

Supplementary Information for “On the Sample Complexity of Quantum Boltzmann Machine Learning”

Luuk Coopmans and Marcello Benedetti

Quantinuum, Partnership House, Carlisle Place, London SW1P 1BX, United Kingdom

CONTENTS

Supplementary Note 1: Some useful mathematical facts and relations	2
Convexity	2
Derivative of a matrix exponential	3
Supplementary Note 2: Properties of the quantum relative entropy for QBMs	4
Strict convexity	4
Strong convexity	5
L -Smoothness	6
Supplementary Note 3: Convergence results of SGD for training QBMs	7
Review of SGD convergence results	7
Proofs of Theorems 1 and 2 in the main text	8
Achieving a desired precision on the quantum relative entropy	10
Additional numerical results	11
Supplementary Note 4: Guaranteed performance improvement by pre-training	12
Proof of Theorem 3 in the main text	12
Pre-training models	13
Mean-Field QBM	13
Gaussian Fermionic QBM	14
Geometrically-Local QBM	15
Supplementary Note 5: Construction of the target state expectation values	15
Classical data encoding	15
Data used for the numerical simulations	16

SUPPLEMENTARY NOTE 1: SOME USEFUL MATHEMATICAL FACTS AND RELATIONS

Here we state and derive some useful mathematical results that are used in the proofs in the other appendices. The definitions and lemmas about convexity are taken from [1]. The derivative of the matrix exponential expressed as quantum belief propagation is a result of [2] and we rederive it in a pedagogical way.

Convexity

Definition 1 (Convexity). *A multivariate function $f : \mathbb{R}^m \mapsto \mathbb{R}$ is said to be convex when*

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y), \quad \forall x, y \in \mathbb{R}^m, \quad t \in [0, 1]. \quad (1)$$

If additionally the gradient $\nabla f(x^)$ is zero only for one unique vector $x^* \in \mathbb{R}^m$, then f is said to be strictly convex.*

The following Lemma can be deduced from the standard definition of convexity [1].

Lemma 1. *Let f be twice continuously differentiable. Then f is convex if*

$$v^T \nabla^2 f(x) v \geq 0, \quad \forall x, v \in \mathbb{R}^m. \quad (2)$$

A stronger version of convexity is used in some of our discussions.

Definition 2 (α -Polyak-Łojasiewicz). *Let $f : \mathbb{R}^m \rightarrow \mathbb{R}$, and $\alpha > 0$. We say that f is α -Polyak-Łojasiewicz if*

$$\frac{1}{2\alpha} \|\nabla f(x)\|^2 \geq f(x) - \min_x f(x), \quad (3)$$

where $\|\cdot\|$ is the Euclidean norm.

An even stronger convexity condition is the following.

Definition 3 (α -strong convexity). *Let $f : \mathbb{R}^m \rightarrow \mathbb{R}$, and $\alpha > 0$. We say that f is α -strongly convex if*

$$\frac{\alpha t(1-t)}{2} \|x - y\|^2 + f(tx + (1-t)y) \leq tf(x) + (1-t)f(y). \quad (4)$$

The latter implies the former.

Lemma 2. *If f is α -strongly convex then f is α -Polyak-Łojasiewicz.*

The strong convexity of a function can be tested as follows.

Lemma 3. *Let f be twice continuously differentiable. Then f is α -strongly convex if*

$$v^T \nabla^2 f(x) v \geq \alpha \|v\|^2, \quad \forall x, v \in \mathbb{R}^m. \quad (5)$$

Besides convexity, we also need to characterize the smoothness of a function.

Definition 4 (L -smoothness). *Let $f : \mathbb{R}^m \rightarrow \mathbb{R}$ and $L > 0$. We say that f is L -smooth if it is differentiable and if the gradient ∇f is L -Lipschitz:*

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|, \quad \forall x, y \in \mathbb{R}^m. \quad (6)$$

For L -smooth functions we have the following useful property [1].

Lemma 4 (Descent lemma). *Let $f : \mathbb{R}^m \rightarrow \mathbb{R}$ be a twice differentiable, L -smooth function, then*

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|^2. \quad (7)$$

Derivative of a matrix exponential

Lemma 5 (Quantum belief propagation [2]). *Let $H = W + \theta V$ be a Hamiltonian ansatz with parameter θ . Let $f(t)$ be the function for which $\int_{-\infty}^{\infty} f(t) e^{-it\omega} dt = \frac{\tanh(\omega/2)}{\omega/2}$, and define the operator Fourier transform as $\Phi(\cdot) = \int_{-\infty}^{\infty} f(t) e^{itH} \cdot e^{-itH} dt$. Then*

$$\partial_{\theta} e^H = \frac{1}{2} \{ \Phi(V), e^H \}, \quad (8)$$

where $\{A, B\} = AB + BA$ is the anti-commutator.

