1.5 Numerical results

We will now compare the performance of our method to the classical shadow protocol [40] to estimate the average of a local observable on a Gibbs state. To ensure that we can still generate samples for a high number of qubits, we will consider the following family of commuting Gibbs states in 1-D:

$$H(\lambda) = \qquad (14)$$
$$- \sum_{k=0}^{n/2-1} S^{2k}(\lambda_k X_0 X_1 + \lambda_{n+k} X_0 X_1 Y_2 Y_3) S^{2k},$$

where S is the shift operator, $\lambda \in B_{\ell_\infty}(0, 1)$ and we assumed n is even. We will then estimate the expectation value of the observable:

$$O = \sum_{k=0}^{n/2-1} S^{2k} X_0 X_1 Y_2 Y_3 X_4 X_5 Y_6 Y_7 S^{2k}. \qquad (15)$$

The results for one particular choice of Gibbs state in this class are shown in Fig. 2. It shows that even for observables of moderate locality like the one in Eq. (15), shadows are outperformed by maximum entropy methods by orders of magnitude. Also note that the quality of our estimates decays like $\sim 1/\sqrt{s}$, where s is the number of samples, showing how the quality of the recovery is essentially independent of the system's size.

We also remark that for obtaining these results, we obtained the expectation values of the $X_i X_{i+1}$ terms by measuring in the X basis on each qubit and of the $XXYY$ terms by measuring in a sequence of $XXYY$ bases followed by the same basis shifted by 2.

2 Conclusion

In this article we have demonstrated that ideas from quantum optimal transport yield provable exponential improvements in the sample complexity required to learn a quantum state when compared to state-of-the-art methods. More precisely, we showcased how the interplay between maximum entropy methods and the Wasserstein distance, which is mediated by a transportation cost inequality, allows for fine-tuning the complexity of observables whose expectation value we wish to estimate and the number of samples required for that. Through our techniques, we essentially settled most questions related to how efficiently it is possible to learn a commuting Gibbs state and significantly advanced our understanding of general Gibbs states. With the impressive growth in

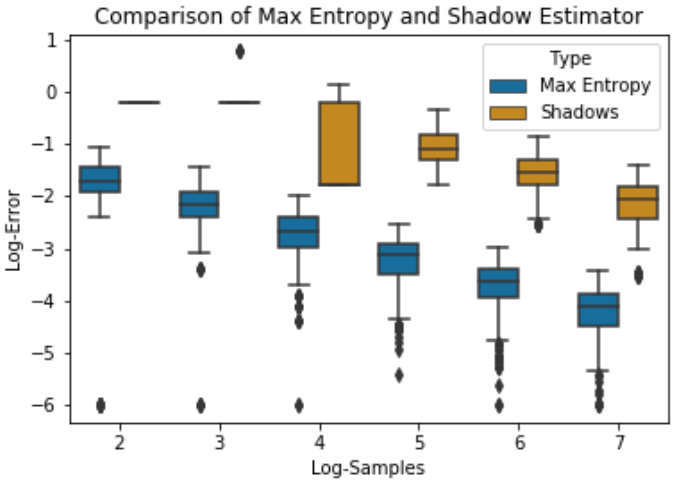

Figure 2: Performance on a Gibbs state from the family of Eq. (14) and the 8-local observable in Eq. (15). We have set the number of qubits to 100, $\beta = 1.1$ and the λ_i uniformly at random between 0.5 and 0.9. The x-axis denotes the logarithm of the number of samples in base 10 and the y the error in absolute value to the true value. We ran each protocol 300 times on the same Gibbs state to see how the estimate varied. We see that even with 10^4 samples the shadows method still has errors of the order 10^0 in the 75 percentile, whereas the maximum entropy already yields good estimates when the number of samples is of order 10^2, showcasing that maximum entropy methods outperform classical shadows by orders of magnitudes for observables of moderate locality like those in Eq. (15).

the size of quantum devices available in the lab over the last years, we believe that the polylogarithmic in system size sample complexities obtained here will come in handy to calibrate and characterize systems containing thousands or millions of qubits.

We believe that the framework and philosophy we began to develop here will find applications in other areas of quantum information theory. Indeed, although a bound in trace distance is the golden standard and one of the most widely accepted and used measures of distance between quantum states in quantum information and computation, we argued that in many physically relevant settings, demanding a trace distance bound might be too strong. More importantly, replacing the trace distance by a Wasserstein distance bound can lead to exponential complexity gains, as in this article. Thus, we believe that this approach is likely to lead to substantial gains and improvements also in other areas like quantum machine learning, process tomography or in quantum many-body problems.

Some of the outstanding open questions raised by this article are establishing that a suitable notion of exponential decay of correlations in general implies a transportation cost inequality and showing that TC holds for a larger class of systems. We believe that our framework should also extend to ground states of gapped Hamiltonians in 1D, however, such a statement would still require us to refine our bounds. The results presented here also make us conjecture that any high-temperature Gibbs state satisfies a TC inequality, even for long-range interactions, which would make our techniques applicable to essentially all physically relevant models at high temperatures.

That being said, to the best of our knowledge, all states that are known to satisfy a TCI can also be simulated efficiently classically. However, there is a-priori no reason to believe that a TCI necessarily implies that the underlying states can be simulated classically and we believe that this is more of a reflection of the fact that the study of TCI is still incipient. Furthermore, although we explained how to do the tomography of the Gbibs state with the maximum entropy principle, other Hamiltonian learning procedures also apply, as all we need is to upper-bound the relative entropy through Eq. (9). Thus, breakthroughs in Hamiltonian learning would then also apply to the setting in this paper. In particular, through efficient Hamiltonian learning methods, it might be possible to use the results derived here to verify if a quantum device correctly implemented a shallow circuit.

Moreover, it would also be interesting to investigate other applications of Gaussian concentration in many-body physics [3, 15, 49, 50, 62] from the angle of transportation cost inequalities.

3 Methods

3.1 Summary of the maximum entropy procedure and contributions

Now that we have discussed how our results yield better sample complexities for some classes of states, we discuss the maximum entropy algorithm in more detail and comment on how our results equip it with better performance guarantees.

As the maximum entropy principle in Eq. (20) corresponds to solving a convex optimization problem, it should come as no surprise that promises on the strong convexity of the underlying functions being optimized can be leveraged to give improved performance guarantees [14]. For the specific case of the maximum entropy principle, strong convexity guarantees translate to bounds of the form

$$LI \leq \nabla^2 \log(\mathcal{Z}(\mu)) \leq UI \tag{16}$$

for constants $L, U > 0$ and all $\mu \in B_{\ell_\infty}(0,1)$. We refer to Sec. A of the supplemental material for a thorough discussion. We note that in [5] the authors show such results in a more general setting, although with U, L polynomial in n, which is not sufficient for our purposes. For us it will be important to ensure that the condition number of the log-partition function is at most polylogarithmic in system size (i.e. $L^{-1}U = \tilde{\mathcal{O}}(1)$).

If we define the function $f : \mu \mapsto \log(\mathcal{Z}(\mu)) + \beta \sum_{i=1}^{m} \mu_i e_i(\lambda)$, then $\nabla f = \beta(e(\lambda) - e(\mu))$. It then follows from standard properties of strongly convex functions that:

$$\|\lambda - \mu\|_{\ell_2} = L^{-1}\beta \|e(\lambda) - e(\mu)\|_{\ell_2}$$

That is, whenever the expectation values are close, the underlying parameters must be close as well. In this case, we have from $\|e(\lambda) - e(\mu)\|_{\ell_2} = \mathcal{O}(\epsilon\sqrt{m})$ that:

$$\begin{aligned} &-\beta \langle \lambda - \mu | e(\lambda) - e(\mu) \rangle \\ &\leq \beta \|\lambda - \mu\|_{\ell_2} \|e(\lambda) - e(\mu)\|_{\ell_2} = \mathcal{O}(L^{-1}\beta^2\epsilon^2 m). \end{aligned} \tag{17}$$

As we will see in the proposition below, $L^{-1} = \mathcal{O}(e^\beta \beta^{-2})$ for commuting Hamiltonians, which gives a quadratic improvement in ϵ in Eq. (17) and yields the expected ϵ^{-2} scaling for the sample complexity in terms of the accuracy.

Proposition 3.1 (Strengthened strong convexity constant for commuting Hamiltonians)**.** *For each $\mu \in B_{\ell_\infty}(0,1)$, let $\sigma(\mu)$ be a Gibbs state at inverse temperature β corresponding to the commuting Hamiltonian $H(\mu) = \sum_j \mu_j E_j$ on the hypergraph $G = (V, E)$, where $\operatorname{tr}[E_i E_j] = 0$ for all i, j and each local operator E_j is traceless on its support. Then for β such that the states $\sigma(\mu)$ satisfy exponential decay of correlations, the Hessian of the log-partition function is bounded by*

$$\Omega(\beta^2 e^{-c\beta}) I \leq \nabla^2 \log \mathcal{Z}(\mu) \leq \mathcal{O}(\beta^2) I. \tag{18}$$

for some constant $c > 0$.

After the completion of the first version of this work, in [34, Corollary 4.4] Tang et al proved strong convexity bounds for high-temperature, (not necessarily geometrically) local Hamiltonians. More precisely, for $\beta = \mathcal{O}(k^{-8})$, where k is the maximal number of qudits each term acts on, they show that $L^{-1} \leq 2\beta^{-2}$. Although the result in Eq. (18) has the advantage of giving an estimate at any temperature, we see that also for noncommuting Gibbs states strong convexity holds at high enough temperatures.

The flowchart in Figure 3 gives the general scheme behind the maximum entropy method. Besides the exponential improvements in sample complexity laid out in Table 2, we also provide structural improvements which we elaborate on while also explaining the general scheme:

Input: The input consists of m linearly independent operators E_i of operator norm 1, some upper bound $\beta > 0$, precision parameter $\epsilon > 0$ and step size η. Moreover, we are given the promise that the state of interest satisfies (7). Although we will be mostly concerned with the case in which the observables are local, we show the convergence of the algorithm in general in Sec. A. The step size should be picked as $\eta = \mathcal{O}(U^{-1})$ with U satisfying (16), as explained in Sec. A.1.

Require: We assume that we have access to copies of $\sigma(\lambda)$ and that we can perform measurements to estimate the expectation values of the observables E_i up to precision $\epsilon > 0$. For most of the examples considered here, this will mostly require implementing simple, few-qudit measurements.

Output: The output is in the form of a vector of parameters μ of a Gibbs state $\sigma(\mu)$ as in Eq. (7). Note that unlike [5], our goal is not to estimate the vector of parameters λ, but rather obtain an approximation of the state satisfying $\sigma(\lambda) \simeq \sigma(\mu)$. Here we will focus on quantifying the output's quality in relative entropy. More precisely, the output of the algorithm is guaranteed to satisfy $D(\sigma(\mu)\|\sigma(\lambda)) = \mathcal{O}(\epsilon n)$.

Step 1: In this step, we estimate the expectation values of each observable E_i on the state $\sigma(\lambda)$ up to an error ϵ. The resources to be optimized here are the number of samples of $\sigma(\lambda)$ we require and the complexity to implement the measurements. Using shadow tomography or Pauli grouping methods [13, 22, 40] we can do so requiring $\mathcal{O}(4^{r_0}\epsilon^{-2}\operatorname{polylog}(m))$ samples and Pauli or $1-$qubit measurements, where r_0 is maximum number of qubits the E_i act on. This is discussed in more detail in Sec. C.

Step 2: The maximum entropy problem in Eq. (8) can be solved with gradient descent, as it corresponds to a strictly convex optimization problem [14]. At this step we simply initialize the algorithm to start at the maximally mixed state.

Step 3: It turns out that the gradient of the maximal entropy problem target function at μ_t is proportional to $e(\mu_t) - e(\lambda)$. Thus, to implement an iteration of gradient descent, it is necessary to compute $e(\mu_t)$, as we assumed we obtained an approximation of $e(\lambda)$ in Step 1. Moreover, it is imperative to ensure that the algorithm also converges with access to approximations to $e(\mu)$ and $e(\lambda)$. This is because most algorithms to compute $e(\mu_t)$ only provide approximate values [35, 37, 46, 51] . In addition, they usually have a poor scaling in the accuracy [51], making it necessary to show that the process converges with rough approximations to $e(\mu_t)$ and $e(\lambda)$. Here we show that it is indeed possible to perform this step with only approximate computations of expectation values. This allows us to identify classes of states for which the postprocessing can be done efficiently. These results are discussed in more detail in Sec. A.2.

Convergence loop: Now that we have seen how to compute one iteration of gradient descent, the next natural question is how many iterations are required to reach the stopping criterium. As this is a strongly convex problem, the convergence speed depends on the eigenvalues of the Hessian of the function being optimized [14, Section 9.1.2]. For max-entropy, this corresponds to bounding the eigenvalues of a generalized covariance matrix. In [5] the authors already showed such bounds for local Hamiltonians implying the convergence of the algorithm in a number of steps depending polynomially in m and logarithmically on the tolerance ϵ for a fixed β. Here we improve their bound in several directions. First, we show that the algorithm converges after a polynomial in m number of iterations for arbitrary E_i, albeit with a polynomial dependence on the error, as discussed in Sec. A.2. We then specialize to certain classes of states to obtain various improvements. For high-temperature, commuting Hamiltonians we provide a complete picture and show that the condition number of the Hessian is constant in Prop. 3.1. This implies that gradient descent converges in a number of iterations that scales logarithmically in system size and error.

Stopping condition and recovery guarantees: the stopping condition,

$$\|e(\lambda) - e(\mu_t)\|_{\ell_2} \leq \epsilon\sqrt{m}\,,$$

can be immediately converted to a relative entropy bound between the target state and the current iterate by the identity (9). This justifies its choice as a stopping criterion.

Since we already discussed the sample complexity of the maximum entropy method, let us now discuss some of its computational aspects. There are two quantities that govern the complexity of the algorithm: how many iterations we need to perform until we converge and how expensive each iteration is.

As the maximum entropy problem is strongly convex, one can show that $\mathcal{O}(UL^{-1}\log(n\epsilon^{-1}))$ iterations suffice to converge. Here again U, L are bounds on the Hessian as in Eq. (16). Nevertheless, we also show how to bypass requiring such bounds in Sec. A.1 and obtain that the maximum entropy algorithm converges after $\mathcal{O}(m\epsilon^{-2})$ iterations without any locality assumptions on the E_i or strong convexity guarantees. That is, the number of iterations is at most polynomial in m.

Let us now discuss the cost of implementing each iteration of the algorithm on a classical computer. This boils down to estimating $e(\mu_t)$ for the current iterate, which can be achieved in various ways. In the worst case, it is possible to just compute the matrix exponential and the expectation values directly, which yields a complexity of $\mathcal{O}(d^{3n}m)$. However, for many of the classes considered here it is possible to do this computation in polynomial time. For instance, in [51] the authors show that for high-temperature Gibbs states it is possible to approximate the partition function efficiently. Thus, for the case of high-temperature Gibbs states, not necessarily commuting ones, we can do the postprocessing efficiently. It is also worth mentioning tensor network techniques to estimate $e(\mu_t)$. As we only require computing the expectation value of local observables, recent works show that it is possible to find tensor network states of constant bond dimension that approximate all expectation values of a given locality well [2, 25, 41]. From such a representation it is then possible to compute $e(\mu_t)$ efficiently in the 1D case by contracting the underlying tensor network. Unfortunately, however, in higher dimensions the contraction still takes exponential time. Table 2 provides a summary of the complexity of the postprocessing for various classes.

It is also worth considering the complexity of the postprocessing with access to a quantum computer, especially for commuting Hamiltonians. As all high-temperature Gibbs states satisfy exponential decay of correlations, the results of [17] imply that high-temperature Gibbs states can be prepared with a circuit of depth logarithmic in system size. Thus, by using the same method we used to estimate $e(\lambda)$ we can also estimate $e(\mu_t)$ by using the copies provided by the quantum computer. The complexity of the postprocessing for shadows is linear in system size. Thus, with access to a quantum computer we can perform the post-processing for each iteration in time $\tilde{O}(m\epsilon^{-2})$. As in this case we showed that the number of iterations is $\tilde{\mathcal{O}}(1)$, we conclude that we can perform the postprocessing in time $\tilde{\mathcal{O}}(m\epsilon^{-2})$. That is, for this class of systems our results give an arguably complete picture regarding the postprocessing, as it can be performed in a time comparable to writing down the vector of parameters, up to polylogarithmic factors. Furthermore, given that the underlying Gibbs states are known to satisfy TC and Prop. 3.1 gives es-

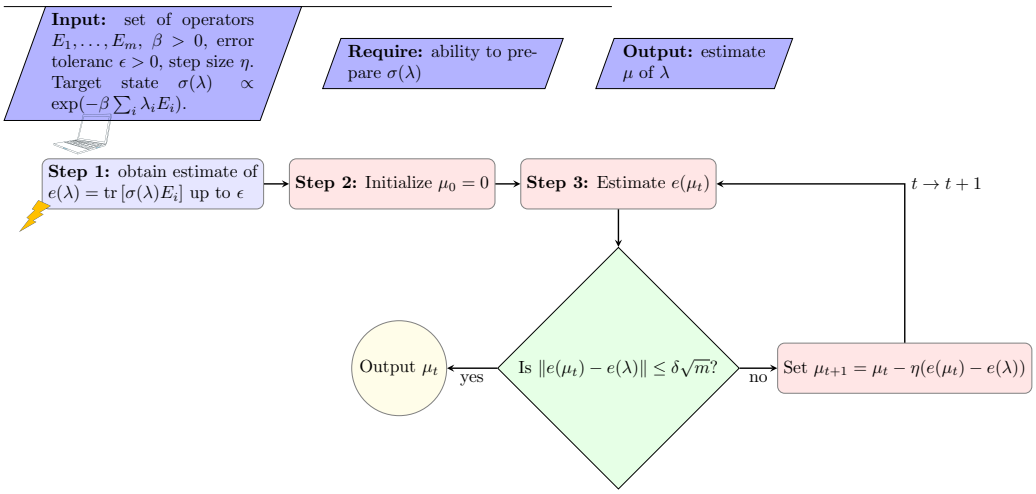

Figure 3: Flowchart for general maximum entropy algorithms.

sentially optimal bounds on the covariance matrices, we believe that the present work essentially settles the question of how efficiently we can learn such Hamiltonians and corresponding Gibbs states. We discuss this in more detail in Sec. F.

Finally, an example of our bounds is illustrated in Fig. 4, where we show that the number of samples required to estimate a local observable to relative precision is essentially system-size independent.

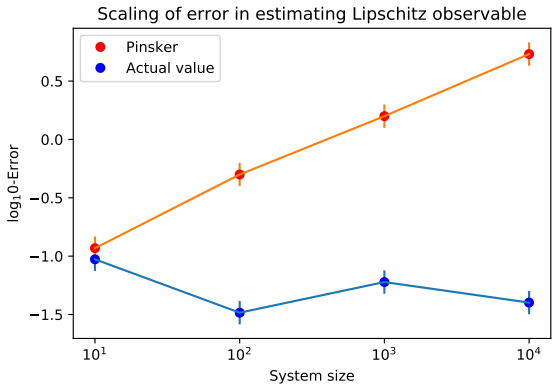

Figure 4: Error in estimating a Lipschitz observable after performing the maximum entropy reconstruction method. The underlying state is a classical 1D-Gibbs state with randomly chosen nearest-neighbor interactions and inverse temperature $\beta = 1$. We estimated all the Z_iZ_{i+1} expectation values from the original state based from 10^3 samples of the original state. We then computed the upper bound on the trace distance predicted by Eq. (9) and Pinsker's inequality and compared it to the actual of discrepancy for a Lipschitz observable on the reconstructed and actual state. The Lipschitz observable was chosen as $\sum_i n^{-1} U Z_i Z_{i+2} U^\dagger$, where we picked U as a depth 3 quantum circuit. We observe that the error incurred is essentially independent of system size, and we get good predictions even when the number of samples is smaller than it.

4 Acknowledgements

DSF was supported by the VILLUM FONDEN via the QMATH Centre of Excellence under Grant No. 10059 and from the European Research Council (grant agreement no. 81876). The research of CR has been supported by project QTraj (ANR-20-CE40-0024-01) of the French National Research Agency (ANR) and by a Junior Researcher START Fellowship from the MCQST. DSF and CR are grateful to Richard Kueng, Fernando Brandão and Giacomo De Palma for interesting discussions.

References

[1] Scott Aaronson. Shadow tomography of quantum states. In *STOC'18—Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 325–338. ACM, New York, 2018. DOI: 10.1145/3188745.3188802.

[2] Álvaro M. Alhambra and J. Ignacio Cirac. Locally accurate tensor networks for thermal states and time evolution. *PRX Quantum*, 2 (4), November 2021. ISSN 2691-3399. DOI: 10.1103/prxquantum.2.040331.