Proof. The derivative of the matrix exponential e^H with respect to a parameter is given by Duhamel's formula [3]

$$\partial_{\theta} e^H = \int_0^1 e^{(1-s)H} (\partial_{\theta} H) e^{sH} ds. \quad (9)$$

Taking $H = W + \theta V$, with simple manipulations we find a useful alternative expression

$$\begin{aligned} \partial_{\theta} e^H &= e^H \int_0^1 e^{-sH} V e^{sH} ds \\ &= \sum_{j,k} |j\rangle\langle k| \langle j|V|k\rangle e^{\lambda_j} \int_0^1 e^{s(\lambda_k - \lambda_j)} ds \\ &= \sum_{j,k} |j\rangle\langle k| V_{jk} e^{\lambda_j} \frac{e^{\lambda_k - \lambda_j} - 1}{\lambda_k - \lambda_j}. \end{aligned} \quad (10)$$

Here we use the basis diagonalizing the Hamiltonian, $H = \sum_j \lambda_j |j\rangle\langle j|$, and we introduce the notation $V_{jk} = \langle j|V|k\rangle$. The above expression is valid also for the diagonal entries, $k = j$, since $\lim_{x \rightarrow 0} \frac{e^x - 1}{x} = 1$. Now,

$$e^{\lambda_j} \frac{e^{\lambda_k - \lambda_j} - 1}{\lambda_k - \lambda_j} = e^{\lambda_j} \frac{e^{\lambda_k - \lambda_j} - 1}{e^{\lambda_k - \lambda_j} + 1} \frac{e^{\lambda_k - \lambda_j} + 1}{\lambda_k - \lambda_j} = \frac{\tanh\left(\frac{\lambda_k - \lambda_j}{2}\right)}{\frac{\lambda_k - \lambda_j}{2}} \frac{e^{\lambda_k} + e^{\lambda_j}}{2}. \quad (11)$$

With the notation $\hat{f}(\omega) = \frac{\tanh(\omega/2)}{\omega/2}$ we can write

$$\partial_{\theta} e^H = \sum_{j,k} |j\rangle\langle k| V_{jk} \hat{f}(\lambda_k - \lambda_j) \frac{e^{\lambda_k} + e^{\lambda_j}}{2}. \quad (12)$$

Let us interpret $\hat{f}(\omega)$ as the Fourier transform of another function: $\hat{f}(\omega) = \int_{-\infty}^{\infty} f(t) e^{-it\omega} dt$. Plugging this in the previous expression we obtain

$$\begin{aligned} \partial_{\theta} e^H &= \sum_{j,k} |j\rangle\langle k| V_{jk} \int_{-\infty}^{\infty} f(t) e^{-it(\lambda_k - \lambda_j)} dt \left(\frac{e^{\lambda_k}}{2} + \frac{e^{\lambda_j}}{2} \right) \\ &= \frac{1}{2} \int_{-\infty}^{\infty} f(t) \sum_j e^{it\lambda_j} |j\rangle\langle j| V \sum_k e^{-it\lambda_k + \lambda_k} |k\rangle\langle k| dt + \frac{1}{2} \int_{-\infty}^{\infty} f(t) \sum_j e^{it\lambda_j + \lambda_j} |j\rangle\langle j| V \sum_k e^{-it\lambda_k} |k\rangle\langle k| dt \\ &= \frac{1}{2} \int_{-\infty}^{\infty} f(t) e^{itH} V e^{-itH} dt e^H + \frac{1}{2} e^H \int_{-\infty}^{\infty} f(t) e^{itH} V e^{-itH} dt \\ &= \frac{1}{2} \{ \Phi(V), e^H \}. \end{aligned} \quad (13)$$

We have recovered a result from [2] by different means. $\square$

SUPPLEMENTARY NOTE 2: PROPERTIES OF THE QUANTUM RELATIVE ENTROPY FOR QBMS

In this appendix, we prove some properties of the quantum relative entropy $S(\eta\|\rho_\theta)$ of a generic Quantum Boltzmann Machine (QBM) ρ_θ with respect to some arbitrary target η . These properties are used for the proof of the theorems in the main text. We start by showing the convexity and afterward, we show the L -smoothness.

Strict convexity

Lemma 6. *Let $\mathcal{H}_\theta = \sum_{i=1}^m \theta_i H_i$ be a Hamiltonian ansatz where H_i are non-commuting operators, and $\rho_\theta = \frac{e^{\mathcal{H}_\theta}}{Z}$ the corresponding QBM. For all states η , the quantum relative entropy $S(\eta\|\rho_\theta)$ is a strictly convex function of θ .*

Proof. In order to show strict convexity of S we use Lemma 1. We first show that the Hessian of the quantum relative entropy with respect to the QBM parameters, $\nabla^2 S$, is positive semidefinite. Afterwards, we show that S has a unique global optimizer θ^* for which $\nabla S(\eta\|\rho_{\theta^*}) = 0$, and apply the Lemma.

Using the derivative of the matrix exponential in Eq. (13) we have

$$\begin{aligned} \frac{\partial S}{\partial \theta_i} &= \frac{\partial}{\partial \theta_i} \text{Tr}[\eta (\log \eta - \mathcal{H}_\theta + \log \text{Tr}[e^{\mathcal{H}_\theta}])] \\ &= -\text{Tr}[\eta H_i] + \frac{\text{Tr}[\{\Phi(H_i), e^{\mathcal{H}_\theta}\}]}{2 \text{Tr}[e^{\mathcal{H}_\theta}]} \\ &= -\text{Tr}[\eta H_i] + \text{Tr}[\rho_\theta \Phi(H_i)] \\ &= -\text{Tr}[\eta H_i] + \text{Tr}[\rho_\theta H_i]. \end{aligned} \tag{14}$$

In the last step, we use the cyclic property of the trace. This is Eq. (4) in the main text. Next, we take the second derivative

$$\begin{aligned} \frac{\partial^2 S}{\partial \theta_i \partial \theta_j} &= \frac{\partial}{\partial \theta_j} \text{Tr}[\rho_\theta H_i] \\ &= \text{Tr} \left[\left(\frac{\{\Phi(H_j), e^{\mathcal{H}_\theta}\}}{2 \text{Tr}[e^{\mathcal{H}_\theta}]} - \frac{e^{\mathcal{H}_\theta} \text{Tr}[\{\Phi(H_j), e^{\mathcal{H}_\theta}\}]}{2(\text{Tr}[e^{\mathcal{H}_\theta}])^2} \right) H_i \right] \\ &= \frac{1}{2} \text{Tr}[\rho_\theta \{\Phi(H_j), H_i\}] - \text{Tr}[\rho_\theta H_j] \text{Tr}[\rho_\theta H_i]. \end{aligned} \tag{15}$$

In the last step we used $\text{Tr}[A\{B, C\}] = \text{Tr}[C\{A, B\}]$ to rearrange the terms. We now show that the Hessian is positive semidefinite and satisfies Eq. (2). Note that one can arrive at the same result using Lemma 27 in [4]. For any vector $v \in \mathbb{R}^m$, we have

$$\begin{aligned} v^T \nabla^2 S v &= \sum_{n,m} v_n v_m \left(\frac{1}{2} \text{Tr}[\rho_\theta \{\Phi(H_n), H_m\}] - \text{Tr}[\rho_\theta H_n] \text{Tr}[\rho_\theta H_m] \right) \\ &= \frac{1}{2} \text{Tr} \left[\rho_\theta \left\{ \Phi \left(\sum_n v_n H_n \right), \sum_m v_m H_m \right\} \right] - \text{Tr} \left[\rho_\theta \sum_n v_n H_n \right] \text{Tr} \left[\rho_\theta \sum_m v_m H_m \right] \\ &= \frac{1}{2} \text{Tr}[\rho_\theta \{\Phi(W), W\}] - \text{Tr}[\rho_\theta W] \text{Tr}[\rho_\theta W] \\ &= \text{Re} \text{Tr}[\rho_\theta \Phi(W)W] - \text{Tr}[\rho_\theta W]^2, \end{aligned} \tag{16}$$

where we define the Hermitian operator $W = \sum_n v_n H_n$ and use that $\Phi(W)$ is also Hermitian. We denote the real

part of a complex number by Re . Let us inspect the first term in the basis diagonalizing the Hamiltonian

$$\begin{aligned}
\text{Re Tr}[\rho_\theta \Phi(W)W] &= \text{Re Tr} \left[\left(\sum_j \frac{e^{\lambda_j}}{\text{Tr}[e^{\mathcal{H}_\theta}]} |j\rangle\langle j| \right) \int_{-\infty}^{\infty} f(t) \left(\sum_k e^{it\lambda_k} |k\rangle\langle k| \right) W \left(\sum_l e^{-it\lambda_l} |l\rangle\langle l| \right) dt W \right] \\
&= \text{Re} \sum_{j,k,l} \frac{e^{\lambda_j}}{\text{Tr}[e^{\mathcal{H}_\theta}]} \int_{-\infty}^{\infty} f(t) e^{it\lambda_k} e^{-it\lambda_l} dt \langle j|k\rangle W_{kl} W_{lj} \\
&= \text{Re} \sum_{j,l} \frac{e^{\lambda_j}}{\text{Tr}[e^{\mathcal{H}_\theta}]} \int_{-\infty}^{\infty} f(t) e^{-it(\lambda_l - \lambda_j)} dt W_{jl} W_{lj} \\
&= \text{Re} \sum_{j,l} \frac{e^{\lambda_j}}{\text{Tr}[e^{\mathcal{H}_\theta}]} \frac{\tanh\left(\frac{\lambda_l - \lambda_j}{2}\right)}{\frac{\lambda_l - \lambda_j}{2}} |W_{jl}|^2 \\
&= \sum_j \frac{e^{\lambda_j}}{\text{Tr}[e^{\mathcal{H}_\theta}]} |W_{jj}|^2 + \sum_{j,l \neq j} \frac{e^{\lambda_j}}{\text{Tr}[e^{\mathcal{H}_\theta}]} \frac{\tanh\left(\frac{\lambda_l - \lambda_j}{2}\right)}{\frac{\lambda_l - \lambda_j}{2}} |W_{jl}|^2 \\
&\geq \left(\sum_j \frac{e^{\lambda_j}}{\text{Tr}[e^{\mathcal{H}_\theta}]} W_{jj} \right)^2.
\end{aligned} \tag{17}$$

Here we use the definition of the quantum belief propagation operator to evaluate the integral. The last step follows from Jensen's inequality applied to the first term, and by ignoring the second term (which is always positive). Plugging this in Eq. (16) we obtain the desired result

$$v^T \nabla^2 S v = \text{Re Tr}[\rho_\theta \Phi(W)W] - \text{Tr}[\rho_\theta W]^2 \geq \left(\sum_j \frac{e^{\lambda_j}}{\text{Tr}[e^{\mathcal{H}_\theta}]} W_{jj} \right)^2 - \text{Tr}[\rho_\theta W]^2 = 0. \tag{18}$$