[3] Anurag Anshu. Concentration bounds for quantum states with finite correlation length on quantum spin lattice systems. *New Journal of Physics*, 18(8):083011, aug 2016. DOI: 10.1088/1367-2630/18/8/083011.

[4] Anurag Anshu, Srinivasan Arunachalam, Tomotaka Kuwahara, and Mehdi Soleimanifar. Efficient learning of commuting hamiltonians on lattices. Unpublished notes avaible at Anurag Anshu's website, (link to note).

[5] Anurag Anshu, Srinivasan Arunachalam, Tomotaka Kuwahara, and Mehdi Soleimanifar. Sample-efficient learning of interacting quantum

systems. *Nature Physics*, may 2021. ISSN 1745-2473. DOI: 10.1038/s41567-021-01232-0.

[6] Huzihiro Araki. Gibbs states of a one dimensional quantum lattice. *Communications in Mathematical Physics*, 14(2):120–157, 1969. DOI: 10.1007/bf01645134.

[7] Ivan Bardet, Angela Capel, and Cambyse Rouzé. Approximate tensorization of the relative entropy for noncommuting conditional expectations. *arXiv preprint arXiv:2001.07981*, 2020. DOI: 10.1007/s00023-021-01088-3.

[8] Ivan Bardet, Ángela Capel, and Cambyse Rouzé. Approximate tensorization of the relative entropy for noncommuting conditional expectations. *Annales Henri Poincaré*, 23(1): 101–140, July 2021. ISSN 1424-0661. DOI: 10.1007/s00023-021-01088-3.

[9] Thomas Barthel and Martin Kliesch. Quasilocality and efficient simulation of markovian quantum dynamics. *Physical review letters*, 108(23):230504, 2012. DOI: 10.1103/physrevlett.108.230504.

[10] T Baumgratz, A Nüßeler, M Cramer, and M B Plenio. A scalable maximum likelihood method for quantum state tomography. *New Journal of Physics*, 15(12):125004, dec 2013. DOI: 10.1088/1367-2630/15/12/125004.

[11] Salman Beigi, Nilanjana Datta, and Cambyse Rouzé. Quantum reverse hypercontractivity: its tensorization and application to strong converses. *Communications in Mathematical Physics*, 376 (2):753–794, 2020. DOI: 10.1007/s00220-020-03750-z.

[12] S.G Bobkov and F Götze. Exponential Integrability and Transportation Cost Related to Logarithmic Sobolev Inequalities. *Journal of Functional Analysis*, 163(1):1–28, April 1999. ISSN 00221236. DOI: 10.1006/jfan.1998.3326.

[13] Xavier Bonet-Monroig, Ryan Babbush, and Thomas E O'Brien. Nearly optimal measurement scheduling for partial tomography of quantum states. *Physical Review X*, 10(3):031064, 2020. DOI: 10.1103/physrevx.10.031064.

[14] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[15] Fernando G. S. L. Brandao and Marcus Cramer. Equivalence of Statistical Mechanical Ensembles for Non-Critical Quantum Systems. *arXiv:1502.03263 [cond-mat, physics:quant-ph]*, February 2015. arXiv: 1502.03263.

[16] Fernando G. S. L. Brandão, Amir Kalev, Tongyang Li, Cedric Yen-Yu Lin, Krysta M. Svore, and Xiaodi Wu. Quantum SDP solvers: large speed-ups, optimality, and applications to quantum learning. In *46th International Colloquium on Automata, Languages, and Programming*, volume 132 of *LIPIcs. Leibniz Int. Proc. Inform.*, pages Art. No. 27, 14. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2019.

[17] Fernando G. S. L. Brandão and Michael J. Kastoryano. Finite Correlation Length Implies Efficient Preparation of Quantum Thermal States. *Communications in Mathematical Physics*, 365 (1):1–16, January 2019. ISSN 0010-3616, 1432-0916. DOI: 10.1007/s00220-018-3150-8.

[18] Ángela Capel, Cambyse Rouzé, and Daniel Stilck França. The modified logarithmic sobolev inequality for quantum spin systems: classical and commuting nearest neighbour interactions. *arXiv preprint arXiv:2009.11817*, 2020.

[19] Eric A Carlen and Jan Maas. An analog of the 2-wasserstein metric in non-commutative probability under which the fermionic fokker–planck equation is gradient flow for the entropy. *Communications in mathematical physics*, 331(3): 887–926, 2014. DOI: 10.1007/s00220-014-2124-8.

[20] Eric A Carlen and Jan Maas. Gradient flow and entropy inequalities for quantum Markov semigroups with detailed balance. *Journal of Functional Analysis*, 273(5):1810–1869, 2017. DOI: 10.1016/j.jfa.2017.05.003.

[21] Eric A. Carlen and Jan Maas. Non-commutative calculus, optimal transport and functional inequalities in dissipative quantum systems. *Journal of Statistical Physics*, 178(2):319–378, nov 2019. DOI: 10.1007/s10955-019-02434-w.

[22] Jordan Cotler and Frank Wilczek. Quantum overlapping tomography. *Physical Review Letters*, 124(10):100401, 2020. DOI: 10.32657/10356/170876.

[23] 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. *Nature Communications*, 1(1):149, dec 2010. ISSN 2041-1723. DOI: 10.1038/ncomms1147.

[24] Ophelia Crawford, Barnaby van Straaten, Daochen Wang, Thomas Parks, Earl Campbell, and Stephen Brierley. Efficient quantum measurement of pauli operators in the presence of finite sampling error. *Quantum*, 5:385, January 2021. ISSN 2521-327X. DOI: 10.22331/q-2021-01-20-385. URL `http://dx.doi.org/10.22331/q-2021-01-20-385`.

[25] Alexander M. Dalzell and Fernando G. S. L. Brandão. Locally accurate MPS approximations for ground states of one-dimensional gapped local Hamiltonians. *Quantum*, 3:187, September 2019. ISSN 2521-327X. DOI: 10.22331/q-2019-09-23-187.

[26] Giacomo De Palma and Cambyse Rouzé. Quantum concentration inequalities. *Annales Henri Poincaré*, 23(9):3391–3429, April 2022.

ISSN 1424-0661. DOI: 10.1007/s00023-022-01181-1. URL http://dx.doi.org/10.1007/s00023-022-01181-1.

[27] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The minimax learning rates of normal and Ising undirected graphical models. *Electronic Journal of Statistics*, 14(1), January 2020. DOI: 10.1214/20-ejs1721.

[28] Roland L Dobrushin and Senya B Shlosman. Completely analytical interactions: constructive description. *Journal of Statistical Physics*, 46(5): 983–1014, 1987. DOI: 10.1007/bf01011153.

[29] Jens Eisert, Dominik Hangleiter, Nathan Walk, Ingo Roth, Damian Markham, Rhea Parekh, Ulysse Chabaud, and Elham Kashefi. Quantum certification and benchmarking. *Nature Reviews Physics*, 2(7):382–390, jul 2020. ISSN 2522-5820. DOI: 10.1038/s42254-020-0186-4.

[30] Daniel Stilck França, Fernando G.S L. Brandão, and Richard Kueng. Fast and robust quantum state tomography from few basis measurements. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021. DOI: 10.4230/LIPICS.TQC.2021.7.

[31] Li Gao, Marius Junge, and Nicholas LaRacuente. Fisher Information and Logarithmic Sobolev Inequality for Matrix-Valued Functions. *Annales Henri Poincaré*, 21(11):3409–3478, November 2020. ISSN 1424-0637, 1424-0661. DOI: 10.1007/s00023-020-00947-9.

[32] Nathael Gozlan and Christian Léonard. Transport inequalities. a survey. *arXiv preprint arXiv:1003.3852*, 2010.

[33] Jeongwan Haah, Aram Wettroth Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu. Sample-optimal tomography of quantum states. *IEEE Trans. Inf. Theory*, 63(9):5628–5641, 2017. DOI: 10.1109/TIT.2017.2719044.

[34] Jeongwan Haah, Robin Kothari, and Ewin Tang. Optimal learning of quantum Hamiltonians from high-temperature Gibbs states, 2021. arXiv:2108.04842v1.

[35] Aram W Harrow, Saeed Mehraban, and Mehdi Soleimanifar. Classical algorithms, correlation decay, and complex zeros of partition functions of quantum many-body systems. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 378–386, 2020. DOI: 10.1145/3357713.3384322.

[36] M. B. Hastings. Quantum belief propagation: An algorithm for thermal quantum systems. *Physical Review B*, 76(20), nov 2007. DOI: 10.1103/physrevb.76.201102.

[37] M. B. Hastings. Quantum belief propagation: An algorithm for thermal quantum systems. *Physical Review B*, 76(20):201102, November 2007. ISSN 1098-0121, 1550-235X. DOI: 10.1103/PhysRevB.76.201102.

[38] M. B. Hastings. Locality in Quantum Systems. *arXiv:1008.5137 [math-ph, physics:quant-ph]*, August 2010. arXiv: 1008.5137.

[39] Matthew B. Hastings and Tohru Koma. Spectral Gap and Exponential Decay of Correlations. *Communications in Mathematical Physics*, 265 (3):781–804, August 2006. ISSN 0010-3616, 1432-0916. DOI: 10.1007/s00220-006-0030-4.

[40] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics*, June 2020. DOI: 10.1038/s41567-020-0932-7.

[41] Yichen Huang. Locally accurate matrix product approximation to thermal states. *Science Bulletin*, 66(24):2456–2457, December 2021. ISSN 2095-9273. DOI: 10.1016/j.scib.2021.08.011. URL http://dx.doi.org/10.1016/j.scib.2021.08.011.

[42] E. T. Jaynes. Information Theory and Statistical Mechanics. *Physical Review*, 106(4):620–630, May 1957. ISSN 0031-899X. DOI: 10.1103/PhysRev.106.620.

[43] Andrew Jena, Scott Genin, and Michele Mosca. Pauli Partitioning with Respect to Gate Sets. *arXiv:1907.07859 [quant-ph]*, July 2019. arXiv: 1907.07859.

[44] Michael J. Kastoryano and Fernando G. S. L. Brandão. Quantum gibbs samplers: The commuting case. *Communications in Mathematical Physics*, 344(3):915–957, May 2016. ISSN 1432-0916. DOI: 10.1007/s00220-016-2641-8.

[45] Bobak Toussi Kiani, Giacomo De Palma, Milad Marvian, Zi-Wen Liu, and Seth Lloyd. Quantum earth mover's distance: A new approach to learning quantum data. *arXiv preprint arXiv:2101.03037*, 2021.

[46] M. Kliesch, C. Gogolin, M.J. Kastoryano, A. Riera, and J. Eisert. Locality of Temperature. *Physical Review X*, 4(3):031019, July 2014. ISSN 2160-3308. DOI: 10.1103/PhysRevX.4.031019.

[47] Martin Kliesch, Christian Gogolin, and Jens Eisert. Lieb-Robinson Bounds and the Simulation of Time-Evolution of Local Observables in Lattice Systems. In Volker Bach and Luigi Delle Site, editors, *Many-Electron Approaches in Physics, Chemistry and Mathematics*, pages 301–318. Springer International Publishing, Cham, 2014. ISBN 978-3-319-06378-2 978-3-319-06379-9. DOI: 10.1007/978-3-319-06379-9_17. Series Title: Mathematical Physics Studies.

[48] Christian Kokail, Rick van Bijnen, Andreas Elben, Benoît Vermersch, and Peter Zoller. Entanglement hamiltonian tomography in quantum simulation. *Nature Physics*, 17(8):936–942, June 2021. ISSN 1745-2481. DOI: 10.1038/s41567-021-01260-w.

[49] Tomotaka Kuwahara and Keiji Saito. Gaussian

concentration bound and ensemble equivalence in generic quantum many-body systems including long-range interactions. *Annals of Physics*, 421: 168278, 2020. DOI: 10.1016/j.aop.2020.168278.

[50] Tomotaka Kuwahara and Keiji Saito. Eigenstate Thermalization from the Clustering Property of Correlation. *Physical Review Letters*, 124(20): 200604, May 2020. ISSN 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.124.200604.

[51] Tomotaka Kuwahara, Kohtaro Kato, and Fernando G. S. L. Brandão. Clustering of conditional mutual information for quantum Gibbs states above a threshold temperature. *Phys. Rev. Lett.*, 124:220601, Jun 2020. DOI: 10.1103/PhysRevLett.124.220601.

[52] Tomotaka Kuwahara, Álvaro M. Alhambra, and Anurag Anshu. Improved Thermal Area Law and Quasilinear Time Algorithm for Quantum Gibbs States. *Physical Review X*, 11(1):011047, March 2021. ISSN 2160-3308. DOI: 10.1103/PhysRevX.11.011047.

[53] B. P. Lanyon, C. Maier, M. Holzäpfel, T. Baumgratz, C. Hempel, P. Jurcevic, I. Dhand, A. S. Buyskikh, A. J. Daley, M. Cramer, M. B. Plenio, R. Blatt, and C. F. Roos. Efficient tomography of a quantum many-body system. *Nature Physics*, 13(12):1158–1162, September 2017. DOI: 10.1038/nphys4244.

[54] Andrea Montanari et al. Computational implications of reducing data to sufficient statistics. *Electronic Journal of Statistics*, 9(2):2370–2390, 2015. DOI: 10.1214/15-ejs1059.

[55] Ryan O'Donnell and John Wright. Efficient quantum tomography. In Daniel Wichs and Yishay Mansour, editors, *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 899–912. ACM, 2016. DOI: 10.1145/2897518.2897544.

[56] Giacomo De Palma, Milad Marvian, Dario Trevisan, and Seth Lloyd. The quantum wasserstein distance of order 1. *IEEE Transactions on Information Theory*, pages 1–1, 2021. DOI: 10.1109/tit.2021.3076442.

[57] David Poulin. Lieb-Robinson Bound and Locality for General Markovian Quantum Dynamics. *Physical Review Letters*, 104(19):190401, May 2010. ISSN 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.104.190401.

[58] Maxim Raginsky and Igal Sason. *Concentration of Measure Inequalities in Information Theory, Communications, and Coding*. Now Publishers, Norwell, MA, 2014. ISBN 978-1-60198-907-9. OCLC: 1193114287.

[59] Cambyse Rouzé and Nilanjana Datta. Concentration of quantum states from quantum functional and transportation cost inequalities. *Journal of Mathematical Physics*, 60(1):012202, January 2019. DOI: 10.1063/1.5023210.

[60] Brian Swingle and Isaac H. Kim. Reconstructing quantum states from local data. *Physical Review Letters*, 113(26):260501, December 2014. ISSN 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.113.260501. arXiv: 1407.2658.

[61] M. Talagrand. Transportation cost for Gaussian and other product measures. *Geometric and Functional Analysis*, 6(3):587–600, may 1996. DOI: 10.1007/bf02249265.

[62] Hal Tasaki. On the Local Equivalence Between the Canonical and the Microcanonical Ensembles for Quantum Spin Systems. *Journal of Statistical Physics*, 172(4):905–926, August 2018. ISSN 0022-4715, 1572-9613. DOI: 10.1007/s10955-018-2077-y.

[63] Kristan Temme, Fernando Pastawski, and Michael J Kastoryano. Hypercontractivity of quasi-free quantum semigroups. *Journal of Physics A: Mathematical and Theoretical*, 47(40):405303, October 2014. ISSN 1751-8113, 1751-8121. DOI: 10.1088/1751-8113/47/40/405303.

[64] Giacomo Torlai, Guglielmo Mazzola, Juan Carrasquilla, Matthias Troyer, Roger Melko, and Giuseppe Carleo. Neural-network quantum state tomography. *Nature Physics*, 14(5):447–450, may 2018. ISSN 1745-2473. DOI: 10.1038/s41567-018-0048-5.

[65] Cédric Villani. *Optimal Transport*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-540-71049-3 978-3-540-71050-9. DOI: 10.1007/978-3-540-71050-9.

[66] Marc Vuffray, Sidhant Misra, Andrey Lokhov, and Michael Chertkov. Interaction screening: Efficient and sample-optimal learning of Ising models. In *Advances in Neural Information Processing Systems*, pages 2595–2603, 2016.

[67] Jun Wang, Zhao-Yu Han, Song-Bo Wang, Zeyang Li, Liang-Zhu Mu, Heng Fan, and Lei Wang. Scalable quantum tomography with fidelity estimation. *Phys. Rev. A*, 101:032321, Mar 2020. DOI: 10.1103/PhysRevA.101.032321.

[68] Akram Youssry, Christopher Ferrie, and Marco Tomamichel. Efficient online quantum state estimation using a matrix-exponentiated gradient method. *New J. Phys.*, 21(3):033006, 2019. DOI: 10.1088/1367-2630/ab0438.

Supplemental Material

This is the supplemental material to "Learning many-body states from very few copies". We will start in Sec. A with a review of the basic properties of the maximum entropy principle to learn quantum states. This is followed by a discussion of Lipschitz constants, Wasserstein distances and transportation cost inequalities in Sec. B. After that, in Sec. C we more explicitly discuss the interplay between the maximum entropy method and transportation cost inequalities. We then briefly discuss scenarios in which the postprocessing required for the maximum entropy method can be performed efficiently in Sec. E. In Sec. F we discuss a class of examples where we show that all technical results required to obtain the strongest guarantees of our work hold, that is, Gibbs states of commuting Hamiltonians at high enough temperature and 1D commuting Hamiltonians. Finally, in Sec. G we discuss lower bounds on the sample complexity of both shadow protocols and many-body algorithms that focus on a recovery in trace distance.

We start by setting some basic notations. Throughout this article, we denote by $\mathcal{M}_k$ the algebra of $k \times k$ matrices on $\mathbb{C}^k$, whereas $\mathcal{M}_k^{\mathrm{sa}}$ denotes the subspace of self-adjoint matrices. The set of quantum states over $\mathbb{C}^k$ is denoted by $\mathcal{D}_k$. Typically, k will be taken as d^n for n qudit systems. The trace on $\mathcal{M}_k$ is denoted by tr. Given two quantum states ρ, σ, we denote by $S(\rho) = -\operatorname{tr}[\rho \log(\rho)]$ the von Neumann entropy of ρ, and by $D(\rho\|\sigma)$ the relative entropy between ρ and σ, i.e. $D(\rho\|\sigma) = \operatorname{tr}[\rho(\log(\rho) - \log(\sigma))]$ whenever the support of ρ is contained in that of σ and $+\infty$ otherwise. The trace distance is denoted by $\|\rho - \sigma\|_{tr} := \operatorname{tr}[|\rho - \sigma|]$ and the operator norm of an observable by $\|O\|_\infty$. Scalar products are denoted by $\langle \cdot | \cdot \rangle$. Moreover, we denote the ℓ_p norm of vectors by $\| \cdot \|_{\ell_p}$, and for $x \in \mathbb{R}^m$ and $r \in \mathbb{R}$, $B_{\ell_p}(x, r)$ denotes the ball of radius r in ℓ_p norm around x. The identity matrix is denoted by I. The adjoint of an operator A is denoted by $A^\dagger$ and that of a channel Φ with respect to the trace inner product by Φ^*. For a hypergraph $G = (V, E)$ we will denote the distance between subsets of vertices induced by the hypergraph by dist.

A Maximum entropy principle for quantum Gibbs states

One of the main aspects of this work concerns the effectiveness of the maximum entropy method for the tomography of quantum Gibbs states in various settings and regimes. Thus, we start by recalling some basic properties of the maximum entropy method. Our starting assumption is that the target state is well-described by a quantum Gibbs state with respect to a known set of operators and that we are given an upper bound on the inverse temperature:

Definition A.1 (Gibbs state with respect to observables)**.** *Given a set of observables $\mathcal{E} = \{E_i\}_{i=1}^m$, $E_1, \ldots, E_m \in \mathcal{M}_{d^n}^{\mathrm{sa}}$ being linearly independent with $\|E_i\|_\infty \leq 1$, we call a state $\sigma \in \mathcal{D}_{d^n}$ a Gibbs state at inverse temperature $\beta > 0$ if there exists a vector $\lambda \in \mathbb{R}^m$ with $\|\lambda\|_{\ell_\infty} \leq 1$ such that:*

$$\sigma = \exp\left(-\beta \sum_{i=1}^m \lambda_i E_i\right) / \mathcal{Z}(\lambda), \qquad \text{where} \qquad \mathcal{Z}(\lambda) = \operatorname{tr}\left[\exp\left(-\beta \sum_{i=1}^m \lambda_i E_i\right)\right] \tag{19}$$

denotes the partition function. In what follows, we will denote σ by $\sigma(\lambda)$ and $\sum_i \lambda_i E_i = H(\lambda)$, where the dependence of $\sigma(\lambda)$ on β is implicitly assumed.