We conclude that the quantum relative entropy is convex. We now show strict convexity by a contradiction argument, following Proposition 17 in [4]. Assume we have found one set of parameters θ^* with $\nabla S(\eta \| \rho_{\theta^*}) = 0$. Then from Eq. (14) we have $\langle H_i \rangle_\eta = \langle H_i \rangle_{\rho_{\theta^*}}$ for all H_i . Note that we can always find at least one such θ^* by Jaynes' principle [5]. Next, assume there exists a different set of parameters, $\chi \neq \theta^*$, with $\langle H_i \rangle_\eta = \langle H_i \rangle_{\rho_\chi}$ for all H_i . Then

$$\begin{aligned}
S(\rho_\chi \| \rho_{\theta^*}) &= \text{Tr}[\rho_\chi \log \rho_\chi] - \text{Tr}[\rho_\chi \log \rho_{\theta^*}] \\
&= \text{Tr}[\rho_\chi \log \rho_\chi] - \sum_i \theta_i^* \text{Tr}[\rho_\chi H_i] + \log Z_{\theta^*} \\
&= \text{Tr}[\rho_\chi \log \rho_\chi] - \sum_i \theta_i^* \text{Tr}[\rho_{\theta^*} H_i] + \log Z_{\theta^*} \\
&= \text{Tr}[\rho_\chi \log \rho_\chi] - \text{Tr}[\rho_{\theta^*} \log \rho_{\theta^*}] \\
&\geq 0.
\end{aligned} \tag{19}$$

Similarly, by swapping ρ_χ and ρ_{θ^*} , we find

$$S(\rho_{\theta^*} \| \rho_\chi) = \text{Tr}[\rho_{\theta^*} \log \rho_{\theta^*}] - \text{Tr}[\rho_\chi \log \rho_\chi] \geq 0. \tag{20}$$

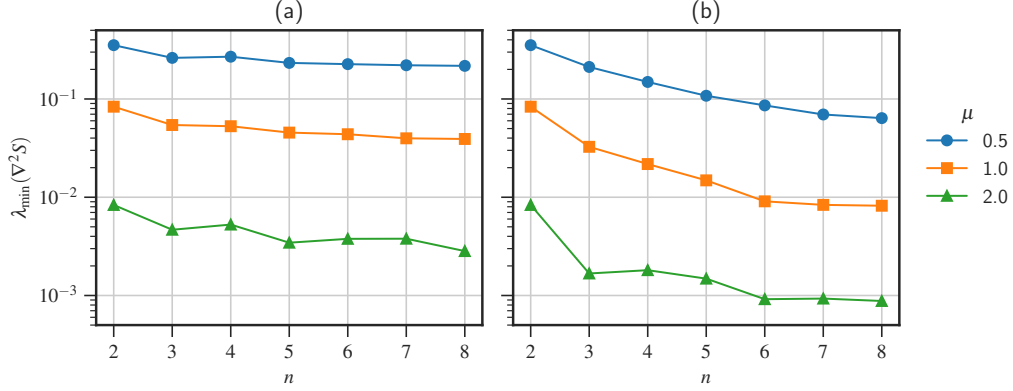
It follows that $S(\rho_{\theta^*} \| \rho_\chi) = 0$, implying $\rho_{\theta^*} = \rho_\chi$. Since the operators H_i are orthogonal we have that $\theta^* = \chi$. But this contradicts the assumption made at the beginning ($\theta^* \neq \chi$). Thus we can only have one unique θ^* such that $\nabla S(\eta \| \rho_{\theta^*}) = 0$, i.e., S is strictly convex by Definition 1. $\square$

Strong convexity

To show α -strong convexity of S one can use Lemma 3. To the best of our knowledge, there is no proof in the literature showing that the quantum relative entropy of Gibbs states is strongly convex in general. On the other hand, this property has been proven for particular classes of Hamiltonians. Anshu et al. [4] proves strong convexity for k -local

Hamiltonians defined on a finite dimensional lattice. They show that in this case $\alpha \in \Omega(1/n)$, a polynomial decrease with respect to the system size. Haah et al. [6] proves strong convexity for the more general class of low-intersection Hamiltonians. Low-intersection Hamiltonians have terms that act non-trivially only on a constant number of qubits, and each term intersects non-trivially with a constant number of other terms.

In this section, we use differentiable programming [7] to numerically analyze the smallest eigenvalue of the Hessian, $\lambda_{\min}(\nabla^2 S)$, seeking evidence for strong convexity. We consider a 1D nearest-neighbor Hamiltonian, $\mathcal{H} = \sum_{i=1}^{n-1} h_{i,i+1}^{xx} X_i X_{i+1} + h_{i,i+1}^{yy} Y_i Y_{i+1} + h_{i,i+1}^{zz} Z_i Z_{i+1}$, and a fully-connected one $\mathcal{H} = \sum_{i=1}^{n-1} \sum_{j>i}^n h_{i,j}^{xx} X_i X_j + h_{i,j}^{yy} Y_i Y_j + h_{i,j}^{zz} Z_i Z_j$. We randomly sample coefficients uniformly in $[-\mu, \mu]$ where μ is a scale parameter and determines the max-norm of the vector of the coefficients. Supplementary Figure 1 shows mean and standard deviation over 25 instances. The smallest eigenvalue decreases with the number of qubits until convergence to a fixed value. The fully-connected Hamiltonian has $m \in O(n^2)$ parameters and yields smaller eigenvalues than the 1D Hamiltonian which has $m \in O(n)$ parameters instead. These results provide evidence of strong convexity with α decreasing polynomially with the system size. We leave more extensive numerical studies for future work.



Supplementary Figure 1. Minimum eigenvalue of the Hessian, as a function of the number of qubits. We show the median of 25 random instances for the 1D nearest-neighbor Hamiltonian (a), and fully-connected Hamiltonian (b). The scale parameter μ determines the maximum size of random parameters. We observe that in all cases the smallest eigenvalues shrink with the number of qubits, but appear to plateau to a positive value.