We are mostly interested in the regime where $m \ll d^n$. Then the above condition can be interpreted as imposing that the matrix $\log(\sigma)$ is sparse with respect to a known basis $\mathcal{E}$. A canonical example for such states are Gibbs states of local Hamiltonians on a lattice, for which $m = \mathcal{O}(n)$ and the observables E_i are taken as tensor products of Pauli matrices acting on neighboring sites. But we could also consider a basis consisting of quasi-local operators or some subspace of Pauli strings.

Next, we review some basic facts about quantum Gibbs states. One of their main properties is that they satisfy a well-known maximum entropy principle [42]. This allows us to simultaneously show that the expectation values of the observables $\mathcal{E}$ completely characterize the state $\sigma(\lambda)$ and further provides us with a variational principle to learn a description from which we can infer an approximation of other expectation values. Let us start with the standard formulation of the maximum entropy principle:

Proposition A.1 (Maximum entropy principle)**.** *Let $\sigma(\lambda) \in \mathcal{D}_{d^n}$ be a quantum Gibbs state (19) with respect to the basis $\mathcal{E}$ at inverse temperature β and introduce $e_i(\lambda) := \operatorname{tr}[\sigma(\lambda) E_i]$ for $i = 1, \ldots, m$. Then $\sigma(\lambda)$ is the unique optimizer of the maximum entropy problem:*

$$\begin{aligned} &\underset{\rho \in \mathcal{D}_{d^n}}{\text{maximize}} \quad S(\rho) \\ &\text{subject to} \quad \operatorname{tr}[E_i \rho] = e_i(\lambda) \quad \text{for all } i = 1, \ldots, m. \end{aligned} \tag{20}$$

Moreover, $\sigma(\lambda)$ optimizes:

$$\min_{\substack{\mu\in\mathbb{R}^m \\ \|\mu\|_{\ell_\infty}\leq 1}} \log(\mathcal{Z}(\mu)) + \beta \sum_{i=1}^m \mu_i e_i(\lambda)\,. \tag{21}$$

Proof. The proof is quite standard, but we include it for completeness. Note that for any state $\rho \neq \sigma$ that is a feasible point of Eq. (20) we have that:

$$\begin{aligned} S(\sigma(\lambda)) - S(\rho) &= D(\rho\|\sigma(\lambda)) + \operatorname{tr}\left[(\rho - \sigma(\lambda))\log(\sigma(\lambda))\right] \\ &= D(\rho\|\sigma(\lambda)) - \beta \sum_{i=1}^m \lambda_i \operatorname{tr}\left[E_i\left(\rho - \sigma(\lambda)\right)\right] \\ &= D(\rho\|\sigma(\lambda)) > 0\,, \end{aligned}$$

where we have used the fact that $\operatorname{tr}\left[E_i\left(\rho - \sigma(\lambda)\right)\right] = 0$ for all feasible points and that the relative entropy between two different states is strictly positive. This shows that $\sigma(\lambda)$ is the unique solution of (20). Eq. (21) is nothing but the dual program of Eq. (20). □

Eq. (20) above gives a variational principle to find a quantum Gibbs state corresponding to certain expectation values. As it is well-known, one can use gradient descent to solve the problem in Eq. (21), as it is a strongly convex problem. Various recent works have discussed learning of Gibbs states [5, 48, 60] and it is certainly not a new idea to do so through maximum entropy methods. Nevertheless, we will discuss how to perform the postprocessing in more detail, as some recent results allow us to give this algorithm stronger performance guarantees. Finally, it should be said that although we draw inspiration from [5], our main goal will be to learn a set of parameters $\mu \in \mathbb{R}^m$ such that the Gibbs states $\sigma(\mu)$ and $\sigma(\lambda)$ are approximately the same on sufficiently regular observables while optimizing the sample complexity. This is in contrast to the goal of [5], which was to learn the vector of parameters λ. Learning λ corresponds to a stronger requirement, in the sense that if the vectors of parameters are close, then the underlying states are also close, as made precise in the following Prop. A.2.

One of the facts that we are going to often exploit is that it is possible to efficiently estimate the relative entropy between two Gibbs states $\sigma(\lambda)$ and $\sigma(\mu)$ given the parameters λ, μ and the expectation values of observables in $\mathcal{E}$. This also yields an efficiently computable bound on the trace distance. Indeed, as observed in [5], we have that:

Proposition A.2. *Let $\sigma(\mu), \sigma(\lambda) \in \mathcal{D}_{d^n}$ be Gibbs states with respect to a set of observables $\mathcal{E}$ at inverse temperature β. Denote $e(\lambda) = (\operatorname{tr}\left[\sigma(\lambda)E_i\right])_i \in \mathbb{R}^m$. Then*

$$\|\sigma(\mu) - \sigma(\lambda)\|_{tr}^2 \leq D(\sigma(\mu)\|\sigma(\lambda)) + D(\sigma(\lambda)\|\sigma(\mu)) = -\beta\,\langle \lambda - \mu | e(\lambda) - e(\mu)\rangle. \tag{22}$$

Proof. The equality in Eq. (22) follows from a simple manipulation. Indeed:

$$D(\sigma(\mu)\|\sigma(\lambda)) + D(\sigma(\lambda)\|\sigma(\mu)) = \beta \sum_{i=1}^m (\lambda_i - \mu_i) \operatorname{tr}\left[[\sigma(\mu) - \sigma(\lambda)]E_i\right].$$

The bound on the trace distance then follows by applying Pinkser's inequality. □

The statement of Proposition A.2 allows us to obtain quantitative estimates on how well a given Gibbs state approximates another one in terms of the expectation values of known observables. In particular, a simple application of Hölder's inequality shows that if two Gibbs states are such that $|\operatorname{tr}\left[E_i[\sigma(\mu) - \sigma(\lambda)]\right]| \leq \epsilon$, then the sum of their relative entropies is at most

$$\beta\,|\langle \lambda - \mu | e(\lambda) - e(\mu)\rangle| \leq \beta\,\|\lambda - \mu\|_{\ell_\infty}\, m\epsilon \leq 2m\epsilon\,\beta\,, \tag{23}$$

where the outer bound arises from our assumption that $\|\lambda\|_{\ell_\infty}, \|\mu\|_{\ell_\infty} \leq 1$. Moreover, it is straightforward to relate the difference of the target function in Eq. (21) evaluated at two vectors to the difference of relative entropies between the target state and their corresponding Gibbs states:

Lemma A.1. *Let $\sigma(\lambda) \in \mathcal{D}_{d^n}$ be a Gibbs state with respect to a set of observables $\mathcal{E}$ at inverse temperature $\beta > 0$ and define, for any $\mu \in \mathbb{R}^m$,*

$$f(\mu) := \log(\mathcal{Z}(\mu)) + \beta \sum_{i=1}^m e_i(\lambda)\,\mu_i\,. \tag{24}$$

Then for any other two vectors $\mu, \xi \in \mathbb{R}^m$ with $\|\mu\|_{\ell_\infty}, \|\xi\|_{\ell_\infty} \leq 1$:

$$f(\mu) - f(\xi) = D(\sigma(\lambda)\|\sigma(\mu)) - D(\sigma(\lambda)\|\sigma(\xi)).$$

Proof. The proof follows from straightforward manipulations. □

Thus, we see that a decrease of the target function f when solving the max entropy problem is directly connected to the decrease of the relative entropy between the target state and the current iterate. We will later use this to show the convergence of gradient descent for solving the max entropy problem with arbitrary $\mathcal{E}$. However, before that we discuss how the convergence of the state $\sigma(\mu)$ to $\sigma(\lambda)$ is related to the convergence of the parameters μ to λ.

A.1 Strong convexity and convergence guarantees

The maximum entropy problem (20) being a convex problem, it should come as no surprise that properties of the Hessian of the function being optimized are vital to understanding its complexity and stability. For the maximum entropy problem, the Hessian at a point is given by a generalized covariance matrix corresponding to the underlying Gibbs state. As the results of [5] showcase, the eigenvalues of such covariance matrices govern both the stability of Eq. (21) with respect to μ and the convergence of gradient descent to solve it. To see why, we recall some basic notions of optimization of convex functions and refer to [14] for an overview.

Definition A.2 (Strong convexity)**.** *Let $C \subset \mathbb{R}^m$ be a convex set. A twice differentiable function $f : C \to \mathbb{R}$ is called strongly convex with parameters $U, L > 0$ if we have for all $x \in C$ that:*

$$UI \geq \nabla^2 f(x) \geq LI.$$

The optimization of strongly convex functions is well-understood. Indeed, we have:

Proposition A.3. *Let $C \subset \mathbb{R}^m$ be a convex set and $f : C \to \mathbb{R}$ be strongly convex with parameters L, U as in the definition above. Then, for all $\epsilon > 0$, the optimal value $\alpha := \min_{x \in C} f(x)$ is achieved up to error ϵ by the gradient descent algorithm initiated at $x^0 \in C$ with step size U^{-1} after at most S steps for*

$$S \leq \frac{U}{L} \log\left(\frac{f(x^0) - \alpha}{\epsilon}\right). \tag{25}$$

Moreover, the gradient norm satisfies $\|\nabla f(x^k)\|_{\ell_2}^2 \leq \delta$ after at most S_∇ steps with

$$S_\nabla \leq \frac{U}{L} \log\left(\frac{2L(f(x^0) - \alpha)}{\delta}\right).$$

Finally, we have for all $\mu, \lambda \in C$ that:

$$\|\mu - \lambda\|_{\ell_2} \leq L^{-1} \|\nabla f(\mu) - \nabla f(\lambda)\|_{\ell_2}. \tag{26}$$

Proof. These are all standard results that can be found e.g. in [14, Section 9]. □

To see the relevance of these results for the maximum entropy problem, we recall the following Lemma:

Lemma A.2. *Let $C = B_{\ell_\infty}(0, 1) \subset \mathbb{R}^m$, let $\sigma(\lambda) \in \mathcal{D}_{d^n}$ be a Gibbs state with respect to a set of operators $\mathcal{E}$ at inverse temperature β and define $f : C \to \mathbb{R}$ as in Eq. (24). Then:*

$$(\nabla f(\mu))_i = \beta \operatorname{tr}\left[(\sigma(\lambda) - \sigma(\mu))E_i\right] \tag{27}$$

and

$$\left(\nabla^2 f(\mu)\right)_{ij} = \frac{\beta^2}{2} \operatorname{tr}\left[\{E_j, \Phi_{H(\mu)}(E_i)\}\sigma(\mu)\right] - \beta^2 e_i(\mu) e_j(\mu), \tag{28}$$

with

$$\Phi_{H(\mu)}(E_i) = \int_{-\infty}^{+\infty} \nu_\beta(t) e^{-iH(\mu)t} E_i e^{iH(\mu)t} dt$$

where $\nu_\beta(t)$ is a probability density function whose Fourier transform is given by:

$$\hat{\nu}_\beta(\omega) = \frac{2 \tanh\left(\frac{\beta\omega}{2}\right)}{\beta\omega}.$$

Proof. The quantum belief propagation theorem [36] states that:

$$\frac{\partial}{\partial \lambda_i} e^{-\beta H(\lambda)} = -\frac{\beta}{2}\{e^{-\beta H(\lambda)}, \Phi_{H(\lambda)}(E_i)\}\,.$$

The claim then follows from a simple computation. □

Thus, we see that in order to compute the gradient of the target function f for the maximum entropy problem, we simply need to compute the expectation values of observables $\mathcal{E}$ on the current state and on the target state. Moreover, the Hessian is given by a generalized covariance matrix of the quantum Gibbs state. That this should indeed be interpreted as a covariance matrix is most easily seen by considering commuting Hamiltonians. Then indeed we have:

$$\left(\nabla^2 f(\mu)\right)_{ij} = \beta^2 \left[\mathrm{tr}\left[\sigma(\mu) E_i E_j\right] - e_i(\mu) e_j(\mu)\right]\,.$$

For any Gibbs state it holds that:

Proposition A.4. *For all $\mu \in B_{\ell_\infty}(0,1) \subset \mathbb{R}^m$, inverse temperature $\beta > 0$ and set of operators $\mathcal{E}$ of cardinality m, we have:*

$$\nabla^2 f(\mu) \leq 2\beta^2 m\, I\,.$$

Proof. Note that

$$\left|\mathrm{tr}\left[E_i \sigma(\mu) e^{iH(\mu)t} E_j e^{-iH(\mu)t}\right]\right| \leq 1\,,$$

by Hölder's inequality, the submultiplicativity and unitary invariance of the operator norm and the fact that $\|E_i\|_\infty \leq 1$. Similarly, we have that $|e_i(\mu)|, |e_j(\mu)| \leq 1$. Thus, by Lemma A.2, $|\left(\nabla^2 f(\mu)\right)_{ij}| \leq 2\beta^2$. As $\nabla^2 f(\mu)$ is an $m \times m$ matrix, it follows from Gershgorin's circle theorem that $\nabla^2 f(\mu) \leq 2\beta^2 m I$. □

The proof above also showcases how exponential decay of correlations can be used to sharpen estimates on the maximal eigenvalue of $\nabla^2 f$, since in that case $\left(\nabla^2 f(\mu)\right)_{ij}$ will have exponentially decaying entries. We discuss this in more detail when we focus on many-body states, for which we also consider the more challenging question of lower bounds.

A.2 Convergence with approximate gradient computation and expectation values

Proposition A.3 already establishes the convergence of gradient descent whenever we can compute the gradient exactly and have a bound on L. Moreover, we see from Lemma A.2 that, in order to compute the gradient of the function f above, it suffices to estimate local expectation values. Moreover, it is a standard result that gradient descent is strictly decreasing for strictly convex problems [14].

However, in many settings it is only possible or desirable to compute the expectation values of quantum Gibbs states approximately. Moreover, the expectation values of the target state are only known up to statistical fluctuations. It is then not too difficult to see that gradient descent still offers the convergence guarantees if we only approximately compute the gradient. We state the exact convergence guarantees and precision requirement for completeness.

Theorem A.1 (Computational complexity and convergence guarantees)**.** *Let $\sigma(\lambda) \in \mathcal{D}_{d^n}$ be a quantum Gibbs state at inverse temperature β with respect to a set of operators $\mathcal{E}$ and $C_{\mathcal{E}}$ be the computational cost of computing $e'(\mu)$ satisfying*

$$\|e'(\mu) - e(\mu)\|_{\ell_2} \leq \delta_\mu$$

for $\mu \in B_{\ell_\infty}(0,1)$ and $\delta_\mu > 0$. Moreover, assume that we are given an estimate $e'(\lambda)$ of $e(\lambda)$ satisfying

$$\|e(\lambda) - e'(\lambda)\|_\infty \leq \epsilon \tag{29}$$

and that the partition function is strongly convex with parameters U, L. Then gradient descent starting at $\mu = 0$ with step size $\frac{1}{cU}$ and input data $e'(\lambda)$ converges to a state $\sigma(\mu_*)$ satisfying:

$$\begin{aligned}\|\sigma(\lambda) - \sigma(\mu_*)\|_{tr}^2 &\leq D(\sigma(\lambda)\|\sigma(\mu_*)) + D(\sigma(\mu_*)\|\sigma(\lambda)) \\ &= \mathcal{O}(\beta\delta_\mu \min\{\sqrt{m}, \beta L^{-1}\delta_\mu\} + \beta\epsilon \min\{1, L^{-1}\beta\epsilon\}m).\end{aligned}$$

in

$$\mathcal{O}\left(\min\left\{UC_{\mathcal{E}}\beta^{-2}n\log(d)\delta_{\mu}^{-2}, \frac{UC_{\mathcal{E}}}{L}\log(n\epsilon^{-1})\right\}\right)$$

time.

We will prove this theorem at the end of this section, as before we will need some auxiliary statements. But the reader familiar with basic concepts from convex optimization should feel comfortable to skip them.

Proposition A.5 (Convergence of gradient descent with constant relative precision)**.** *Let $\sigma(\lambda) \in \mathcal{D}_{d^n}$ be a quantum Gibbs state at inverse temperature β with respect to a set of operators $\mathcal{E}$, and for a Gibbs state $\sigma(\mu)$ let $z(\mu) \in \mathbb{R}^m$ be a vector such that*

$$\|z(\mu) - \beta(e(\lambda) - e(\mu))\|_{\ell_2} \leq \frac{\beta}{4c}\|e(\mu) - e(\lambda)\|_{\ell_{\ell_2}} \tag{30}$$

for some $c > 10$. Then we have that:

$$D(\sigma(\lambda)\|\sigma(\mu - \tfrac{z(\mu)}{cU})) - D(\sigma(\lambda)\|\sigma(\mu)) \leq -\frac{9\beta^2\|e(\mu) - e(\lambda)\|_{\ell_2}^2}{10\,c\,U}\,, \tag{31}$$

where U is a uniform bound on the operator norm of the Hessian of the function f defined in Eq. (24).

Proof. From a Taylor expansion and strong convexity we have for any two points μ, ξ that:

$$f(\xi) \leq f(\mu) + \langle \nabla f(\mu) | (\xi - \mu)\rangle + \frac{U}{2}\|\xi - \mu\|_{\ell_2}^2.$$

Note that $\nabla f(\mu) = \beta\,(e(\lambda) - e(\mu))$ by Eq. (27). Setting $\xi = \mu - \frac{z}{cU} = \mu + \frac{1}{cU}\,(-\nabla f(\mu) + \nabla f(\mu) - z)$ we obtain:

$$\begin{aligned}
&f\left(\mu - \frac{z}{cU}\right) \\
&\quad\leq f(\mu) - \frac{1}{cU}\|\nabla f\|_{\ell_2}^2 + \frac{1}{cU}\langle \nabla f(\mu) | \nabla f(\mu) - z\rangle + \frac{1}{2c^2U}\| - \nabla f(\mu) + \nabla f(\mu) - z\|_{\ell_2}^2 \\
&\quad\leq f(\mu) - \frac{1}{cU}\|\nabla f\|_{\ell_2}^2 + \frac{1}{cU}\|\nabla f(\mu)\|_{\ell_2}\,\|\nabla f(\mu) - z\|_{\ell_2} + \frac{1}{2c^2U}\left(\|\nabla f(\mu)\|_{\ell_2} + \|\nabla f(\mu) - z\|_{\ell_2}\right)^2,
\end{aligned}$$

where in the last step we used the Cauchy-Schwarz inequality. By our assumption in Eq. (30) for $z \equiv z(\mu)$ we have:

$$\begin{aligned}
&f(\mu) - \frac{1}{cU}\|\nabla f\|_{\ell_2}^2 + \frac{1}{cU}\|\nabla f(\mu)\|_{\ell_2}\|\nabla f(\mu) - z\|_{\ell_2} + \frac{1}{2c^2U}\left(\|\nabla f(\mu)\|_{\ell_2} + \|\nabla f(\mu) - z\|_{\ell_2}\right)^2 \\
&\quad\leq f(\mu) - \frac{1}{cU}\|\nabla f\|_{\ell_2}^2 + \frac{1}{4c^2U}\|\nabla f(\mu)\|_{\ell_2}^2 + \frac{(1 + (4c)^{-1})^2}{2c^2U}\|\nabla f(\mu)\|_{\ell_2}^2 \\
&\quad= f(\mu) - \frac{1}{cU}\left(1 - \frac{1}{4c} - \frac{(1 + (4c)^{-1})^2}{2c}\right)\|\nabla f(\mu)\|_{\ell_2}^2
\end{aligned}$$

and it can be readily checked that $\frac{1}{4c} + \frac{(1+(4c)^{-1})^2}{2c} \leq \frac{1}{10}$ for $c \geq 10$. To conclude the proof, note that by Lemma A.1:

$$D(\sigma(\lambda)\|\sigma(\mu - \tfrac{z}{cU})) - D(\sigma(\lambda)\|\sigma(\mu)) = f(\mu - \tfrac{z}{cU})) - f(\mu)\,,$$

and insert $\|\nabla f(\mu)\|_{\ell_2}^2 = \beta^2\|e(\mu) - e(\lambda)\|_{\ell_2}^2$. □

Thus, we see that we make constant progress for the gradient descent algorithm if we only compute the derivative up to constant relative precision. We show now how to pick our stopping criterium based on approximate computations of the gradient which ensure convergence in polynomial time.