L-Smoothness

Lemma 7. Let $\mathcal{H}_\theta = \sum_{i=1}^m \theta_i H_i$ be a Hamiltonian ansatz where H_i are non-commuting operators, and $\rho_\theta = \frac{e^{\mathcal{H}_\theta}}{Z}$ the corresponding QBM. For all states η , the quantum relative entropy $S(\eta||\rho_\theta)$ is a *L*-smooth function of θ with $L = 2m \max_j \|H_j\|_2^2$.

Proof. We look for an upper bound on the largest eigenvalue of the Hessian in Eq. (15). We begin with the following property

$$\|\Phi(V)\| = \left\| \sum_{j,k} |j\rangle\langle k| V_{jk} \hat{f}(\lambda_k - \lambda_j) \right\| \leq \hat{f}_{\max} \left\| \sum_{j,k} |j\rangle\langle k| V_{jk} \right\| = \|V\|, \quad (21)$$

where we use that $\hat{f} > 0$ and $\hat{f}_{\max} = 1$. In what follows, we make use of this bound with the spectral norm $\|\cdot\|_2$, which is the operator norm induced by the Euclidean vector norm ($p = 2$). Let us bound the entries of the Hessian

$$\begin{aligned} \left| \frac{\partial^2 S}{\partial \theta_j \partial \theta_k} \right| &= \left| \frac{1}{2} \text{Tr}[\rho_\theta \{\Phi(H_j), H_k\}] - \text{Tr}[\rho_\theta H_j] \text{Tr}[\rho_\theta H_k] \right| \\ &\leq \frac{1}{2} \|\{\Phi(H_j), H_k\}\|_2 + \|H_j\|_2 \|H_k\|_2 \\ &\leq \|\Phi(H_j)\|_2 \|H_k\|_2 + \|H_j\|_2 \|H_k\|_2 \\ &\leq 2\|H_j\|_2 \|H_k\|_2. \end{aligned} \quad (22)$$

Here we use the triangle inequality, that expectations are bounded by the spectral norm, the sub-multiplicative property of the spectral norm, and Eq. (21). We are now able to upper-bound the largest eigenvalue $\lambda_{\max}$ of the Hessian

$$\begin{aligned}
\|\nabla^2 S\|_2 &= |\lambda_{\max}(\nabla^2 S)| \\
&\leq \max_j \sum_{k=1}^m \left| \frac{\partial^2 S}{\partial \theta_j \partial \theta_k} \right| \\
&\leq m \max_j \max_k \left| \frac{\partial^2 S}{\partial \theta_j \partial \theta_k} \right| \\
&\leq 2m \max_j \max_k \|H_j\|_2 \|H_k\|_2 \\
&= 2m \max_j \|H_j\|_2^2.
\end{aligned} \tag{23}$$

The first inequality is a consequence of the Gershgorin circle theorem. We can now prove the L -smoothness of the quantum relative entropy. Let us define a function $h(t) = \nabla S(\eta\|\rho_{y+t(x-y)})$. Then we have

$$\begin{aligned}
\|\nabla S(\eta\|\rho_x) - \nabla S(\eta\|\rho_y)\|_2 &= \|h(1) - h(0)\|_2 \\
&= \left\| \int_0^1 h'(t) dt \right\|_2 \\
&\leq \int_0^1 \|h'(t)\|_2 dt \\
&= \int_0^1 \|\nabla^2 S(\eta\|\rho_{y+t(x-y)}) (x-y)\|_2 dt \\
&\leq \int_0^1 \|\nabla^2 S(\eta\|\rho_{y+t(x-y)})\|_2 \|x-y\|_2 dt \\
&\leq 2m \max_j \|H_j\|_2^2 \|x-y\|_2,
\end{aligned} \tag{24}$$

where in the last step we used Eq. (23). Thus the quantum relative entropy is L -smooth with $L = 2m \max_j \|H_j\|_2^2$. $\square$

SUPPLEMENTARY NOTE 3: CONVERGENCE RESULTS OF SGD FOR TRAINING QBMS

In this appendix we first review useful results from the machine learning literature, then prove Theorems 1 and 2 in the main text. We also discuss a few upper bounds for the relative entropy in the context of QBM learning. Finally, we provide additional numerical results for the scaling behavior of our theorems.

Review of SGD convergence results

We begin by stating three convergence results from the Stochastic Gradient Descent (SGD) literature. Consider a loss function $f: \mathbb{R}^m \rightarrow \mathbb{R}$ that is L -smooth (Def. 4) and bounded from below by $f_{\inf} \in \mathbb{R}$. The stochastic gradient is unbiased, i.e., $\mathbb{E}[\hat{g}] = \nabla f$, and satisfies

$$\mathbb{E}\|\hat{g}(x)\|^2 \leq 2A(f(x) - f_{\inf}) + B\|\nabla f(x)\|^2 + C, \tag{25}$$

for some $A, B, C \geq 0$ and all $x \in \mathbb{R}^m$. SGD iteratively minimizes f according to the update rule $x^t = x^{t-1} - \gamma^t \hat{g}_{x^{t-1}}$ at time step t . Khaled and Richtárik [8] proved the following SGD convergence result.