Proposition A.6. *Let $\sigma(\lambda) \in \mathcal{D}_{d^n}$ be a quantum Gibbs state at inverse temperature β with respect to a set of operators $\mathcal{E}$. Suppose that at each time step t of gradient descent we compute an estimate $e'(\mu_t)$ of $e(\mu_t)$ that satisfies*

$$\|e'(\mu_t) - e(\mu_t)\|_{\ell_{\ell_2}} \leq \delta_{\mu}\,,$$

and set the stopping criterion to be:

$$\|e(\lambda) - e'(\mu_*)\|_{\ell_2} < (4c+1)\delta_\mu \,.$$

for some constant $c > 10$. Then gradient descent starting at $\mu = 0$ with update rule $\mu_{t+1} := \mu_t - \frac{\beta(e(\lambda)-e'(\mu_t))}{cU}$ will converge to a state $\sigma(\mu_*)$ satisfying:

$$\|\sigma(\lambda) - \sigma(\mu_*)\|_{tr}^2 \leq D(\sigma(\lambda)\|\sigma(\mu_*)) + D(\sigma(\mu_*)\|\sigma(\lambda)) \leq 2(4c+1)\beta\delta_\mu\sqrt{m}$$

after at most $\mathcal{O}(U\beta^{-2}n\log(d)\delta_\mu^{-2})$ iterations.

Proof. First, we show that the relative precision bound required for Proposition A.5 holds under these assumptions. By our choice of the stopping criterium, at each time step we have the property that, while we did not stop,

$$\begin{aligned}\|e(\lambda) - e(\mu_t)\|_{\ell_2} &= \|e(\mu) - e'(\mu_t) + e'(\mu_t) - e(\lambda)\|_{\ell_2} \\ &\geq \|e'(\mu_t) - e(\lambda)\|_{\ell_2} - \|e(\mu_t) - e'(\mu_t)\|_{\ell_2} \\ &\geq (4c+1)\delta_\mu - \delta_\mu \\ &= 4c\delta_\mu\end{aligned}$$

by the reverse triangle inequality. As we assumed we have that $\|e'(\mu_t) - e(\mu_t)\|_{\ell_2} \leq \delta_\mu$, it follows that $(4c)^{-1}\|e(\lambda) - e(\mu_t)\|_{\ell_2} \geq \|(e'(\mu_t) - e(\lambda)) - (e(\mu_t) - e(\lambda))\|_{\ell_2}$. Multiplying the inequality by β, we see that the conditions of Proposition A.5 are satisfied for $z(\mu_t) := \beta(e(\lambda) - e'(\mu_t))$. Let us now show the convergence. By our choice of initial point, we have that:

$$D(\sigma(\lambda)\|\sigma(0)) \leq n\log(d)\,.$$

Now, suppose that we did not stop before T iterations. It follows from a telescopic sum argument and Proposition A.5 that:

$$D(\sigma(\lambda)\|\sigma(\mu_T)) \leq n\log(d) - T\frac{9\beta^2(4c+1)^2\delta_\mu^2}{10\,c\,U}\,,$$

since $\|e'(\mu_t) - e(\lambda)\| \geq (4c+1)\delta_\mu$ at all iterations because we did not halt. As the relative entropy is positive, it follows that $T = \mathcal{O}(\beta^{-2}Uc^{-1}n\log(d)\delta_\mu^{-2})$ before the stopping criterium is met. The recovery guarantee whenever the stopping criterium is met follows from Proposition A.2, the Cauchy-Schwarz inequality and the equivalence of norms $\|\lambda - \mu\|_{\ell_2} \leq \sqrt{m}\|\lambda - \mu\|_{\ell_\infty} \leq 2\sqrt{m}$. □

Since we proved in Proposition A.4 that $U = \mathcal{O}(\beta^2 m)$, it follows that the number of iterations is $\mathcal{O}(nm)$. Thus, we see that having a lower bound on the Hessian is not necessary to ensure convergence, but it can speed it up exponentially:

Proposition A.7 (Exponential convergence of gradient descent with approximate gradients)**.** *In the same setting as Proposition A.5 we have:*

$$f\left(\mu - \frac{z(\mu)}{cU}\right) - f(\lambda) \leq \left(1 - \frac{18L}{10cU}\right)(f(\mu) - f(\lambda)). \tag{32}$$

In particular, gradient descent with approximate gradient computations starting at $\mu_0 = 0$ converges after $\mathcal{O}\left(\frac{U}{L}\log(n\epsilon^{-1})\right)$ iterations to μ such that $f(\mu) - f(\lambda) \leq \epsilon$.

Proof. For any strongly convex function f and points $\mu, \xi \in C$ we have that:

$$f(\xi) \geq f(\mu) + \langle \nabla f(\mu) | (\xi - \mu)\rangle + \frac{L}{2}\|\mu - \xi\|_{\ell_2}^2.$$

As explained in [14, Chapter 9], the R.H.S. of the equation above is a convex quadratic function of ξ for μ fixed. One can then easily show that its minimum is achieved at $\tilde{\xi} = \mu - \frac{1}{L}\nabla f(\mu)$. From this we obtain:

$$f(\mu) - f(\lambda) \geq -\frac{1}{2L}\|\nabla f(\mu)\|_{\ell_2}^2 = -\frac{\beta^2\|e(\mu) - e(\lambda)\|_{\ell_2}^2}{2L}\,,$$

where the last identity follows from Eq. (27). By subtracting $f(\lambda)$ from both sides of the inequality (31) in Proposition A.5 and rearranging the terms we have that:

$$\begin{aligned} f\left(\mu - \frac{z(\mu)}{cU}\right) - f(\lambda) &\leq f(\mu) - f(\lambda) - \frac{9\beta^2\|e(\mu) - e(\lambda)\|_{\ell_2}^2}{10cU} \\ &\leq f(\mu) - f(\lambda) - \frac{18L}{10cU}(f(\mu) - f(\lambda)) \\ &= \left(1 - \frac{18L}{10cU}\right)(f(\mu) - f(\lambda)). \end{aligned}$$

This yields the claim in Eq. (32). To obtain the second claim, note that applying Eq. (32) iteratively yields that after k iterations we have that, for $\mu_k = \mu_{k-1} - \frac{z(\mu_{k-1})}{cU}$,

$$(f(\mu_k) - f(\lambda)) \leq \left(1 - \frac{18L}{10cU}\right)^k (f(0) - f(\lambda)).$$

By our choice of initial point and Lemma A.1 we have that $f(0) - f(\lambda) = \mathcal{O}(n)$, which yields the claim solving for k and noting that $-\log\left(1 - \frac{18L}{10cU}\right) = \Omega(\frac{L}{U})$. □

Remark A.1 (Comparison to mirror descent). It is also worth noting that the convergence guarantees of Proposition A.6 and update rules of gradient descent are very similar to the ones of mirror descent with the von Neumann entropy as potential, another algorithm used for learning of quantum states [1, 16, 30, 68]. In this context, mirror descent would use a similar update rule. However, instead of computing the whole gradient, i.e. all expectation values of the basis, for one iteration, mirror descent just requires us to find one i such that $|e_i(\lambda) - e_i(\mu)| \geq \delta$ and updates the Hamiltonian in the direction i. This implies that the algorithm can be run online while we still estimate some other e_i, but we will not analyze this variation more in detail here.

Finally, we assumed so far that we knew the expectation values of the target state, $e(\lambda)$, exactly. However, it follows straightforwardly from Proposition A.6 that knowing each expectation value up to an error ϵ is sufficient to ensure that the additional error due to statistical fluctuations is at most ϵm. More precisely, if we have that $\|e(\lambda) - e'(\lambda)\|_\infty \leq \epsilon$ for some $\epsilon > 0$, then any Gibbs state $\sigma(\mu_*)$ satisfying $\|e(\mu_*) - e'(\lambda)\|_{\ell_2} \leq \delta$ satisfies:

$$D(\sigma(\lambda)\|\sigma(\mu_*)) + D(\sigma(\mu_*)\|\sigma(\lambda)) \leq 2\beta\delta\sqrt{m} + 2\beta\epsilon m$$

by Proposition A.2 and a Cauchy-Schwarz inequality. With these statements at hand we are finally ready to prove Thm. A.1.

Proof of Thm. A.1. We will show in Propositions A.6, A.7 that under the conditions outlined above, the maximum entropy problem will converge to a μ_* that satisfies:

$$\|e'(\lambda) - e(\mu_*)\|_{\ell_2} \leq (4c+1)\delta_\mu.$$

Without making any assumptions on L we can then bound

$$\begin{aligned} D(\sigma(\lambda)\|\sigma(\mu_*)) + D(\sigma(\mu_*)\|\sigma(\lambda) &= \beta\,|\langle\lambda - \mu_*|e(\lambda) - e(\mu_*)\rangle| \\ &\leq \beta\,(|\langle\lambda - \mu_*|e(\lambda) - e(\mu_*)\rangle| + |\langle\lambda - \mu_*|e'(\lambda) - e(\mu_*)\rangle|) \\ &\leq (4c+1)\delta_\mu\sqrt{m} + 2\beta\epsilon m \end{aligned}$$

by Hölder inequality and our assumptions on $e'(\lambda)$. Let us now discuss how strong convexity can improve these estimates. First note that by strong convexity and Cauchy-Schwarz we have:

$$\begin{aligned} D(\sigma(\lambda)\|\sigma(\mu_*)) + D(\sigma(\mu_*)\|\sigma(\lambda) &= \beta\,|\langle\lambda - \mu_*|e(\lambda) - e(\mu_*)\rangle| \\ &\leq \beta\|e(\lambda) - e(\mu_*)\|_{\ell_2}\|\lambda - \mu_*\|_{\ell_2} \\ &\leq L^{-1}\beta^2\|e(\lambda) - e(\mu_*)\|_{\ell_2}^2 \\ &\leq L^{-1}\beta^2\left(\|e'(\lambda) - e(\mu_*)\|_{\ell_2} + \|e(\lambda) - e'(\lambda)\|_{\ell_2}\right)^2 \end{aligned}$$

which yields the claim. □

In short, we see that we can perform the recovery by simply computing the gradient approximately. In particular, as already hinted at in [5], this implies that recent methods developed to approximately compute the partition function of high-temperature quantum Gibbs states can be used to perform the postprocessing in polynomial time [35, 46, 49, 51]. This and other methods to compute the gradient are discussed in more detail in Sec. E. Furthermore, it should be noted that usually $L = \Omega(\beta^{-2})$ in the high temperature regime, making the bound independent of β for such states. We refer to Sec. D for a summary of the cases for which bounds on L are known.

B Lipschitz constants and transportation cost inequalities

In this section, we identify conditions under which it is possible to estimate all expectation values of k-local observables up to an error ϵ by measuring $\mathcal{O}(\mathrm{poly}(k, \log(n), \epsilon^{-1}))$ copies of it, where n is the system size, which constitutes an exponential improvement in some regimes. To obtain this result, we combine the maximum entropy method introduced in Section A with techniques from quantum optimal transport. In order to formalize and prove the result claimed, we resort to transportation cost inequalities and the notion of Lipschitz constants of observables, which we now introduce.

B.1 Lipschitz constants and Wasserstein metrics

Transportation cost inequalities, introduced by Talagrand in the seminal paper [61], constitute one of the strongest tools available to show concentration of measure inequalities. In the quantum setting, their study was initiated in [19, 20, 56, 59]. Here we are going to show how they can also be used in the context of quantum tomography. On a high level, a transportation cost inequality for a state σ quantifies by how much the relative entropy with respect to another state ρ is a good proxy to estimate to what extent the expectation values of sufficiently regular observables differ on the states. As maximum entropy methods allow for a straightforward control of the convergence of the learning in relative entropy (cf. Section A), they can be combined to derive strong recovery guarantees. But first we need to define what we mean by a regular observable.

We start by a short discussion of Lipschitz constants and the Wasserstein-1 distance. To obtain an intuitive grasp of these concepts, one way is to first recall the variational formulation of the trace distance of two quantum states σ, ρ:

$$\|\rho - \sigma\|_{tr} = \sup_{P=P^\dagger, \|P\|_\infty \leq 1} \mathrm{tr}\left[P(\rho - \sigma)\right].$$

Seeing probability distributions as diagonal quantum states, we recover the variational formulation of the total variation distance by noting that we may restrict to diagonal operators P. Thus, the total variation distance quantifies by how much the expectation values of arbitrary bounded functions can differ under the two distributions. However, in many situations we are not interested in expectation values of arbitrary bounded observables, but rather observables that are sufficiently regular. E.g., most observables of physical interest are (quasi)-local. Thus, it is natural to look for distance measures between quantum states that capture the notion that two states do not differ by much when restricting to expectation values of sufficiently regular observables. These concerns are particularly relevant in the context of tomography protocols, as they should be designed to efficiently obtain a state that reflects the expectation values of extensive observables of the system. As we will see, one of the ways of ensuring that the sample complexity of the tomography algorithm reflects the regularity of the observables we wish to recover is through demanding a good recovery in the Wasserstein distance of order 1 [56, 59].

In the classical setting [58, Chapter 3], one way to define a Wasserstein-1 distance between two probability distributions is by replacing the optimization over all bounded diagonal observables by that over those that are sufficiently regular: given a metric d on a sample space Ω, we define the Lipschitz constant of a function $f : \Omega \to \mathbb{R}$ to be:

$$\|f\|_{\mathrm{Lip}} := \sup_{x,y\in\Omega} \frac{|f(x) - f(y)|}{d(x,y)}.$$

Denoting the Wasserstein-1 distance by W_1, it is given for two probability measures p, q on Ω by

$$W_1(p,q) := \sup_{f:\|f\|_{\mathrm{Lip}}\leq 1} |\mathbb{E}_p(f) - \mathbb{E}_q(f)|. \tag{33}$$

That is, this metric quantifies by how much the expectation values of sufficiently regular functions can vary under p and q, in clear analogy to the variational formulation of the trace distance. We refer to [58, 65] for other interpretations and formulations of this metric.

B.1.1 Quantum Hamming Wasserstein distance

It is not immediately clear how to generalize these concepts to noncommutative spaces. There are by now several definitions of transport metrics for quantum states [26, 45, 56, 59]. As already noted in the main text, de Palma et al. defined the Lipschitz constant of an observable $O \in \mathcal{M}_{d^n}$ as [56]:

$$\|O\|_{\mathrm{Lip},\square} = \sqrt{n} \max_{1\leq i\leq n} \max_{\substack{\rho,\sigma\in\mathcal{D}_{d^n} \\ \mathrm{tr}_i[\rho]=\mathrm{tr}_i[\sigma]}} \mathrm{tr}\left[O(\rho - \sigma)\right]. \tag{34}$$

That is, the Lipschitz constant quantifies the amount by which the expectation value of an observable can change when evaluated on two states that only differ on one site. This is in analogy with the Lipschitz constants induced by the Hamming distance on the hypercube, so we denote it with $\square$. Note that in our definition we added the $\sqrt{n}$ factor, which will turn out to be convenient later. Armed with this definition, we can immediately obtain an analogous definition of the Wasserstein distance in Eq. (33) for two states ρ, σ:

$$W_{1,\square}(\rho,,\sigma) := \sup_{O=O^\dagger, \|O\|_{\mathrm{Lip},\square}\leq 1} \mathrm{tr}\left[O(\rho-\sigma)\right] . \tag{35}$$

The authors of [56] also put forth the following equivalent expression for the norm:

$$W_{1,\square}(\rho,\sigma) = \frac{\min\left\{\sum_{i=1}^n \|X^{(i)}\|_1 : \rho-\sigma = \sum_{i=1}^n X^{(i)}, X^{(i)} \in \mathcal{M}_{d^n}^{\mathrm{sa}}, \mathrm{tr}_i[X^{(i)}] = 0\right\}}{2\sqrt{n}} . \tag{36}$$

It follows from an application of Hölder's inequality combined with the variational formulation in Eq. (34) that $\|O\|_{\mathrm{Lip},\square} \leq 2\sqrt{n}\|O\|_\infty$. However, it can be the case that $\|O\|_{\mathrm{Lip}} \ll \sqrt{n}\,\|O\|_\infty$, which are exactly those observables that should be thought of as regular. This is because this signals that changing the state locally leads to significantly smaller changes to the expectation value of the observable than global ones. Two examples of this behavior are given by the observables:

$$O_1 = \sum_{i=1}^n Z_{i^c}, \qquad \text{and} \qquad O_2 = \sum_{i=1}^n Z_i ,$$

where Z_{i^c} acts as identity on i and Z_i else, i.e. $Z_{1^c} = I_1 \otimes Z_2 \otimes Z_3 \otimes \cdots \otimes Z_n$. Clearly, $\|O_1\|_\infty = \|O_2\|_\infty = n$. On the other hand, by considering the states $\rho = |0\rangle\langle 0|^{\otimes n}$ and $\sigma = |1\rangle\langle 1| \otimes |0\rangle\langle 0|^{\otimes n-1}$, we see that $\|O_1\|_{\mathrm{Lip},\square} \geq \sqrt{n}\,(2n-2)$ whereas $\|O_2\|_{\mathrm{Lip},\square} = 2\sqrt{n}$. More generally, it is not difficult to see that if $O = \sum_{i=1}^n O_i$ with $\|O_i\|_\infty \leq 1$ and we denote by $\mathrm{supp}(O_i)$ the qudits on which O_i acts nontrivially, then:

$$\|O\|_{\mathrm{Lip},\square} \leq 2\sqrt{n} \max_{1\leq j\leq n} \sum_i |\,\mathrm{supp}(O_i) \cap \{j\}| .$$

That is, the maximal number of intersections of the support of O_i on one qubit. From these examples we see that for local observables, the ratio of the operator norm and Lipschitz constant reflects the locality of the observable.

B.1.2 Quantum differential Wasserstein distance

The Wasserstein distance $W_{1,\square}$ generalizes the classical Orstein distance, that is the Wasserstein distance corresponding to the Hamming distance on bit strings. Another definition of a Lipschitz constant and its attached Wasserstein distance was put forth in [59], where the construction is based on a differential structure that bears more resemblance to that of the Lipschitz constant of a differentiable function on a continuous sample space, e.g. a smooth manifold [58]. Let us now define the notion of a noncommutative differential structure (see [21]):

Definition B.1 (Differential structure)**.** *A set of operators $L_k \in \mathcal{M}_{d^n}$ and constants $\omega_k \in \mathbb{R}$ defines a differential structure $\{L_k, \omega_k\}_{k\in\mathcal{K}}$ for a full rank state $\sigma \in \mathcal{D}_{d^n}$ if*

1 *$\{L_k\}_{k\in\mathcal{K}} = \{L_k^\dagger\}_{k\in\mathcal{K}}$;*

2 *$\{L_k\}_{k\in\mathcal{K}}$ consists of eigenvectors of the modular operator $\Delta_\sigma(X) := \sigma X \sigma^{-1}$ with*

$$\Delta_\sigma(L_k) = e^{-\omega_k} L_k . \tag{37}$$

3 *$\|L_k\|_\infty \leq 1$.*

Such a differential structure can be used to provide the set of matrices with a Lipschitz constant that is tailored to σ, see e.g. [21, 59] for more on this. In order to distinguish that constant from the one defined in (34), we will refer to it as the differential Lipschitz constant and denote it by $\|X\|_{\mathrm{Lip},\nabla}$. It is given by:

$$\|X\|_{\mathrm{Lip},\nabla} := \left(\sum_{k\in\mathcal{K}} (e^{-\omega_k/2} + e^{\omega_k/2}) \|[L_k, X]\|_\infty^2\right)^{1/2} . \tag{38}$$

The quantity $[L_k, X]$ should be interpreted as a partial derivative and is sometimes denoted by $\partial_k X$ for that reason. Then, the gradient of a matrix A, denoted by ∇A with a slight abuse of notations, refers to the vector of operator-valued coordinates $(\nabla A)_i = \partial_i A$. For ease of notation, we will denote the differential structure by the couple (∇, σ). The notion of a differential structure is also intimately connected to that of the generator of a quantum dynamical semigroup converging to σ [21], and properties of that semigroup immediately translate to properties of the metric. This is because the differential structure can be used to define an operator that behaves in an analogous way to the Laplacian on a smooth manifold, which in turn induces the heat semigroup. We refer to [21, 59] for more details.

To illustrate the differential structure version of the Lipschitz constant, it is instructive to think of the maximally mixed state. In this case, one possible choice would consist of picking the L_k to be all $1-$local Pauli strings and $\omega_j = 0$. Then the Lipschitz constant turns out to be given by:

$$\|X\|_{\mathrm{Lip},\nabla} = \left(\sum_{k \in \mathcal{K}} \|P_k X P_k - X\|_\infty^2 \right)^{1/2}, \tag{39}$$

where P_k are all $1-$local Pauli matrices. Thus, we see that this measures how much the operator changes if we act locally with a Pauli unitary on it. If we think of an operator as a function and conjugating with a Pauli as moving in a direction, the formula above indeed looks like a derivative. In fact, it is possible to make this connection rigorous, see [21].

As before, the definition in Eq. (38) yields a metric on states by duality:

$$W_{1,\nabla}(\rho, \sigma) := \sup_{X = X^\dagger,\, \|X\|_{\mathrm{Lip},\nabla} \leq 1} |\mathrm{Tr}\left(X(\rho - \sigma)\right)|\,.$$

It immediately follows from the definitions that for any observable X:

$$|\mathrm{tr}\left[X(\rho - \sigma)\right]| \leq \|X\|_{\mathrm{Lip},\nabla}\, W_{1,\nabla}(\rho, \sigma)\,. \tag{40}$$

Although this geometric interpretation opens up other interesting mathematical connections for this definition, the differential Wasserstein distance has the downside of being state dependent. It however induces a stronger topology than the quantum Hamming Wasserstein distance in some situations (see [26, Proposition 5])). In particular, the results of [26, Proposition 5]) imply that for commutative Gibbs states a TC inequality for $W_{1,\nabla}$ implies the corresponding statement for $W_{1,\square}$.

B.2 Local evolution of Lipschitz observables

As already discussed in Subsections B.1.1 and B.1.2 when we defined $\|.\|_{\mathrm{Lip},\nabla}$ and $\|.\|_{\mathrm{Lip},\square}$, Lipschitz constants can be easily controlled by making assumptions on the locality of the operators. Indeed, if we apply a local circuit to a local observable, it is straightforward to control the growth of the Lipschitz constant in terms of the growth of the support of the observable under the circuit. More precisely, in [56, Proposition 13] the authors show such a statement for discrete time evolutions with exact lightcones: if we denote by $|L|$ the size of the largest lightcone of one qubit under a channel Φ, then for any observable $O \in \mathcal{D}_{d^n}$, $\|\Phi^*(O)\|_{\mathrm{Lip},\square} \leq 2|L|\|O\|_{\mathrm{Lip},\square}$. Here we will extend such arguments to the evolution under local Hamiltonians or Lindbladians. By resorting to Lieb-Robinson bounds, we show that the Lipschitz constants $\|.\|_{\mathrm{Lip},\nabla}$ and $\|.\|_{\mathrm{Lip},\square}$ of initially local observables evolving according to a quasi-local dynamics increase at most with the effective lightcone of the evolution. Thus, short-time dynamics and shallow-depth quantum channels can only mildly increase the Lipschitz constant. This further justifies the intuition that observables with small Lipschitz constant reflect physical observables.

Lieb-Robinson (LR) bounds in essence assert that the time evolution of local observables under (quasi)-local short-time dynamics have an effective lightcone. There are various formulations of Lieb-Robinson bounds. Reviewing those in detail is beyond the scope of this work and we refer to [9, 38, 47, 57] and references therein for more details. For studying the behaviour of $\|.\|_{\mathrm{Lip},\nabla}$ under local evolutions, the most natural formulation is the *commutator version*: the generator $\mathcal{L}$ of a quasi-local dynamics $t \mapsto \Phi_t = e^{t\mathcal{L}}$ on n qudits arranged on a graph $G = (V, E)$, with graph distance dist, is said to satisfy a LR bound with LR velocity v if for any observable O_A supported on a region A and any other observable B supported on a region B, we have:

$$\|[\Phi_t^*(O_A), O_B]\|_\infty \leq c\,(e^{vt} - 1)\, g(\mathrm{dist}(A, B))\, \|O_A\|_\infty \|O_B\|_\infty\,, \tag{LR1}$$

for $g : \mathbb{N} \to \mathbb{R}_+$ a function such that $\lim_{x \to \infty} g(x) = 0$. We then have:

Proposition B.1 (Growth of differential Lipschitz constant for local evolutions)**.** *Let (∇, σ) be a differential structure on $\mathcal{M}_{d^n}$ and let $O = \sum_i O_i$ be an observable with $\|O_i\|_\infty \leq 1$. Let A_i denote the support of each O_i and B_j that of each L_j. Moreover, let $t \mapsto \Phi_t$ be an evolution satisfying Eq. (LR1) and set $o(i,j)(t) = 2$ if $A_i \cap B_j \neq \varnothing$ and $c\,(e^{vt}-1)\,g(\mathrm{dist}(A_i, B_j))$ else. Then:*

$$\|\Phi_t^*(O)\|^2_{\mathrm{Lip},\nabla} \leq \sum_{k\in\mathcal{K}} (e^{\omega_j} + e^{-\omega_j}) \left(\sum_i o_{i,j}(t)\right)^2 .$$

Proof. The proof follows almost by definition. We have:

$$\|\Phi_t^*(O)\|^2_{\mathrm{Lip},\nabla} = \sum_{j\in\mathcal{J}} (e^{\omega_j} + e^{-\omega_j}) \|[\Phi_t(O), L_j]\|^2_\infty .$$

By a triangle inequality we have that:

$$\|[\Phi_t^*(O), L_i]\|^2_\infty \leq \left(\sum_i \|[\Phi_t^*(O_i), L_j]\|_\infty\right)^2$$

For any term in the sum we have $\|[\Phi_t^*(O_i), L_j]\|_\infty \leq 2$ by the submultiplicativity of the operator norm, a triangle inequality and $\|L_j\|_\infty \leq 1$. In case O_i and L_j do not overlap, the stronger bound in Eq. (LR1) holds. □

To illustrate this bound more concretely, let us take $O = \sum_{i=1}^n Z_i$, L_j acting on $[j, j+k]$ for $j = 1, \ldots, n-k$, and $g(\mathrm{dist}(i,j)) = e^{-\mu|i-j|}$, for some constant μ. I.e. we have a 1D differential structure and a local time evolution on a 1D lattice. Then for any j:

$$\sum_i o_{i,j}(t) = k + (e^{vt}-1)\left(\sum_{i=1}^{j-1} e^{-\mu|i-j|} + \sum_{i=j+k+1}^{n} e^{-\mu|i-j|}\right) \leq k + \frac{(e^{vt}-1)e^{-\mu}}{1-e^{-\mu}}.$$

Thus,

$$\|\Phi_t^*(O)\|_{\mathrm{Lip},\nabla} \leq \sqrt{n-k}\left(k + \frac{(e^{vt}-1)e^{-\mu}}{1-e^{-\mu}}\right).$$

We see that for constant times the Lipschitz constant is still of order $\sqrt{n}k$.

Let us now derive a similar, yet somehow stronger, version of Prop. B.1 for $\|.\|_{\mathrm{Lip},\square}$. In some situations, bounds like (LR1) can be further exploited in order to prove the quasi-locality of Markovian quantum dynamics [9]: for any region $C \subset D \subset V$, there exists a local Markovian evolution $t \mapsto \Phi_t^{(D)}$ that acts non-trivially only on region D, and such that for any observable O_C supported on region C,

$$\|\left(\Phi_t^* - (\Phi_t^{(D)})^*\right)(O_C)\|_\infty \leq c'\,(e^{vt}-1)\,h(d(C, V\backslash D))\,\|O_C\|_\infty\,, \tag{LR2}$$

for some other function $h : \mathbb{N} \to \mathbb{R}_+$ such that $\lim_{x\to\infty} h(x) = 0$ and constant $c' > 0$. In other words, at small times, the channels Φ_t^* can be well-approximated by local Markovian dynamics when acting on local observables. Let us now estimate the growth of Lipschitz constants for the definition of [56] given a Lieb-Robinson bound:

Proposition B.2. *Assume that Φ_t satisfies the bound (LR2). Then, for any two quantum states $\rho, \sigma \in \mathcal{D}_{d^n}$ and any ordering $\{1, \ldots, n\}$ of the graph:*

$$W_{1,\square}(\Phi_t(\rho), \Phi_t(\sigma)) \leq \left(8 + 2\,c'\,(e^{vt}-1)\sum_{i=3}^{n} h(d(\{i\cdots n\}, \{1\}))\right) W_{1,\square}(\rho, \sigma)\,. \tag{41}$$

Moreover, for any observable $H \in \mathcal{M}_{d^n}$,

$$\|\Phi_t^*(H)\|_{\mathrm{Lip},\square} \leq \left(8 + 2\,c'\,(e^{vt}-1)\sum_{i=3}^{n} h(d(\{i\cdots n\}, \{1\}))\right) \|H\|_{\mathrm{Lip},\square}\,. \tag{42}$$

Proof. From [56], the Wasserstein distance $W_{1,\square}$ arises from a norm $\|.\|_{W_1}$, i.e. $W_{1,\square}(\rho,\sigma) = \|\rho-\sigma\|_{W_1}$. Moreover, the norm $\|.\|_{W_1}$ is uniquely determined by its unit ball $\mathcal{B}_n$, which in turn is the convex hull of the set of the differences between couples of neighboring quantum states:

$$\mathcal{N}_n = \bigcup_{i\in V} \mathcal{N}_n^{(i)}\,, \qquad \mathcal{N}_n^{(i)} = \{\rho-\sigma:\ \rho,\sigma\in\mathcal{D}_{d^n},\ \mathrm{tr}_i(\rho)=\mathrm{tr}_i(\sigma)\}\,.$$

Now by convexity, the contraction coefficient for this norm is equal to

$$\|\Phi_t\|_{W_1\to W_1} = \max\left(\|\Phi_t(X)\|_{W_1}:\ X\in\mathcal{M}_{d^n}^{\mathrm{sa},0},\ \|X\|_{W_1}\le 1\right) = \max_{X\in\mathcal{N}_n}\|\Phi_t(X)\|_{W_1}\,,$$

where $\mathcal{M}_{d^n}^{\mathrm{sa},0}$ denotes the set of self-adjoint, traceless observables. Let then $X\in\mathcal{N}_n$. By the expression (36), and choosing without loss of generality an ordering of the vertices such that $\mathrm{tr}_1(X)=0$, we have

$$\begin{aligned}
\|\Phi_t(X)\|_{W_1} &\le \frac{1}{2\sqrt{n}}\sum_{i=1}^n \left\| \frac{I}{d^{i-1}}\otimes \mathrm{tr}_{1\cdots i-1}\circ\Phi_t(X) - \frac{I}{d^i}\otimes\mathrm{tr}_{1\cdots i}\circ\Phi_t(X)\right\|_1 \\
&= \frac{1}{2\sqrt{n}}\sum_{i=1}^n \left\| \int d\mu(U_i)\ \mathrm{tr}_{1\cdots i-1}\circ(\Phi_t(X) - U_i\Phi_t(X)U_i^\dagger)\right\|_1 \\
&\le \frac{1}{2\sqrt{n}}\sum_{i=1}^n \int d\mu(U_i)\,\|[U_i,\mathrm{tr}_{1\cdots i-1}\circ\Phi_t(X)]\|_1 \\
&\le \frac{1}{\sqrt{n}}\sum_{i=1}^n \|\,\mathrm{tr}_{1\cdots i-1}\circ\Phi_t(X)\|_1 \\
&\overset{(1)}{=} \frac{1}{\sqrt{n}}\sum_{i=1}^n \|\,\mathrm{tr}_{1\cdots i-1}\circ(\Phi_t-\Phi_t^{(i-k\cdots n)})(X)\|_1
\end{aligned} \tag{43}$$

where μ denotes the Haar measure on one qudit, and where (1) follows from the fact that $\mathrm{tr}_1(X)=0$, with $\Phi_t^{(i-k\cdots n)}\equiv\Phi_t^{(\{i-k,\cdots,n\})}$ defined as in Eq. (LR2) with $k<i-1$. Next, by the variational formulation of the trace distance and Eq. (LR2), we have for $i\ge 3$ that

$$\begin{aligned}
\|\,\mathrm{tr}_{1\cdots i-1}\circ(\Phi_t-\Phi_t^{(i-k\cdots n)})(X)\|_1 &= \max_{\|O_{i\cdots n}\|_\infty\le 1}\left|\,\mathrm{tr}\left[X(\Phi_t^*-\Phi_t^{(i-k\cdots n)*})(O_{i\cdots n})\right]\right| \\
&\le \max_{\|O_{i\cdots n}\|_\infty\le 1}\|(\Phi_t^*-\Phi_t^{(i-k\cdots n)*})(O_{i\cdots n})\|_\infty\|X\|_1 \\
&\le c'\,(e^{vt}-1)\,h(\mathrm{dist}(\{i\cdots n\},\{1\cdots i-k-1\}))\,\|X\|_1 \\
&\overset{(2)}{\le} 2\,c'\,(e^{vt}-1)\,h(\mathrm{dist}(\{i\cdots n\},\{1\cdots i-k-1\}))\,\sqrt{n}\,\|X\|_{W_1}\,,
\end{aligned}$$

where (2) follows from [56, Proposition 6]. By picking $k=i-2$ and inserting this estimate into Eq. (43) for $i\ge 3$ and the trivial estimate $\|\,\mathrm{tr}_{1\cdots i-1}\circ(\Phi_t-\Phi_t^{(i-k\cdots n)})(X)\|_1\le 2\|X\|_1$ for $i=1,2$ we obtain Eq. (41). Eq. (42) follows by duality. □

B.3 Transportation cost inequalities

Although interesting on their own, the relevance of the Lipschitz constants introduced above becomes clearer in our context when we also have a *transportation cost* inequality [32, 58]. A quantum state σ satisfies a transportation cost inequality with constant $\alpha>0$ if for any other state ρ it holds that:

$$W_1(\rho,\sigma)\le\sqrt{\frac{1}{2\alpha}\,D(\rho\|\sigma)}\,, \tag{44}$$

where $W_1\in\{W_{1,\square},W_{1,\nabla},W_{1,\mathrm{loc}}\}$. In what follows, we simply write $\|.\|_{\mathrm{Lip}}$ and W_1 to denote either of the Lipschitz constants, and their corresponding Wasserstein metrics, defined above. This inequality should be thought of as a stronger version of Pinsker's inequality that is tailored to a state σ and the underlying Wasserstein distance.

One of the well-established techniques to establish a transportation cost inequality for $W_{1,\nabla}$ is by exploiting the fact that it is implied by a functional inequality called the modified logarithmic Sobolev inequality. It is

beyond the scope of this paper to explain this connection and we refer to e.g. [59] and references therein for a discussion on these topics. But for our purposes it is important to note that in [18] the authors and Capel show modified logarithmic Sobolev inequalities for several classes of Gibbs states. More recently, one of the authors and De Palma derived such transportation cost inequalities for $W_{1,\square}$ in [26]. In Theorem B.1 below we summarize the regimes for which transportation cost inequalities are known to hold:

Theorem B.1 (transportation cost for commuting Hamiltonians [8, 18, 26])**.** *Let $E_1, \dots, E_m \subset \mathcal{M}_{d^n}$ be a set of k-local linearly independent commuting operators with $\|E_i\|_\infty \leq 1$. Then $\sigma(\lambda)$ satisfies a transportation cost inequality with constant $\alpha > 0$ for all $\lambda \in B_{\ell_\infty}(0,1)$ in the following cases:*

(i) *The E_i are classical or nearest neighbour (i.e. $k = 2$) on a regular lattice and the inverse temperature $\beta < \beta$ where β_c only depends on k and the dimension of the lattice, for both $W_{1,\nabla}, W_{1,\square}$ and $\alpha = \Omega(1)$ [18].*

(ii) *The operators E_i are local with respect to a hypergraph $G = (V, E)$ and the inverse temperature satisfies $\beta < \beta_c$, where β_c only depends on k and properties of the interaction hypergraph for $W_{1,\square}$ and $\alpha = \Omega(1)$ [26, Theorem 3, Proposition 4].*

(iii) *The E_i are one-dimensional and $\beta > 0$, for both $W_{1,\nabla}$ and $W_{1,\square}$ and $\alpha = \Omega(\log(n)^{-1})$ [8].*

Moreover, the underlying differential structure (∇, σ) consists of L_k acting on at most $\mathcal{O}(k)$ qudits.

Theorem B.1 establishes that transportation cost inequalities are satisfied for most classes of commuting Hamiltonians known to have exponential decay of correlations.

Remark B.1. In [26, Proposition 5], the authors show that $W_{1,\square} \leq c(k)\, W_{1,\nabla}$ holds up to some constant $c(k)$ depending on the locality of the differential structure. This implies that a transportation cost inequality for $W_{1,\nabla}$ implies one for $W_{1,\square}$ up to $c(k)$. Thus, although the authors of [8, 18] only obtain the result for $W_{1,\nabla}$, we can use it to translate it to $W_{1,\square}$. We conclude that for commuting Hamiltonians TC are available for essentially all classes of local Hamiltonians for which they are expected to hold.

C Combining the maximum entropy method with transportation cost inequalities

With these tools at hand, we are now ready to show that by resorting to transportation cost inequalities it is possible to obtain exponential improvements in the sample complexity required to learn a Gibbs state. First, let us briefly review shadow tomography or Pauli regrouping techniques [22, 24, 40, 43]. Although these methods all work under slightly different assumptions and performance guarantees, they have in common that they allow us to learn the expectation value of M k-local observables $O_1, \dots, O_M \in \mathcal{M}_{2^n}$ such that $\|O_i\|_\infty \leq 1$ up to an error ϵ and failure probability at most $1 - \delta$ by measuring $\mathcal{O}(e^{\mathcal{O}(k)} \log(M\delta^{-1})\epsilon^{-2})$ copies of the state.

For instance, for the shadow methods of [40], we obtain a $\mathcal{O}(4^k \log(M\delta^{-1})\epsilon^{-2})$ scaling by measuring in 1-qubit random bases. The estimate is then obtained by an appropriate median-of-means procedure for the expectation value of each output string. The computation for obtaining the expectation value of E_i through this method then entails evaluating the expectation value of the observables on $\mathcal{O}(4^k \log(M\delta^{-1})\epsilon^{-2})$ product states. For k-local observables E_i, evaluating the expectation value of E_i on a product state takes time $\mathcal{O}(e^{ck} \log(M\delta^{-1})\epsilon^{-2})$ for some $c > 0$. Thus, we see that for $k = \mathcal{O}(\log(n))$ also the postprocessing can be done efficiently.

The application of such results to maximum entropy methods is then clear: given $E_1, \dots, E_m$ assumed to be at most k-local, with probability at least $1 - \delta$ we can obtain an estimate $e'(\lambda)$ of $e(\lambda)$ satisfying:

$$\|e'(\lambda) - e(\lambda)\|_{\ell_p} \leq m^{\frac{1}{p}} \epsilon \tag{45}$$

using $\mathcal{O}(4^k \log(m\delta^{-1})\epsilon^{-2})$ copies of $\sigma(\lambda)$. We then finally obtain our main theorem, Theorem 1.1, restated here for the sake of clarity:

Theorem C.1 (Fast learning of Gibbs states)**.** *Let $\sigma(\lambda) \in \mathcal{D}_{2^n}$ be an n-qubit Gibbs state at inverse temperature β with respect to a set of k-local operators $\mathcal{E} = \{E_i\}_{i=1}^m$ that satisfies a transportation cost inequality with constant α. Moreover, assume that $\mu \mapsto \log(\mathcal{Z}(\mu))$ is L, U strongly convex in $B_{\ell_\infty}(0,1)$. Then*

$$\mathcal{O}\left(4^k \alpha^{-1} \epsilon^{-2} \beta \log(m\delta^{-1}) \min\left\{\frac{m\beta}{Ln}, \frac{m^2}{\alpha\, n^2 \epsilon^2}\right\}\right)$$

samples of $\sigma(\lambda)$ are sufficient to obtain a state $\sigma(\mu)$ that satisfies:

$$|\operatorname{tr}[O(\sigma(\lambda) - \sigma(\mu))]| \leq \epsilon\, n^{\frac{1}{2}} \|O\|_{\mathrm{Lip}} \tag{46}$$

for all Lipschitz observables O with probability at least $1 - \delta$.