Lemma 8 (restatement of Corollary 1 in [8]). *Choose precision $\epsilon > 0$ and step size $\gamma = \min\{\frac{1}{\sqrt{LAT}}, \frac{1}{LB}, \frac{\epsilon}{2LC}\}$, and set $\delta_0 = \mathbb{E}[f(x^0)] - f_{\inf}$. Then provided that*

$$T \geq \frac{12\delta_0 L}{\epsilon^2} \max\left\{B, \frac{12\delta_0 A}{\epsilon^2}, \frac{2C}{\epsilon^2}\right\}, \tag{26}$$

we have that SGD converges with

$$\min_{1 \leq t \leq T} \mathbb{E}\|\nabla f(x^t)\| \leq \epsilon. \tag{27}$$

Here $\mathbb{E}[\cdot]$ denotes the expectation with respect to x^t , which is a random variable due to the stochasticity in the gradient. Let us now consider a loss function which, in addition to the previous conditions, is also α -Polyak-Łojasiewicz (Def. 2). We consider the following iterative learning rate scheme for γ_t .

Lemma 9 (restatement of Lemma 3 in [8]). *Consider a sequence $(r_t)_t$ satisfying $r_{t+1} \leq (1 - a\gamma_t)r_t + c\gamma_t^2$, where $\gamma_t \leq \frac{1}{b}$ for all $t \geq 0$ and $a, c \geq 0$ with $a \leq b$. Fix $T > 0$ and let $k_0 = \lceil \frac{T}{2} \rceil$. Then choosing the stepsize as*

$$\gamma_t = \begin{cases} \frac{1}{b}, & \text{if } T \leq \frac{b}{a} \text{ or } t < k_0, \\ \frac{2}{a(s+t-k_0)}, & \text{if } T \geq \frac{b}{a} \text{ and } t > k_0, \end{cases} \quad (28)$$

with $s = \frac{2b}{a}$ gives $r_T \leq \exp\{-\frac{aK}{2b}\}r_0 + \frac{9c}{a^2T}$

For this learning rate scheme, Khaled and Richtárik [8] proved the following SGD convergence result.

Lemma 10 (restatement of Corollary 2 in [8]). *Choose precision $\epsilon > 0$ and step size γ_t following Lemma 9 with $\gamma_t \leq \min\{\frac{\alpha}{2AL}, \frac{1}{2BL}\}$. Then provided that*

$$T \geq \frac{L}{\alpha} \max \left\{ \frac{2A}{\alpha} \log \frac{2\delta_0}{\epsilon}, \frac{2B}{\alpha} \log \frac{2\delta_0}{\epsilon}, \frac{9C}{2\alpha\epsilon} \right\} \quad (29)$$

we have that SGD converges with

$$\mathbb{E}|f(x^T) - f_{\inf}| \leq \epsilon. \quad (30)$$

Proofs of Theorems 1 and 2 in the main text

We prove Theorem 1, which is repeated here for completeness.

Theorem 1 (QBM training). *Given a QBM defined by a set of n -qubit Pauli operators $\{H_i\}_{i=1}^m$, a precision κ for the QBM expectations, a precision ξ for the data expectations, and a target precision ϵ such that $\kappa^2 + \xi^2 \geq \frac{\epsilon}{2m}$. After*

$$T = \frac{48\delta_0 m^2 (\kappa^2 + \xi^2)}{\epsilon^4} \quad (31)$$

iterations of stochastic gradient descent on the relative entropy $S(\eta||\rho_\theta)$ with constant learning rate $\gamma^t = \frac{\epsilon}{4m^2(\kappa^2 + \xi^2)}$, we have

$$\min_{t=1, \dots, T} \mathbb{E} |\langle H_i \rangle_{\rho_{\theta^t}} - \langle H_i \rangle_\eta| \leq \epsilon, \quad \forall i, \quad (32)$$

where $\mathbb{E}[\cdot]$ denotes the expectation with respect to the random variable θ^t . Each iteration $t \in \{0, \dots, T\}$ requires

$$N \in \mathcal{O} \left(\frac{1}{\kappa^4} \log \frac{m}{1 - \lambda^{\frac{1}{T}}} \right) \quad (33)$$

preparations of the Gibbs state ρ_{θ^t} , and the success probability of the full algorithm is λ . Here, $\delta_0 = S(\eta||\rho_{\theta^0}) - S(\eta||\rho_{\theta^{\text{opt}}})$ is the relative entropy difference with the optimal model $\rho_{\theta^{\text{opt}}}$.

Proof. Recall from Supplementary Note 2 that the quantum relative entropy is L -smooth with $L = 2m \max_i \|H_i\|_2^2$ and that for Pauli operators $\|H_i\|_2 = 1$. Then, we can minimize the relative entropy by SGD and apply the convergence result in Lemma 8.

For the SGD algorithm, we need an unbiased gradient estimator with bounded variance. We recall that the gradient of the relative entropy is given by $\partial_{\theta_i} S(\eta||\rho_\theta) = \langle H_i \rangle_{\rho_\theta} - \langle H_i \rangle_\eta$. The target expectation values $\langle H_i \rangle_\eta$ are estimated as $\hat{h}_{i,\eta}$ from the data set, as described in Supplementary Note 5. Note that $|\langle H_i \rangle_\eta - \hat{h}_{i,\eta}| \leq \xi$, where $\xi > 0$ is limited by the size of the data set. One can improve on ξ by collecting more data, as long as the number of samples is polynomial in n .