Proof. Using the aforementioned methods of [40] we can obtain an estimate $e'(\lambda)$ satisfying the guarantee of Eq. (29) with probability at least $1-\delta$. We will now resort to the results of Thm. A.1 to obtain guarantees on the output of the maximum entropy algorithm. Now we solve the maximum entropy problem with $e'(\lambda)$ and set the stopping criterion for the output μ_* as

$$\|e'(\lambda) - e(\mu_*)\|_{\ell_2} < (4c+1)\,\epsilon\sqrt{m}.$$

for $c > 10$. Then it follows from Thm. A.1 that:

$$D(\sigma(\mu_*)\|\sigma(\lambda)) \leq D(\sigma(\mu_*)\|\sigma(\lambda)) + D(\sigma(\lambda)\|\sigma(\mu_*)) = \mathcal{O}\left(\beta \min\{\beta L^{-1}\epsilon^2 m, \epsilon m\}\right). \tag{47}$$

This can then be combined with transportation cost inequalities. Indeed, we have:

$$|\operatorname{tr}\left[O(\sigma(\lambda) - \sigma(\mu_*)\right]| \leq \|O\|_{\mathrm{Lip}} W_1(\sigma(\mu_*), \sigma(\lambda)) \leq \|O\|_{\mathrm{Lip}} \sqrt{\frac{D(\sigma(\mu_*), \sigma(\lambda))}{2\alpha}}.$$

Inserting our bound on the relative entropy in Eq. (47) we obtain:

$$|\operatorname{tr}\left[O(\sigma(\lambda) - \sigma(\mu_*)\right]| = \mathcal{O}\left(\|O\|_{\mathrm{Lip}} \sqrt{\beta} \min\{(\alpha L)^{-\frac{1}{2}} \epsilon \sqrt{\beta\, m}, \sqrt{\epsilon m \alpha^{-1}}\}\right).$$

We conclude by suitably rescaling ϵ. □

The theorem above yields exponential improvements in the sample complexity to learn the expectation value of certain observables for classes of states that satisfy a transportation cost inequality with $\alpha = \Omega(\log(n)^{-1})$. As discussed in Sec. B.2, extensive observables that can be written as a sum of l-local observables have a Lipschitz constant that satisfies $\|O\|_{\mathrm{Lip}} = \mathcal{O}(l\sqrt{n})$. Shadow-like methods would require $\mathcal{O}(e^{\mathcal{O}(l)} \log(m\delta^{-1})\,\epsilon^{-2})$ samples to learn such observables up to a relative error of $n\epsilon$. Our methods, however, require $\mathcal{O}(e^{\mathcal{O}(k)} \operatorname{poly}(l, \epsilon^{-1}) \log(m\delta^{-1}))$, which yields exponential speedups in the regime $l = \operatorname{poly}(\log(n))$. Of course it should also be mentioned that classical shadows do not require any assumptions on the underlying state.

Furthermore, considering the exponential dependency of the sample complexity in the locality for shadow-like methods, we believe that our methods yield practically significant savings already in the regime in which we wish to obtain expectation values of observables with relatively small support. For instance, for high-temperature Gibbs state of nearest neighbor Hamiltonians and observables supported on 15 qubits, shadows require a factor of $\sim 10^7$ more samples than solving the maximum entropy problem for obtaining the same precision.

On the other hand, previous guarantees on learning quantum many-body states [10, 23, 53, 67] required a polynomial in system size precision to obtain a nontrivial approximation, which implies a polynomial-time complexity. Thus, our recovery guarantees are also exponentially better compared to standard many-body tomography results.

C.1 Results for shallow circuits

Let us be more explicit on how to leverage our results to also cover the outputs of short-depth circuits. To this end, let $G = (V, E)$ be a graph that models the interactions of a unitary circuit and suppose we implement L layers of an unknown unitary circuit consisting of 1 and 2-qubit unitaries laid out according to G. That is, we have an unknown shallow circuit $\mathcal{U}$ of depth L with respect to G. More precisely,

$$\mathcal{U} = \prod_{\ell \in [L]} \bigotimes_{e \in \mathcal{E}_\ell} \mathcal{U}_{\ell,e}, \tag{48}$$

where each $\mathcal{E}_\ell \subset E$ are a subset of the edges such that any $e, e' \in \mathcal{E}_\ell$ do not share a vertex. Our goal is to show how to approximately learn the state

$$|\psi\rangle = \mathcal{U}\,|0\rangle^{\otimes n}.$$

The overall idea consists in finding a Gibbs state approximating $|\psi\rangle$ in Wasserstein distance. We will then find a differential structure for (approximations) of shallow circuits and then showing that the (approximation of the) output satisfies a TC inequality with respect to it. Thus, it suffices to control the relative entropy with this approximation to ensure a good approximation in Wasserstein distance.

Let us find the appropriate approximation. First, note that for $\beta_\epsilon = \log(\epsilon^{-1})$ and $H_0 = -\sum_i^n Z_i$ where Z_i denotes the Pauli matrix $Z = |0\rangle\langle 0| - |1\rangle\langle 1|$ acting on site i, we have that:

$$\begin{aligned} D\Big(\mathcal{U}|0\rangle\langle 0|^{\otimes n}\mathcal{U}^\dagger \Big\| \mathcal{U}\frac{e^{-\beta_\epsilon H_0}}{\operatorname{tr}\left[e^{-\beta_\epsilon H_0}\right]}\mathcal{U}^\dagger\Big) &= D\Big(|0\rangle\langle 0|^{\otimes n}\Big\|\frac{e^{-\beta_\epsilon H_0}}{\operatorname{tr}\left[e^{-\beta_\epsilon H_0}\right]}\Big) \\ &= nD\Big(|0\rangle\langle 0|\Big\|\frac{e^{\beta_\epsilon Z}}{\operatorname{tr}\left[e^{\beta_\epsilon Z}\right]}\Big) \\ &= -n\beta_\epsilon \langle 0|Z|0\rangle + n\log\operatorname{tr}\left[e^{\beta_\epsilon Z}\right] \\ &= n\log\left(1+e^{-2\beta_\epsilon}\right) \\ &\leq ne^{-2\beta_\epsilon} \\ &= \epsilon^2 n. \end{aligned} \tag{49}$$

Thus, if the states $\mathcal{U}\frac{e^{-\beta_\epsilon H_0}}{\operatorname{tr}\left[e^{-\beta_\epsilon H_0}\right]}\mathcal{U}^\dagger$ satisfy a transportation cost inequality with some constant α, this would allow us to conclude that

$$W_1\left(\mathcal{U}|0\rangle\langle 0|^{\otimes n}\mathcal{U}^\dagger, \mathcal{U}\frac{e^{-\beta_\epsilon H_0}}{\operatorname{tr}\left[e^{-\beta_\epsilon H_0}\right]}\mathcal{U}^\dagger\right) \leq \epsilon\sqrt{\frac{n}{2\alpha}}.$$

Moreover, defining $H_\mathcal{U} = -\sum_i^n \mathcal{U}Z_i\mathcal{U}^\dagger$, we see that $\mathcal{U}\frac{e^{-\beta_\epsilon H_0}}{\operatorname{tr}\left[e^{-\beta_\epsilon H_0}\right]}\mathcal{U}^\dagger = \frac{e^{-\beta_\epsilon H_\mathcal{U}}}{\operatorname{tr}\left[e^{-\beta_\epsilon H_0}\right]}$. As we know the geometry of $\mathcal{U}$, we can bound the support of $\mathcal{U}Z_i\mathcal{U}^\dagger$. Thus, we only need to find a suitable transportation cost inequality to see that this approximation fits into our framework.

Let us now find a suitable differential structure for the state $\frac{e^{-\beta_\epsilon H_0}}{\operatorname{tr}\left[e^{-\beta_\epsilon H_0}\right]}$. For simplicity, denote by $\tau_p = p|0\rangle\langle 0| + (1-p)|1\rangle\langle 1|$ and note that $\frac{e^{-\beta_\epsilon H_0}}{\operatorname{tr}\left[e^{-\beta_\epsilon H_0}\right]} = \tau_{p_\epsilon}^{\otimes n}$ with $p_\epsilon = \frac{e^{\beta_\epsilon}}{e^{\beta_\epsilon}+e^{-\beta_\epsilon}}$. Moreover, let

$$a_i = I^{\otimes i-1} \otimes |0\rangle\langle 1| \otimes I^{\otimes n-i-1}$$

be the anihilation operator acting on qubit i. Defining $L_{i,0} = (p(1-p))^{\frac{1}{4}} a_i$ and $L_{i,1} = L_{i,0}^\dagger$, we get a differential structure for $\tau_p^{\otimes n}$ with $\{L_{i,k}, \omega_{i,k}\}$ for $i = 1,\ldots,n$, $k = 0,1$, $\omega_{i,0} = \frac{p}{1-p}$ and $\omega_{i,1} = \frac{1-p}{p}$. That this is indeed a differential structure follows from a simple computation. One can readily check that the induced Lipschitz constant is given by:

$$\begin{aligned} \|O\|_{\mathrm{Lip},\nabla}^2 &= \sum_{i=1}^n \left(\sqrt{\frac{p}{1-p}} + \sqrt{\frac{1-p}{p}}\right)\left(\|[O,L_{i,0}]\|^2 + \|[O,L_{i,1}]\|^2\right) \\ &= \sum_{i=1}^n \|[O,a_i]\|^2 + \|[O,a_i^\dagger]\|^2. \end{aligned}$$

Thus, we see that the Lipschitz constant takes a particularly simple form for this differential structure. Moreover, it is not difficult to see that $\{\mathcal{U}L_{i,k}\mathcal{U}^\dagger, \omega_{i,k}\}$ provides a differential structure for the state $\mathcal{U}\tau_p^{\otimes n}\mathcal{U}^\dagger$. Indeed, it is easily checked that this new differential structure still gives eigenvectors of the modular operator. Importantly, the result of [11, Theorem 19] establish that the state $\tau_p^{\otimes n}$ satisfies a transportation cost inequality with constant $\frac{1}{2}$. Putting all these elements together we have:

Theorem C.1 (transportation cost for shallow circuits)**.** *Let $\mathcal{U}$ be an unknown depth L quantum circuit on n qubits defined on a graph $G = (V,E)$ and $|\psi\rangle = \mathcal{U}|0\rangle^{\otimes n}$. Define for $\epsilon > 0$ $H_\mathcal{U} = -\sum_i \mathcal{U}Z_i\mathcal{U}^\dagger$ and $\sigma(\mathcal{U},\epsilon) = \frac{e^{-\beta_\epsilon H_\mathcal{U}}}{\operatorname{tr}\left[e^{-\beta_\epsilon H_\mathcal{U}}\right]}$ with $\beta_\epsilon = \log(\epsilon^{-1})$. Then for any state ρ and all observables O we have:*

$$|\operatorname{tr}\left[O(|\psi\rangle\langle\psi| - \rho)\right]| \leq \|O\|_{\mathrm{Lip},\nabla}\left(\epsilon\sqrt{n} + \sqrt{D(\rho\|\sigma(\mathcal{U},\epsilon))}\right)$$

with

$$\|O\|_{\mathrm{Lip},\nabla}^2 = \sum_{i=1}^n \|[O,\mathcal{U}a_i\mathcal{U}^\dagger]\|^2 + \|[O,\mathcal{U}a_i^\dagger\mathcal{U}^\dagger]\|^2, \quad a_i = I^{i-1} \otimes |0\rangle\langle 1| \otimes I^{n-i-1}$$

Proof. We have:

$$|\operatorname{tr}\left[O(|\psi\rangle\langle\psi| - \rho)\right]| \leq |\operatorname{tr}\left[O(|\psi\rangle\langle\psi| - \sigma(\mathcal{U}))\right]| + |\operatorname{tr}\left[O(\sigma(\mathcal{U}) - \rho)\right]|.$$

The claim then follows from the discussion above, as $\sigma(\mathcal{U})$ satisfies a transportation cost inequality with constant $\frac{1}{2}$. □

Of course the result above has the downside that the Lipschitz constant depends on the unknown circuit $\mathcal{U}$. However, as we can estimate the locality of each term $\mathcal{U}a_i\mathcal{U}^\dagger$, it is also possible to estimate the Lipschitz constant by controlling the overlap of the observable O with each $\mathcal{U}a_i\mathcal{U}^\dagger$.

The result of Thm. C.1 and the discussion preceding it also give us a method of efficiently learning the outputs of shallow circuits, as illustrated by the following proposition:

Proposition C.1 (Learning the outputs of shallow circuits)**.** *Let $\mathcal{U}$ be an unknown n-qubit quantum circuit with known locality structure as in Eq. (48) and $|\psi\rangle = \mathcal{U}|0\rangle^{\otimes n}$. Moreover, define l_0 as*

$$l_0 = \max_{1\leq i\leq n} \operatorname{supp}|\mathcal{U}Z_i\mathcal{U}^\dagger|.$$

For some $\epsilon > 0$ we have that

$$\mathcal{O}(\epsilon^{-2}4^{3l_0}\log^4(n4^{l_0}\epsilon^{-1})\log(4^{l_0}n\delta^{-1})) \tag{50}$$

samples of $|\psi\rangle$ suffice to learn a Gibbs state $\sigma(\mu)$ that satisfies

$$W_{1,\nabla}(|\psi\rangle\langle\psi|, \sigma(\mu)) \leq \sqrt{\epsilon n} \tag{51}$$

with probability of success at least $1-\delta$. Along similar lines we have that

$$\mathcal{O}(\epsilon^{-2}4^{3l_0}n^2\log^4(n4^{l_0}\epsilon^{-1})\log(4^{l_0}n\delta^{-1})) \tag{52}$$

samples suffice to learn a state $\sigma(\lambda)$ such that

$$\||\psi\rangle\langle\psi| - \sigma(\lambda)\|_{\mathrm{tr}} \leq \sqrt{\epsilon}.$$

Proof. Let $\mathcal{E}_i$ be a basis of Pauli operators for matrices on the support of $\mathcal{U}Z_i\mathcal{U}^\dagger$. By our assumption on l_0, we know that for each i we have that $|\mathcal{E}_i| \leq 4^{l_0}$. For simplicity we will assume that there are no Pauli words that are contained in two distinct $\mathcal{E}_i$ and we will enumerate all different Pauli words as $\{E_{i,j}\}$ for $1 \leq i \leq n$ and $1 \leq j \leq 4^{l_0}$ indicating the elements of the different $\mathcal{E}_i$. Thus, there is a $\lambda \in \mathbb{R}^m$ with $m \leq n4^{l_0}$ and $\|\lambda\|_{\ell_\infty} \leq 1$ such that

$$\sigma(\mathcal{U}, \epsilon) \propto \exp\left(-\beta_\epsilon \sum_{i=1}^{n}\sum_{E_{i,j}\in\mathcal{E}_i} \lambda_{i,j}E_{i,j}\right).$$

Let $\tilde{\epsilon} > 0$ be given. It follows from Eq. (49) and Pinsker's inequality that picking $\epsilon = \frac{\tilde{\epsilon}}{n16}$ is sufficient to ensure that

$$\||\psi\rangle\langle\psi| - \sigma(\mathcal{U}, \tfrac{\tilde{\epsilon}}{n4})\|_{\mathrm{tr}} \leq \frac{\tilde{\epsilon}}{2}. \tag{53}$$

Measuring

$$\mathcal{O}(4^{l_0}\tilde{\epsilon}^{-2}\log(4^{l_0}n\delta^{-1})) \tag{54}$$

copies of $|\psi\rangle$ is sufficient to obtain estimates of $\operatorname{tr}\left[|\psi\rangle\langle\psi|E_{i,j}\right]$ up to an $\tilde{\epsilon}/4$ error. By Eq. (53) and a triangle inequality they are also $\tilde{\epsilon}/2$ away from $\operatorname{tr}\left[\sigma(\mathcal{U}, \frac{\tilde{\epsilon}}{n4})E_{i,j}\right]$.

Thus, running the maximum entropy principle with these estimates and the basis of operators given by $\cup_{i=1}^n \mathcal{E}_i$ will yield us an estimate $\sigma(\mu)$ that satisfies:

$$D(\sigma(\mu)\|\sigma(\mathcal{U}, \tfrac{\tilde{\epsilon}}{n})) \leq \log\left(\frac{n}{4\tilde{\epsilon}}\right)\tilde{\epsilon}m \leq \log\left(\frac{n}{4\tilde{\epsilon}}\right)\tilde{\epsilon}4^{l_0}n.$$

To obtain the estimate in Wasserstein distance in Eq. (50), we can pick $\tilde{\epsilon} = \mathcal{O}(\epsilon/(4^{l_0}\log^2(n4^{l_0}\epsilon^{-1})))$, as in this case

$$\log\left(\frac{n}{4\tilde{\epsilon}}\right)\tilde{\epsilon}m \leq \log\left(\frac{n}{4\tilde{\epsilon}}\right)\tilde{\epsilon}4^{l_0}n = \epsilon\frac{\log\left(n\log^2(n4^{l_0}\epsilon^{-1})\epsilon^{-1}\right)}{\log^2(n4^{l_0}\epsilon^{-1})}n \leq \epsilon n.$$

The claim then follows from the results of Thm. C.1 and substituing $\tilde{\epsilon}$ into Eq. (54).

For the statement on the sample complexity for the trace distance, we may pick $\tilde{\epsilon} = \mathcal{O}(\epsilon/(n4^{l_0}\log^2(n4^{l_0}\epsilon^{-1})))$, which yields the statement after applying Pinsker's inequality. □

This shows that shallow circuits can be learned efficiently as long as $l_0 = \mathcal{O}(\log(n))$. However, it is not immediately obvious how to also estimate the expectation values of the Gibbs states required to run the maximum entropy algorithm. Thus, at least with the methods presented here, the postprocessing takes exponential time in the number of qubits.

Ground states of gapped systems: in light of our results for shallow circuits, it is natural to ask to what extent our framework can be extended to ground states of gapped Hamiltonians, specially in 1D. Thus, let us briefly comment on the technical barriers on the way of such statements. First, notice that the statement of Thm C.1 required inverse temperatures scaling logarithmically in system size for the approximation of the ground state. Most known TC inequalities have an exponential scaling with the inverse temperature and, thus, TC at this inverse temperature the savings compared to Pinsker are not quadratic, hindering a straightforward extension to gapped systems. There are some nontrivial examples of ground states satisfying a TC inequality with the constant depending inverse linearly with the temperature, like graph states [63]. But, as they can also be prepared from a constant depth quantum circuit almost by definition, they fall into the assumptions of the previous statement.

However, the results of [25] assert that k-local density matrices of ground states of gapped Hamiltonians in $1D$ can be approximated by constant depth circuits, giving evidence that our framework should also extend to such states. And to get there, a technical obstacle has to be overcome in the proof of Thm. C.1. Essentially we need to show that local reduced density matrices of the ground state are already well-approximated at inverse temperature $\log(\epsilon^{-1})$. With such a statement we could show that we can still approximate the expectation values of E_i at inverse temperature $\log(\epsilon^{-1})$ from samples from the ground state $|\psi\rangle$.

D Summary of known strong convexity constants

As we see in the statement of Thm. A.1 and C.1, having a bound on the strong convexity constant L^{-1} can give a quadratic improvement in the sample complexity in terms of the error ϵ. Here we will briefly summarize for which cases estimates on L are known in the literature for the classes of states we considered here.

General many-body quantum: first, we should mention the results of [5]. There the authors show bounds on L^{-1} for arbitrary many-body Hamiltonians and temperatures that scales linearly in m. Thus, although certainly nontrivial, these bounds do not improve the sample complexity for the regimes we are interested in this work, namely that of logarithmic sample complexity in system size. To obtain improvements in this regime, L^{-1} needs to be at most polylogarithmic in system size.

In the case of high-temperature Gibbs states, the recent work of [34] shows that this is indeed the case. I.e., in their Corollary 4.4 they show that for $\beta = \mathcal{O}(k^{-8})$, where k is the locality of the Hamiltonian, we indeed have $L^{-1} = \mathcal{O}(\beta^{-2})$. It should also be noted that their results do not hold only for geometrically local Hamiltonians, but rather any Hamiltonian such that each term acts on at most k qudits. This implies that for the high temperature regime, for which we also have the TC inequality in Thm. B.1, the improved sample complexity yielded by our methods holds. Note, however, that there is a slight mismatch between the inverse temperature range for which the two results hold: for the strong convexity we need $\beta = \mathcal{O}(k^{-8})$, whereas for TC $\beta = \mathcal{O}(k^{-1})$ suffices.

Commuting Hamiltonians: as we will prove later in Prop. F.1, in the case of commuting Hamiltonians we have that $L^{-1} = \mathcal{O}(e^{\beta}\beta^{-2})$. Thus, for any constant inverse temperature $\beta > 0$ we have an improved sample complexity. However, in order to analyse ground states in $1D$, our current proof techniques still require an inverse temperature scaling logarithmically in system size, so for such states we do not obtain improvements through strong convexity. We plan to address this gap in future work.

Besides the cases mentioned above, we also considered the case of local circuits in this work. For those there are no nontrivial estimates on L available to the best of our knowledge.

E Regimes of efficient postprocessing

The only question we have still to answer is how to perform the postprocessing efficiently, namely how the parameter $C_{\mathcal{E}}$ appearing in Theorem C.1 scales and how we obtain the bounds in Table 2.

There have been many recent breakthroughs on approximating quantum Gibbs states efficiently on classical computers [35, 46, 49, 51, 52]. The gradient descent method only requires us to estimate the gradient of the

partition function for a Gibbs state at each iteration. Thus, any algorithm that can approximate the log-partition function $\mathcal{Z}(\mu)$ efficiently or approximate $e(\lambda)$ suffices for our purposes.

For Gibbs states on a lattice, the methods of [51] yield that we can perform such approximations on a classical computer in time polynomial in n for temperatures $\beta < \beta_c = 1/(k8e^3)$, where k is the locality of the Hamiltonian, and ϵ inverse polynomial in system size. Thus, we conclude that for this temperature range, which coincides with the range for which the results of Thm. B.1 hold, $C_{\mathcal{E}}$ is polynomial in system size and we can also obtain efficient classical postprocessing.

For the case of Gibbs states of 1D systems, to the best of our knowledge, the best results available are those of [52]. They show how to efficiently obtain efficient tensor network approximations for 1D Gibbs states for $\beta = o(\log(n))$. As such tensor networks can be contracted efficiently as well, this gives an efficient classical algorithm to compute local expectation values of such states, which suffices for our purposes. Thus, the results of [51, 52] ensure that for the systems considered in Thm. B.1 we can also perform the postprocessing efficiently on a classical computer.

It is also interesting to consider what happens if we have access to a quantum computer to estimate the gradient. We will discuss the implications of this in the next section for the case of commuting Hamiltonians.

Finally, for local quantum circuits we are not aware at this stage of any method that could yield a better postprocessing complexity than computing the partition function explicitly. This would yield a postprocessing that is exponential in the system size, as it requires diagonalizing an exponentially large matrix.

F A complete picture: commuting Gibbs states

In this section, we discuss two classes of states for which the strongest version of our theorems holds, namely that of commuting 1D Gibbs states, and the one of high-temperature commuting Gibbs states on arbitrary hypergraphs. We already discussed that they satisfy transportation cost inequalities in Thm. B.1. Thus, the last missing ingredient to obtain an optimal performance is to show that the partition function is indeed strongly convex. More precisely, we will now establish that, for these classes of states, both the upper and lower bounds on the Hessian of the log partition function are order 1. In addition to that, with access to a quantum computer, it is possible to perform the post-processing in time $\tilde{\mathcal{O}}(m)$. As writing down the vector λ takes $\Omega(m)$ time, we conclude that the postprocessing can be performed in a time comparable to writing down a solution to the problem. Thus, our procedure is essentially optimal.

Also in the setting of commuting Gibbs states, it is worth noting that after the completion of this work, we became aware of [4], which gives another method to learn the Gibbs state and its Hamiltonian that neither involves the maximum entropy method nor requires strong convexity guarantees. Their algorithm works by learning local reduced density matrices and showing that the parameters λ of the Hamiltonian of a commuting Gibbs state can be efficiently estimated from that. In principle, obtaining a bound on λ also suffices for our purposes and we could alternatively use their methods to bypass having to solve the maximum entropy problem for such states. In particular, this means that the postprocessing with their methods could be performed even for temperatures at which the partition function cannot be estimated efficiently but we still have access to samples from the state. However, as we ultimately are interested in the regime in which TC inequalities hold, which corresponds to the high-temperature regime, we do not further comment on their results.

In this section, we consider a hypergraph $G = (V, E)$ and assume that there is a maximum radius $r_0 \in \mathbb{N}$ such that, for any hyperedge $A \in E$, there exists a vertex $v \in V$ such that the ball $B(v, r_0)$ centered at v and of radius r_0 includes A. In what follows, we also denote by $S(v, r)$ the sphere of radius r centered at vertex $v \in V$, and define for all $r \in \mathbb{N}$:

$$B(r) := \max_{v \in V} |B(v, r)|\,, \qquad S(r) := \max_{v \in V} |S(v, r)|\,.$$

Next, we consider a Gibbs state $\sigma(\lambda)$ on $\mathcal{H}_V := \bigotimes_{v \in V} \mathcal{H}_v$, where $\dim(\mathcal{H}_v) = d$ for all $v \in V$. More precisely, $\sigma(\lambda) := e^{-\beta H(\lambda)}/\mathcal{Z}(\lambda)$, where with a slight abuse of notations

$$H(\lambda) = \sum_{i=1}^{m} \lambda_i E_i = \sum_{A \in E} \sum_{\alpha_A \in [d^2]^{|A|}} \alpha_A \, E_{\alpha_A}, \tag{55}$$

with $\|E_i\|_\infty = 1$ for all $i \in \{1, \cdots, m\}$, $[E_i, E_j] = 0$ for all $i, j \in \{1, \cdots, m\}$, and where the sets $\{1, \cdots, m\}$ and $\{\alpha_A\}_{A \in E, \alpha_A \in [d^2]^A}$ are in bijection. Note that $B(r_0)$ bounds the maximal locality of the Hamiltonian.

We also denote by $\sigma_A(\lambda)$ the Gibbs state corresponding to the restriction of H onto the region A, i.e.

$$\sigma_A(\lambda) := \frac{e^{-\beta H_A(\lambda)}}{\operatorname{tr}\left[e^{-\beta H_A(\lambda)}\right]}\,, \qquad \text{where} \qquad H_A(\lambda) := \sum_{i \,|\, \operatorname{supp}(E_i) \cap A \neq 0} \lambda_i E_i\,.$$

Note that, in general, $\sigma_A(\lambda) \neq \mathrm{tr}_{A^c}(\sigma(\lambda))$.

F.1 Upper bound on Hessian for commuting Gibbs states with decay of correlations

In this section, we prove tightened strong convexity constants for the log partition function in the case when the Gibbs state arises from a local commuting Hamiltonian at high-temperature. In fact, the upper constant found in Proposition A.4 can be tightened into a system size-independent one under the condition of exponential decay of the correlations in the Gibbs state. Several notions of exponential decay of correlations exist in the literature [18]. Here, we will say that a Gibbs state σ has correlation length ξ if for all observables O_A, O_B supported on non-overlapping regions A and B respectively, we have that:

$$|\operatorname{tr}[\sigma\, O_A \otimes O_B] - \operatorname{tr}[\sigma\, O_A] \operatorname{tr}[\sigma\, O_B]| \leq c\, \|O_A\|_\infty \|O_B\|_\infty e^{-\xi \operatorname{dist}(A,B)}$$

for some constant $c > 0$. In the classical setting, this condition is known to hold at any inverse temperature for 1D systems, and below a critical inverse temperature β_c that depends on the locality of the Hamiltonian when $D \geq 2$ [28]. The same bound also holds in the quantum case for 1D systems [6], or above a critical temperature on regular lattices when $D \geq 2$ [35, 46]. Using these bounds, we obtain the following improvement of Proposition A.4 which shows that for this class of states $U = \mathcal{O}(1)$.

Lemma F.1 (Strengthened upper strong convexity at high-temperature and for 1D systems)**.** *For each $\mu \in B_{\ell_\infty}(0,1)$, let $\sigma(\mu)$ be a Gibbs state at inverse temperature $\beta < \beta_c$ corresponding to the Hamiltonian defined on the hypergraph $G = (V, E)$ in (55). Then*

$$\nabla^2 \log(\mathcal{Z}(\mu)) \leq \left(c\,\beta^2\, B(r_0)\, B(2r_0)\, d^{2B(r_0)} \sum_{r=1}^{\infty} e^{-\xi r} S(r)\right) I\,,$$

where ξ is the correlation length of the state. Moreover, this result holds for all $\beta > 0$ in 1D.

Proof. Let us first use the assumption of commutativity to simplify the expression for the Hessian. We find for all $\alpha_A \in [d^2]^{|A|}, \alpha'_B \in [d^2]^{|B|}$, $A, B \in E$,

$$(\nabla^2 \log(\mathcal{Z}(\mu)))_{\alpha_A \alpha'_B} = \beta^2 \operatorname{tr}\left[\sigma(\mu)\,(E_{\alpha_A} - \operatorname{tr}[\sigma(\mu)\, E_{\alpha_A}])\,(E_{\alpha'_B} - \operatorname{tr}\left[\sigma(\mu)\, E_{\alpha'_B}\right])\right]. \tag{56}$$

The rest follows similarly to Proposition A.4 from Gershgorin's circle theorem together with the decay of correlations arising at $\beta < \beta_c$: for all α_A,

$$\sum_{\alpha'_B \neq \alpha_A} \left|\left(\nabla^2 \log(\mathcal{Z}(\mu))\right)_{\alpha_A \alpha'_B}\right| \leq c\,\beta^2 \sum_{\alpha'_B \neq \alpha_A} e^{-\xi \operatorname{dist}(A,B)}\,,$$

where we also used that the basis operators $\{E_i\}$ have operator norm 1. The claim then follows by bounding the number of basis operators whose support is at a distance r of A for each $r \in \mathbb{N}$: the latter is bounded by the product of (i) the number of vertices in A; (ii) the number of vertices at a distance r of a given vertex; (iii) the number of hyperedges containing a given vertex; and (iv) the number of basis operators corresponding to a given hyperedge. A simple estimate of each of these quantities gives the bound $B(r_0)\, S(r)\, B(2r_0)\, d^{2B(r_0)}$. Therefore:

$$\sum_{\alpha'_B \neq \alpha_A} \left|\left(\nabla^2 \log(\mathcal{Z}(\mu))\right)_{\alpha_A \alpha'_B}\right| \leq c\,\beta^2\, B(r_0)\, B(2r_0)\, d^{2B(r_0)} \sum_{r=1}^{\infty} e^{-\xi r} S(r)\,.$$

□

Note that for D-regular graphs and $r_0 = \mathcal{O}(1)$ we have $B(r_0)\, B(2r_0)\, d^{2B(r_0)} = \mathcal{O}(1)$ and $S(r) = \mathcal{O}(r^{D-1})$, giving a scaling of $\nabla^2 \log(\mathcal{Z}(\mu)) = \mathcal{O}(\beta^2 \xi^D)$.

F.2 Lower bound on Hessian for commuting Gibbs states

Whenever the Gibbs state is assumed to be commuting, the lower strong convexity constant L can also be made independent of m, by a direct generalization of the classical argument as found in [66][Lemma 7] or [54] (see also [5]).

Before we state and prove our result, let us introduce a few useful concepts: given a full-rank quantum state σ, we denote the weighted 2-norm of an observable Y as

$$\|Y\|_{2,\sigma} := \operatorname{tr}\left[\left|\sigma^{\frac{1}{4}} Y \sigma^{\frac{1}{4}}\right|^2\right]^{1/2} .$$

and refer to the corresponding non-commutative weighted L_2 space with inner product $\langle X, Y\rangle_\sigma := \operatorname{tr}\left[\sigma^{\frac{1}{2}} X^\dagger \sigma^{\frac{1}{2}} Y\right]$ as $L_2(\sigma)$. The Petz recovery map corresponding to a subregion $A \subset V$ with respect to $\sigma(\mu)$ is defined as

$$\mathcal{R}_{A,\sigma(\mu)}(X) = \operatorname{tr}_A(\sigma(\mu))^{-1/2} \operatorname{tr}_A(\sigma(\mu)^{1/2} X \sigma(\mu)^{1/2}) \operatorname{tr}_A(\sigma(\mu))^{-1/2} ,$$

We will also need the notion of a conditional expectation $\mathbb{E}_A$ with respect to the state $\sigma(\mu)$ into the region $A \subset \Lambda$ (see [7, 44] for more details). For instance, one can choose $\mathbb{E}_A := \lim_{n\to\infty} \mathcal{R}^n_{A,\sigma(\mu)}$, where $\mathcal{R}_{A,\sigma(\mu)}$ is the Petz recovery map of $\sigma(\mu)$. In other words, the map $\mathbb{E}_A$ is a completely positive, unital map that projects onto the algebra $\mathcal{N}_A$ of fixed points of $\mathcal{R}_{A,\sigma(\mu)}$. This algebra is known to be expressed as the commutant [7]

$$\mathcal{N}_A := \{\sigma(\mu)^{it} \mathcal{B}(\mathcal{H}_A) \sigma(\mu)^{-it};\, t \in \mathbb{R}\}' .$$

Moreover, the maps $\mathbb{E}_A$ commute with the modular operator $\Delta_{\sigma(\mu)}(.) := \sigma(\mu)(.)\sigma(\mu)^{-1}$, and for any $X_A, Z_A \in \mathcal{N}_A$ and all $Y \in \mathcal{B}(\mathcal{H}_V)$, $\mathbb{E}_A[X_A Y Z_A] = X_A \mathbb{E}_A[Y] Z_A$. The commutativity condition for $H(\mu)$ implies frustration freeness of the family of conditional expectations $\{\mathbb{E}_A\}_{A\in E}$: for any $X \in L_2(\sigma(\mu))$, $\|\mathbb{E}_A(X)\|^2_{2,\sigma(\mu)} + \|(\mathrm{id} - \mathbb{E}_A)(X)\|^2_{2,\sigma(\mu)} = \|X\|^2_{L_2(\sigma(\mu))}$.

The next technical lemma is essential in the derivation of our strong convexity lower bound. With a slight abuse of notations, we use the simplified notations $\sigma_x(\lambda) := \sigma_{\{x\}}(\lambda)$, $H_x(\lambda) := H_{\{x\}}(\lambda)$ and so on.

Lemma F.2. *Let $H = \sum_j \mu_j E_j$ be a local commuting Hamiltonian on the hypergraph $G = (V, E)$ defined in (55), each local operator E_j is further assumed to be traceless. The following holds for any $x \in V$:*

$$c(x,\beta) := \max_{\mu \in B_{\ell_\infty}(0,1)} \frac{\|\mathcal{R}_{x,\sigma_x(\mu)}(H_x) - \operatorname{tr}[\sigma_x(\mu) H_x]\, I\|_{2,\sigma_x(\mu)}}{\|H_x - \operatorname{tr}[\sigma_x(\mu) H_x]\, I\|_{2,\sigma_x(\mu)}} < 1 .$$

Proof. We first prove that $X = \mathcal{R}_{x,\sigma_x(\mu)}(X)$ is equivalent to $\|\mathcal{R}_{x,\sigma_x(\mu)}(X)\|_{2,\sigma_x(\mu)} = \|X\|_{2,\sigma_x(\mu)}$. One direction trivially holds. For the opposite direction, assume that $X \neq \mathcal{R}_{x,\sigma_x(\mu)}(X)$. This means that $X = Y + Z$, with Y, Z two operators that are orthogonal in $L_2(\sigma_x(\mu))$, with $\mathcal{R}_{x,\sigma_x(\mu)}(Y) = Y$ and $Z \neq 0$. Now, since $\mathcal{R}_{x,\sigma_x(\mu)}$ is self-adjoint and unital, it strictly contracts elements in the orthogonal of its fixed points and we have

$$\begin{aligned}
\|\mathcal{R}_{x,\sigma_x(\mu)}(X)\|^2_{2,\sigma_x(\mu)} &= \|Y\|^2_{2,\sigma_x(\mu)} + \|\mathcal{R}_{x,\sigma_x(\mu)}(Z)\|^2_{2,\sigma_x(\mu)} \\
&< \|Y\|^2_{2,\sigma_x(\mu)} + \|Z\|^2_{2,\sigma_x(\mu)} \\
&= \|X\|^2_{2,\sigma_x(\mu)} ,
\end{aligned}$$

which contradicts the condition of equality of the norms. Now, since the map $\mathcal{R}_{x,\sigma_x(\mu)}$ is unital, it suffices to prove that $\mathcal{R}_{x,\sigma_x(\mu)}(H_x) \neq H_x$, or equivalently $\mathbb{E}_x(H_x) \neq H_x$, in order to conclude. Let us assume instead that equality holds. This means that, for all observables A_x supported on site x, and all $t \in \mathbb{R}$:

$$[H_x, \sigma(\mu)^{it} A_x \sigma(\mu)^{-it}] = \sigma(\mu)^{it} [H_x, A_x] \sigma(\mu)^{-it} = 0 \qquad \Rightarrow \qquad H_x = \frac{I_x}{d} \otimes \operatorname{tr}_x(H_x) .$$

However, this contradicts the fact that H_x is traceless on site x. Therefore $\mathcal{R}_{x,\sigma_x(\mu)}(H_x) \neq H_x$ and the proof follows. □

We are ready to prove our strong convexity lower bound.

Proposition F.1 (Strengthened lower strong convexity constant for commuting Hamiltonians)**.** *For each $\mu \in B_{\ell_\infty}(0,1)$, let $\sigma(\mu)$ be a Gibbs state at inverse temperature β corresponding to the commuting Hamiltonian $H(\mu) = \sum_j \mu_j E_j$ on the hypergraph $G = (V, E)$ defined in (55), where $\operatorname{tr}[E_i E_j] = 0$ for all $i \neq j$ and each local operator E_j is traceless on its support. Then the Hessian of the log-partition function is lower bounded by*

$$\nabla^2 \log \mathcal{Z}(\mu) \geq \beta^2 e^{-\beta(B(2r_0)+2B(4r_0))} d^{-B(2r_0)} (1 - c(\beta)^2)\, I ,$$

where $c(\beta) := \max_{v\in V} c(v,\beta)$.

Proof. We first use the assumption of commutativity in order to simplify the expression for the Hessian. As in (56), we find

$$(\nabla^2 \log(\mathcal{Z}(\mu)))_{ij} = \beta^2 \operatorname{tr}\left[\sigma(\mu)\left(E_i - \operatorname{tr}\left[\sigma(\mu)\, E_i\right]\right)\left(E_j - \operatorname{tr}\left[\sigma(\mu)\, E_j\right]\right)\right].$$

Therefore, for any linear combination $H \equiv H(\lambda) = \sum_j \lambda_j E_j$ of the basis vectors, we have

$$\sum_{ij} \lambda_i \lambda_j \left(\nabla^2 \log(\mathcal{Z}(\mu))\right)_{ij} = \beta^2 \operatorname{Var}_{\sigma(\mu)}(H),$$

with $\operatorname{Var}_{\sigma(\mu)}(X) := \|X - \operatorname{tr}\left[\sigma(\mu)X\right]\|^2_{2,\sigma(\mu)}$. It is thus sufficient to lower-bound the latter. For this, we choose a subregion $A \subseteq V$ such that any basis element E_i has support intersecting a unique vertex in A. We lower-bound the variance by

$$\|H - \operatorname{tr}\left[\sigma(\mu)H\right] I\|^2_{2,\sigma(\mu)} \geq \|(\mathrm{id} - \mathbb{E}_A)(H - \operatorname{tr}\left[\sigma(\mu)H\right] I)\|^2_{2,\sigma(\mu)} = \|H - \mathbb{E}_A[H]\|^2_{2,\sigma(\mu)}, \tag{57}$$

where the first inequality follows by the $L_2(\sigma(\mu))$ contractivity of $\mathrm{id} - \mathbb{E}_A$. Now, the weighted norm can be further simplified into a sum of local weighted norms as follows: first, for any two E_i, E_j whose supports intersect with a different vertex of A, we have

$$\begin{aligned}
\langle \mathbb{E}_A[E_i],\, \mathbb{E}_A[E_j] \rangle_{\sigma(\mu)} &= \operatorname{tr}\left[\sigma(\mu)^{1/2}\mathbb{E}_A[E_i]\sigma(\mu)^{1/2}\mathbb{E}_A[E_j]\right] \\
&= \operatorname{tr}\left[\sigma(\mu)\Delta^{-1/2}_{\sigma(\mu)} \circ \mathbb{E}_A[E_i]\mathbb{E}_A[E_j]\right] \\
&= \operatorname{tr}\left[\sigma(\mu)\mathbb{E}_A[E_i]\mathbb{E}_A[E_j]\right]
\end{aligned} \tag{58}$$

where in the third line we used the commutation of the modular operator $\Delta_{\sigma(\mu)}(X) := \sigma(\mu)X\sigma(\mu)^{-1}$ with $\mathbb{E}_A$ together with the commutativity of $\sigma(\mu)$ and E_i. Now, denoting $\operatorname{supp}(E_i) \cap A = \{x\}$ and $\operatorname{supp}(E_j) \cap A = \{y\}$, we show that

$$\mathbb{E}_A[E_i] = \mathbb{E}_x[E_i] \qquad \text{and} \qquad \mathbb{E}_A[E_j] = \mathbb{E}_y[E_j]. \tag{59}$$