For estimating the QBM expectation values $\langle H_i \rangle_{\rho_\theta}$, we can use a number of techniques. Here we focus on classical shadow tomography. From Theorem 4 in [9], there exists a procedure that returns the expectation values of m different Pauli operators $\{H_i\}$ to precision κ with $\mathcal{O}(\frac{\log m/\tilde{\lambda}}{\kappa^4})$ preparations of ρ_θ , and succeeds with probability $1 - \tilde{\lambda}$. To make use of this result we now restrict the H_i in the QBM Hamiltonian to be Pauli operators. As discussed in the main text,

this is not a limitation, and it is possible to use non-Pauli operators in combination with other shadow tomography protocols. We can therefore obtain estimators $\hat{h}_{i,\rho_\theta}$ such that

$$\max_i |\hat{h}_{i,\rho_\theta} - \langle H_i \rangle_{\rho_\theta}| \leq \kappa. \quad (34)$$

We then use $\hat{g}_{\theta_i} = \hat{h}_{i,\rho_\theta} - \hat{h}_{i,\eta}$ as estimators for the partial derivatives of the quantum relative entropy. The variance of the norm of the gradient estimator is bounded as

$$\begin{aligned} \mathbb{E} \|\hat{g}_\theta - \nabla S(\eta \| \rho_\theta)\|^2 &= \mathbb{E} \sum_{i=1}^m (\hat{h}_{i,\rho_\theta} - \hat{h}_{i,\eta} - \langle H_i \rangle_{\rho_\theta} + \langle H_i \rangle_\eta)^2 \\ &\leq \mathbb{E} \sum_{i=1}^m (\hat{h}_{i,\rho_\theta} - \langle H_i \rangle_{\rho_\theta})^2 + (\hat{h}_{i,\eta} - \langle H_i \rangle_\eta)^2 \\ &\leq m(\kappa^2 + \xi^2). \end{aligned} \quad (35)$$

Since the variance can also be written as $\mathbb{E} \|\hat{g}_\theta\|^2 - \|\nabla S(\eta \| \rho_\theta)\|^2$ we find that our setup is compatible with Eq. (29) for $A = 0, B = 1, C = m(\kappa^2 + \xi^2)$. We choose $\epsilon < 1$ and $\kappa^2 + \xi^2 \geq \frac{\epsilon}{2m}$ in Lemma 8. This yields a learning rate of $\gamma = \frac{\epsilon}{4m^2(\kappa^2 + \xi^2)}$. We conclude that after

$$T \geq \frac{48\delta_0 m^2 (\kappa^2 + \xi^2)}{\epsilon^4} \quad (36)$$

iterations of SGD we have

$$\min_{1 \leq t \leq T} \mathbb{E} \|\nabla S(\eta \| \rho_{\theta^t})\| \leq \epsilon. \quad (37)$$

Here $\delta_0 = S(\eta \| \rho_{\theta^0}) - S(\eta \| \rho_{\theta^{\text{opt}}})$ is the relative entropy at the initialization minus the relative entropy at the optimum. Importantly we note the QBM expectation values are computed with a success probability $1 - \tilde{\lambda}$ at each iteration. Consequently, the total success probability of the whole training is equal to $(1 - \tilde{\lambda})^T$ for T update steps. Then to have a total success probability of λ we need to set $\tilde{\lambda} = (1 - \lambda^{1/T})$ in the shadow tomography protocol. This result, together with the sampling bound on the number of measurements of the shadow tomography, $\mathcal{O}(\frac{\log m/\tilde{\lambda}}{\kappa^4})$, completes the proof of Theorem 1. $\square$

We now provide a proof for Theorem 2, which we restate here.

Theorem 2 (α -strongly convex QBM training). *Given a QBM defined by a Hamiltonian ansatz $\mathcal{H}_\theta$ such that $S(\eta \| \rho_\theta)$ is α -strongly convex, a precision κ for the QBM expectations, a precision ξ for the data expectations, and a target precision ϵ such that $\kappa^2 + \xi^2 \geq \frac{\epsilon}{2m}$. After*

$$T = \frac{18m^2(\kappa^2 + \xi^2)}{\alpha^2 \epsilon^2} \quad (38)$$

iterations of stochastic gradient descent on the relative entropy $S(\eta \| \rho_\theta)$ with learning rate $\gamma^t \leq \frac{1}{4m^2}$, we have

$$\min_{t=1,\dots,T} \mathbb{E} |\langle H_i \rangle_{\rho_{\theta^t}} - \langle H_i \rangle_\eta| \leq \epsilon, \quad \forall i. \quad (39)$$

Each iteration requires the number of samples given in Eq. (33).

Proof. In order to prove this theorem, we first show that η , ρ^{opt} , and ρ_θ are ‘collinear’ with respect to the relative entropy. This is helpful because it allows us to directly bound the difference in relative entropy $S(\eta \| \rho_\theta) - S(\eta \| \rho_{\theta^{\text{opt}}})$ instead of bounding the individual relative entropies.

$$\begin{aligned} S(\eta \| \rho_\theta) - S(\eta \| \rho_{\theta^{\text{opt}}}) &= -\text{Tr}[\eta \log \rho_\theta] + \text{Tr}[\eta \log \rho_{\theta^{\text{opt}}}] \\ &= -\text{Tr}[\eta \mathcal{H}_\theta] + \log Z_\theta + \text{Tr}[\eta \mathcal{H}_{\theta^{\text{opt}}}] - \log Z_{\theta^{\text{opt}}} \\ &= -\text{Tr}[\rho_{\theta^{\text{opt}}} \mathcal{H}_\theta] + \log Z_\theta + \text{Tr}[\rho_{\theta^{\text{opt}}} \mathcal{H}_{\theta^{\text{opt}}}] - \log Z_{\theta^{\text{opt}}} \\ &= -\text{Tr}[\rho_{\theta^{\text{opt}}} \log \rho_\theta] + \text{Tr}[\rho_{\theta^{\text{opt}}} \log \rho_{\theta^{\text{opt}}}] \\ &= S(\rho_{\theta^{\text{opt}}} \| \rho_\theta). \end{aligned} \quad (40)$$