In order to prove these two identities, we simply need to prove for instance that $\mathbb{E}_x[E_i]$ belongs to the image algebra $\mathcal{N}_A$ of $\mathbb{E}_A$ since $\mathcal{N}_A \subseteq \mathcal{N}_x$ by definition. Hence, it is enough to show that $\mathbb{E}_x[E_i]$ commutes with operators of the form $\sigma(\mu)^{it}X_A\sigma(\mu)^{-it}$ for any $t \in \mathbb{R}$ and any $X_A \in \mathcal{B}(\mathcal{H}_A)$. This claim follows from

$$\begin{aligned}
\mathbb{E}_x[E_i]\sigma(\mu)^{it}X_A\sigma(\mu)^{-it} &= \sigma(\mu)^{it}\sigma(\mu)^{-it}\mathbb{E}_x[E_i]\sigma(\mu)^{it}X_A\sigma(\mu)^{-it} \\
&= \sigma(\mu)^{it}\mathbb{E}_x[\sigma(\mu)^{-it}E_i\sigma(\mu)^{it}]X_A\sigma(\mu)^{-it} \\
&= \sigma(\mu)^{it}\mathbb{E}_x[E_i]X_A\sigma(\mu)^{-it} \\
&= \sigma(\mu)^{it}X_A\mathbb{E}_x[E_i]\sigma(\mu)^{-it} \\
&= \sigma(\mu)^{it}X_A\sigma(\mu)^{-it}\mathbb{E}_x[E_i].
\end{aligned}$$

where the fourth line follows from the fact that the support of $\mathbb{E}_x[E_i]$ does not intersect $A\backslash\{i\}$, together with the fact that $\mathbb{E}_x[E_i]$ is locally proportional to I on site x, by definition of $\mathcal{N}_x$. Therefore, using (59) into (58), we get

$$\begin{aligned}
\langle \mathbb{E}_A[E_i],\, \mathbb{E}_A[E_j] \rangle_{\sigma(\mu)} &= \operatorname{tr}\left[\sigma(\mu)\mathbb{E}_x[E_i]\mathbb{E}_y[E_j]\right] \\
&= \operatorname{tr}\left[\sigma(\mu)\mathbb{E}_x\mathbb{E}_y[E_iE_j]\right] \\
&= \operatorname{tr}\left[\sigma(\mu)E_iE_j\right] \\
&= \langle E_i,\, E_j \rangle_{\sigma(\mu)},
\end{aligned}$$

where in the second line we used that $\mathbb{E}_x[E_i]$ is a fixed point of $\mathbb{E}_y$, and then that E_j is a fixed point of $\mathbb{E}_x$, by the support conditions of E_i and E_j. Therefore, the variance on the right-hand side of (57) can be simplified as

$$\begin{aligned}
\|H - \mathbb{E}_A[H]\|^2_{2,\sigma(\mu)} &= \sum_{x\in A} \Big\| \sum_{j|\ \operatorname{supp}(E_j)\ni x} \lambda_j \left(E_j - \mathbb{E}_x[E_j]\right)\Big\|^2_{2,\sigma(\mu)} \\
&= \sum_{x\in A} \left\|H_x - \mathbb{E}_x[H_x]\right\|^2_{2,\sigma(\mu)}
\end{aligned}$$

where we recall that $H_x := \sum_{j|\ \mathrm{supp}(E_j)\ni x} \lambda_j E_j$. Now, for any $x \in V$, we denote $x\partial := \{A \in E | \, x \in A\}$ and decompose the Hamiltonian $H(\mu)$ as

$$H(\mu) = H_x(\mu) + K_x^0(\mu) + K_x^1(\mu)\,, \qquad \text{where} \qquad \begin{cases} \mathrm{supp}(K_x^1(\mu)) \cap x\partial = \emptyset \\ \mathrm{supp}(K_x^0(\mu)) \cap x\partial \neq \emptyset \end{cases}. \tag{60}$$

Clearly,

$$\sigma(\mu) \geq \sigma_x(\mu)\, e^{-2\beta\, \|K_x^0(\mu)\|_\infty}\, \frac{e^{-\beta K_x^1(\mu)}}{\mathrm{tr}_{x\partial^c}\left[e^{-\beta K_x^1(\mu)}\right]}\,. \tag{61}$$

Now,

$$\begin{aligned}
\|H_x - \mathbb{E}_x[H_x]\|_{2,\sigma(\mu)}^2 &= \mathrm{tr}\left[\sigma(\mu)(H_x - \mathbb{E}_x[H_x])^2\right] \\
&\geq e^{-2\beta\, \|K_x^0(\mu)\|_\infty}\ \mathrm{tr}\left[\sigma_x(\mu)\, \frac{e^{-\beta K_x^1(\mu)}}{\mathrm{tr}_{x\partial^c}\left[e^{-\beta K_x^1(\mu)}\right]} (H_x - \mathbb{E}_x[H_x])^2\right] \\
&= e^{-2\beta\, \|K_x^0(\mu)\|_\infty}\, \|H_x - \mathbb{E}_x[H_x]\|_{2,\sigma_x(\mu)}^2 \\
&= e^{-2\beta\, \|K_x^0(\mu)\|_\infty} \left(\|H_x - \mathrm{tr}\left[\sigma_x(\mu) H_x\right] I\|_{2,\sigma_x(\mu)}^2 - \|\mathbb{E}_x[H_x] - \mathrm{tr}\left[\sigma_x(\mu) H_x\right] I\|_{2,\sigma_x(\mu)}^2\right) \\
&\geq e^{-2\beta\, \|K_x^0(\mu)\|_\infty} \left(\|H_x - \mathrm{tr}\left[\sigma_x(\mu) H_x\right] I\|_{2,\sigma_x(\mu)}^2 - \|\mathcal{R}_{x,\sigma(\mu)}[H_x] - \mathrm{tr}\left[\sigma_x(\mu) H_x\right] I\|_{2,\sigma_x(\mu)}^2\right) \\
&\geq (1 - c(\beta)^2)\, e^{-2\beta\, \|K_x^0(\mu)\|_\infty}\, \|H_x - \mathrm{tr}\left[\sigma_x(\mu) H_x\right] I\|_{2,\sigma_x(\mu)}^2\,.
\end{aligned}$$

The first and second identities above follow once again from the commutativity of the Hamiltonian similarly to (58), where for the second one we also use the disjointness of $x\partial$ and $\mathrm{supp}(K_x^1(\mu))$. The first inequality follows from (61). The third identity is a consequence of the fact that $\mathbb{E}_x$ is a projection with respect to $L_2(\sigma_x(\mu))$. The second inequality follows from $\|\mathcal{R}_{x,\sigma(\mu)}[X]\|_{2,\sigma(\mu)} \geq \|\mathbb{E}_x[X]\|_{2,\sigma(\mu)}$ for all X (see Proposition 10 in [44]). The last inequality is a consequence of Lemma F.2. Finally, we further bound the weighted L_2 norm on the last line of the above inequality in terms of the Schatten 2 norm to get

$$\mathrm{Var}_{\sigma(\mu)}(H) \geq (1 - c(\beta)^2) \min_{x\in V} e^{-2\beta \|K_x^0(\mu)\|_\infty}\, \lambda_{\min}(\sigma_x(\mu)) \sum_{x\in A} \|H_x - \mathrm{tr}\left[\sigma_x(\mu)\, H_x\right]\|_2^2 \tag{62}$$

$$\geq (1 - c(\beta)^2) \min_{x\in V} e^{-2\beta \|K_x^0(\mu)\|_\infty}\, \lambda_{\min}(\sigma_x(\mu)) \sum_i \lambda_i^2\,, \tag{63}$$

where $\lambda_{\min}(\sigma_x(\mu))$ denotes the minimum eigenvalue of $\sigma_x(\mu)$. The result follows from the simple estimates $\|K_x^0(\mu)\|_\infty \leq B(4r_0)$ and $\lambda_{\min}(\sigma_x(\mu)) \geq e^{-\beta B(2r_0)} d^{-B(2r_0)}$.

□

F.3 Summary for 1D or high-temperature commuting

Now we genuinely have all the elements in place to essentially give the last word on estimating Lipschitz observables for Gibbs states of nearest neighbor 1D or high-temperature commuting Gibbs states.

Theorem F.1. *For each $\mu \in B_{\ell_\infty}(0,1)$, let $\sigma(\mu)$ be a Gibbs state at inverse temperature β corresponding to the commuting Hamiltonian $H(\mu) = \sum_{j=1}^m \mu_j E_j$ on the hypergraph $G = (V, E)$ defined in (55), where $\mathrm{tr}\left[E_i E_j\right] = 0$ for all $i \neq j$, each local operator E_j is traceless on its support and acts on a constant number of qubits and $m = \mathcal{O}(n)$. Moreover, assume that $\sigma(\lambda)$ satisfies the conditions of Thm. B.1. Then $\mathcal{O}(\log(n)\epsilon^{-2})$ samples of $\sigma(\lambda)$ suffice to obtain a $\mu \in B_{\ell_\infty}(0,1)$ satisfying*

$$\mathrm{tr}\left[O(\sigma(\lambda) - \sigma(\mu))\right] = \mathcal{O}(\epsilon\sqrt{n}\|O\|_{\mathrm{Lip},\nabla}).$$

with probability at least $1 - p$. Moreover, we can find μ in $\mathcal{O}(poly(n, \epsilon^{-1}))$ time on a classical computer. With access to a quantum computer, the postprocessing can be performed in $\tilde{\mathcal{O}}(n\epsilon^{-2})$ time by only implementing quantum circuits of $\tilde{\mathcal{O}}(1)$ depth.

Proof. To obtain the claim on the sample complexity, note that for such systems $L = \Omega(1)$ by Prop. F.1 and they satisfy a transportation cost inequality by Thm. B.1. Moreover, we can learn the expectation of all E_i up to an error ϵ and failure probability δ with $\mathcal{O}(\log(n\delta^{-1})\epsilon^{-2})$ samples using shadows, as they all have constant

support. The claimed sample complexity then follows from Thm. C.1. The classical postprocessing result follows from [51]. The postprocessing with a quantum computer follows from the results of [44], [18] combined with the fact that $L^{-1}U = \mathcal{O}(1)$ by also invoking Lemma F.1. Indeed, [18, 44] asserts that we can approximately prepare any $\sigma(\mu)$ on a quantum computer with a circuit of depth $\tilde{\mathcal{O}}(1)$. Moreover, once again resorting to shadows we can estimate $e(\mu)$ with $\tilde{\mathcal{O}}(\epsilon^{-2})$ samples. We conclude that we can run each iteration of gradient descent in $\tilde{\mathcal{O}}(n\epsilon^{-2})$ time. As $L^{-1}U = \mathcal{O}(1)$, Thm. A.1 asserts that we converge after $\tilde{\mathcal{O}}(1)$ iterations, which yields the claim □

The theorem above nicely showcases how the joint use of transportation cost inequalities and strong convexity bounds to improve maximum entropy methods come together. Moreover, with access to a quantum computer, up to polylogarithmic overhead in system size factors, the computational complexity of learning the Gibbs state is comparable to reading out the results of measurements on the system. Together with the polylog sample complexity bounds that we obtain, this justifies our claiming that the result above almost gives the last word on learning such states.

G Comparison of sample complexity of previous methods

Our work arguably introduces two technical innovations to the literature of learning and tomography of quantum states that underly our exponential speedups in sample complexity. The first is the observation that most observables of physical interest have a small Lipschitz constant, and, thus, it might be more motivated to look for good approximations in Wasserstein distance instead of trace distance. The second is that for states that satisfy a TC inequality, it suffices to obtain an estimate of the state that has a small relative entropy density with the target state to recover Lipschitz observables. And that finding such an estimate can be achieved from a few samples through a combination of maximum entropy and classical shadow methods.

We will now argue that these two innovations are indeed crucial to ensure our exponential speedups. First, we will show that the shadow protocol will yield bad estimates for the expectation value of local observables with high probability even for product states if the number of samples is not exponential in the locality of the underlying observables. This shows that exploiting the locality of the underlying states is crucial to obtaining a polynomial sample complexity in the locality of the observables. After that, we will show in Subsec. G.2 that $\Omega(\sqrt{n}\epsilon^{-2})$ samples are necessary for any algorithm that can recover any Gibbs state on a regular lattice at constant temperature up to trace distance ϵ. This will follow from the results of [27] and showcase the need to focus on the Wasserstein distance instead of the trace distance.

G.1 Lower bounds for sample complexity of classical shadows

One of the main advantages of our results compared with the classical shadows method of [40] is that whenever a TC inequality is available, we can learn all $k-$local observables with a number of samples that grows polynomially in k, whereas classical shadows require a number of samples that grows exponentially in k. However, the classical shadows framework does not assume any structure for the underlying state. Thus, it is natural to ask if this undesired exponential scaling of the shadows framework is due to this broader applicability.

In this section, we will demonstrate that this is not the case even for one of the simplest imaginable classes of states, namely tensor products of Pauli eigenstates. We will show that if the number of samples is not exponential in the locality of the desired observables, there will always be a k-local observable whose estimate will be wrong with constant probability.

But before we show that, let us briefly recall how the shadow method works. To recover local observables on n qubits, the methods of [40] proceed as follows. First, we sample a random unitary $U = \otimes_{i=0}^{n-1} U_i$, where each U_i is an independent rotation into a Pauli basis. Then we proceed to measure the state of interest ρ in the basis defined by U, obtaining a n-bit classical string $b_0 b_1 \dots b_{n-1}$. The shadow corresponding to this measurement $\hat{\rho}$ is then defined to be given by

$$\hat{\rho} = \otimes_{i=1}^{n} (3U_i^\dagger |b_i\rangle\langle b_i| U_i - I).$$

We then repeat this procedure SM times for $S, M \in \mathbb{N}$, obtaining shadows $\{\hat{\rho}_{s,m}\}_{1\leq m\leq M, 1\leq s\leq S}$. We then set our estimate of the expectation value of an observable O to be given by

$$\hat{O}(\{\hat{\rho}_{s,m}\}) = \text{median}\left\{ S^{-1}\sum_{s=1}^{S} \text{tr}\left[\hat{\rho}_{s,m=1}O\right], S^{-1}\sum_{s=1}^{S} \text{tr}\left[\hat{\rho}_{s,m=2}O\right], \dots, S^{-1}\sum_{s=1}^{S} \text{tr}\left[\hat{\rho}_{s,m=M}O\right] \right\}. \tag{64}$$

That is, we partition the set of shadow samples into M subgroups, take the average on each of them and then take their median. The authors of [40] then proceed to show in Theorem 1 that if we take $SM = \mathcal{O}(4^k \log(N\delta^{-1})\epsilon^{-2})$, then with probability at least $1-\delta$ the expectation value any given N k-local observables of bounded operator norm will deviate by at most ϵ from the estimate given in Eq. (64). We will now show that this exponential scaling in k is unavoidable if we want to obtain nontrivial estimates with high probability. More precisely:

Proposition G.1. *Let $\rho = \otimes_{i=1}^{n} |\phi_i\rangle\langle\phi_i|$ be an unknown n qubit state where each $|\phi_i\rangle$ is given by a Pauli eigenstate. For $k \geq \log_3(n)\log(n)$, if $SM \leq n$ then there is an observable O of the form*

$$O = n^{-1} \sum_{i=1}^{n} O_i$$

where each O_i is k-local and such that:

$$\mathrm{tr}\left[O\rho\right] = 1, \quad \hat{O}(\{\hat{\rho}_{s,m}\}) = 0 \tag{65}$$

with probability at least $1 - e^{-1}$.

Proof. We will let $O_i = \otimes_{j=i}^{i+k} P_j$, where we take addition modulo n and let P_i be the Pauli matrix such that $P_i |\phi_i\rangle = 1$, where we assume w.l.o.g. that each Pauli eigenstate corresponds to a $+1$ eigenstate. It is then clear that $\mathrm{tr}\left[O\rho\right] = 1$.

Let us now analyse the performance of shadows. Note that as the random unitaries used to obtain each sample are random Pauli bases, we have that $\mathrm{tr}\left[O_i \hat{\rho}_{s,m}\right] = 3^k$ if the unitary $U^{s,m}$ corresponds to the identity on qubits $i, i+1, \ldots, i+k$ and 0 otherwise. This is because then the string we measure will be rotated to a Pauli basis different from that of P_i on at least one of the qubits in that interval. Thus, as we pick the Pauli bases uniformly at random and there are three different possibilities for each one of them, we see that the probability that a shadow is different from 0 on a given O_i is 3^{-k}. By a union bound, the probability that a shadow is different than 0 on any of the n O_i is at most $n3^{-k}$. As the different shadows are independent, the probability that all of the SM shadows return 0 is at least $(1 - n3^{-k})^{SM}$. As we picked $k \geq \log_3(n)\log(n)$ and $SM \leq n$, we have:

$$(1 - n3^{-k})^{SM} \geq (1 - n^{-1})^{n} \geq \frac{e^{-1}}{2}$$

for n large enough. Thus, as all shadows will have expectation value 0, the median and means procedure will clearly also output 0, which concludes the proof. □

Despite the fact the proof above used quite rough estimates and simple observables and states, it still gives some intuition as to why classical shadows require an exponential number of samples in the locality. We see that the probability that the shadows "look in the right direction" is exponentially small in the locality of the observables, in the sense that the overlap between the state and a random Pauli eigenstate is likely to be exponentially small. And whenever it does look in a direction in which the underlying state has a significant overlap with that basis it has to compensate that direction exponentially. Thus, if the number of samples is not exponential in the locality, it is unlikely that we will measure in a direction that has significant overlap with our state or there will be significant fluctuations due to the exponential rewarding of the "good" directions.

However, by combining shadows with a locality structure and the maximum entropy principle, as we do in this work, we can bypass the need to measure in random directions for observables with a large locality, bypassing this exponential scaling.

We also note that the authors of [40] already proved the optimality of their protocol by only considering product states in Section 8 of their supplemental material. The main difference between their proof and ours is that we focus on a Lipschitz observable, whereas they focus on observables that only depend on k qubits.

G.2 Lower bounds for recovery in trace distance

In the main text, we claimed that one of the reasons why we obtain an exponential speedup compared to usual many-body methods is that we focus on a good recovery in the Wasserstein distance instead of trace distance. Moreover, by combining the maximum entropy method with a TC inequality we are able to obtain good recovery guarantees from a constant *relative entropy density*. That is, as long as two Gibbs states $\sigma(\lambda), \sigma(\mu)$ satisfy

$$D(\sigma(\lambda) \| \sigma(\mu)) \leq \epsilon n, \tag{66}$$

for some $\epsilon > 0$ we already obtain some nontrivial guarantees.

In this section, we will argue that focusing on the Wasserstein recovery instead of trace distance is essential to obtain nontrivial recovery guarantees from a number of samples scaling logarithmically with the system size. To achieve this, we will resort to results of [27, Theorem 1.3]:

Proposition G.2. *[Lower bound of sample complexity in trace distance] Let $G = (V, E)$ be a graph on n vertices and m edges and for $\lambda \in \mathbb{R}^{15m}$, $\|\lambda\|_{\ell_\infty} \leq 1$ let $H(\lambda)$ be defined as*

$$H(\lambda) = \sum_{i\sim j} \sum_{l=1}^{15} \lambda_{i,j}^l H_{i,j}^l,$$

where the $H_{i,j}^l$ correspond to some ordering of the nonidentity Pauli strings acting on sites i, j. Then for any $\beta = \Omega(m^{-\frac{1}{2}})$, let $\hat{\sigma}(\lambda)$ be the estimate of $\sigma(\lambda)$ outputted by an algorithm with access to s samples from a state $\sigma(\lambda)$. Then:

$$\sup_{\lambda \in B_\infty(0,1)} \|\sigma(\lambda) - \hat{\sigma}(\lambda)\|_{\mathrm{tr}} = \Omega\left(\min\left\{1, \sqrt{\frac{m}{s}}\right\}\right). \tag{67}$$

Proof. This statement immediately follows from [27, Theorem 1.3], which shows the analogous statement when restricted to classical Ising models. As our class of Hamiltonians includes those as a subset, any algorithm that could provide an estimate for this more general class also can find one for the classical instances. Moreover, in the proof of [27, Theorem 1.3] the inverse temperature is absorbed into the coefficients of the Hamiltonian, which are assumed to have 2-norm bounded by a constant independent of the system's size. This is easily seen to be satisfied by our conditions since $\beta = \Omega(m^{-\frac{1}{2}})$. □

The statement above implies in particular that any algorithm that finds an estimate that is ϵ close in trace distance for all 2-local Gibbs state on a lattice and constant inverse temperature requires $\Omega(n\epsilon^{-2})$ samples. In contrast, we see that it is possible to obtain an estimate that is $\mathcal{O}(\epsilon\sqrt{n})$ close in Wasserstein distance from $\mathcal{O}(\epsilon^{-2}\log(n))$ samples of the Gibbs states, which is sufficient to already give nontrivial recovery guarantees for Lipschitz observables. Thus, we see from Prop. G.2 that resorting to the Wasserstein distance is essential to obtain recovery guarantees in the regime where the number of samples is logarithmic in the system's size.

Furthermore, it is interesting to note that the proof of [27] is based on a set of Gibbs states of the form:

$$\delta \sum_{i\sim j} s_{i,j} Z_i Z_j, \tag{68}$$

with $\delta = \Theta(m^{-\frac{1}{2}})$ and $s_{i,j} \in \{\pm 1\}$. Their proof then proceeds by finding a large subset of Gibbs states of the form in Eq. (68) which have a trace distance and relative entropy of constant order. The lower bound on the sample complexity then follows from standard information-theoretic arguments. We believe that this class of examples in the proof further illustrates why the trace distance is not necessarily the adequate distance measure when estimating the error on extensive observables. Indeed, for extensive, local observables the class of Gibbs states from the Hamiltonians in Eq. (68) behaves like the maximally mixed state, as each local term converges to 0 as the system size increases.