Here, in the going from the second to the third line, we used the fact that $\text{Tr}[\eta H_i] = \text{Tr}[\rho_{\theta^{\text{opt}}} H_i]$, which follows from setting Eq. (14) to zero. Rearranging the terms we get the collinearity $S(\eta \| \rho_\theta) = S(\eta \| \rho_{\theta^{\text{opt}}}) + S(\rho_{\theta^{\text{opt}}} \| \rho_\theta)$. This is a non-trivial result because the relative entropy is not a distance: it is not symmetric and does not satisfy the triangle inequality in general. With this relation, we are now able to prove Theorem 2.

The relative entropy $S(\rho_{\theta^{\text{opt}}} \| \rho_\theta)$ satisfies all the relevant assumptions for SGD convergence: it is an L -smooth function with $L = 2m \max_i \|H_i\|_2^2$, it is bounded below by 0, and the stochastic gradient has bounded variance [Eqs. (35) and (33) apply]. In addition, the α -strong convexity assumed by the theorem implies that $S(\rho_{\theta^{\text{opt}}} \| \rho_\theta)$ is α -Polyak-Łojasiewicz by Lemma 2. This means we can invoke Lemma 10. As before, we set $A = 0, B = 1, C = m(\kappa^2 + \xi^2)$ and choose $\epsilon' < 1$ in the Lemma, thus obtaining a maximum learning rate $\gamma_t \leq \frac{1}{4m^2}$. Looking at the case where $\frac{2}{\alpha} \log \frac{2\delta_0}{\epsilon'} \leq \frac{9m(\kappa^2 + \xi^2)}{2\alpha\epsilon'}$ we find that after $T \geq \frac{9m^2(\kappa^2 + \xi^2)}{\alpha^2\epsilon'}$ iterations the expected relative entropy is $\mathbb{E}S(\rho_{\theta^{\text{opt}}} \| \rho_{\theta^T}) \leq \epsilon'$. Note that, depending on the problem-specific parameter δ_0 , and the free parameters κ and ξ , one could be in the other case of Lemma 10. Following the same steps shown here one arrives at a different number of iterations T , which is still polynomial in n .

Since the QBM learning problem is phrased in terms of the distance in the expectation values and we only have obtained a bound on the relative entropy we now relate the two. Using Pinsker's and Jensen's inequalities it follows that

$$\begin{aligned} \mathbb{E}S(\rho_{\theta^{\text{opt}}} \| \rho_{\theta^T}) &\geq \frac{1}{2 \ln 2} \mathbb{E} \|\rho_{\theta^{\text{opt}}} - \rho_{\theta^T}\|_1^2 \\ &\geq \frac{1}{2} (\mathbb{E} \|\rho_{\theta^{\text{opt}}} - \rho_{\theta^T}\|_1)^2 \\ &= \frac{1}{2} (\mathbb{E} \max_{-I < U \leq I} |\text{Tr}[U(\rho_{\theta^{\text{opt}}} - \rho_{\theta^T})]|)^2, \end{aligned} \quad (41)$$

where in the last line we used the variational definition of trace distance. The maximization is over unitary matrices. Let us restrict the maximization to unitary matrices defined as $U_i = \frac{H_i}{\|H_i\|} + i\sqrt{I - \frac{H_i^2}{\|H_i\|^2}}$. These have the property that $H_i = \|H_i\|(U_i + U_i^\dagger)/2$. Therefore,

$$\begin{aligned} \sqrt{2\mathbb{E}S(\rho_{\theta^{\text{opt}}} \| \rho_{\theta^T})} &\geq \mathbb{E} \max_{-I < U \leq I} \left| \text{Tr} \left[\frac{1}{2} (U + U^\dagger) (\rho_{\theta^{\text{opt}}} - \rho_{\theta^T}) \right] \right| \\ &\geq \mathbb{E} \max_i \left| \text{Tr} \left[\frac{1}{2} (U_i + U_i^\dagger) (\rho_{\theta^{\text{opt}}} - \rho_{\theta^T}) \right] \right| \\ &\geq \mathbb{E} \max_i \frac{1}{\|H_i\|} |\text{Tr}[H_i \rho_{\theta^{\text{opt}}}] - \text{Tr}[H_i \rho_{\theta^T}]|. \end{aligned} \quad (42)$$

This implies,

$$\sqrt{2\epsilon'} \geq \frac{1}{\|H_i\|} \mathbb{E} |\text{Tr}[H_i \rho_{\theta^{\text{opt}}}] - \text{Tr}[H_i \rho_{\theta^T}]|, \quad \forall i. \quad (43)$$

To solve the QBM learning problem to precision ϵ we choose $\epsilon' = \frac{\epsilon^2}{2}$ and, since $\|H_i\| = 1$ for Pauli operators, we conclude that

$$\mathbb{E} |\text{Tr}[H_i \eta] - \text{Tr}[H_i \rho_{\theta^T}]| \leq \epsilon, \quad \forall i. \quad (44)$$

□

Achieving a desired precision on the quantum relative entropy

In this section, we study the scenario where the reader is interested in obtaining a certain precision on the quantum relative entropy, rather than on the difference in the expectation values. Again, due to a potential model mismatch, we discuss the relative entropy $S(\rho_{\theta^{\text{opt}}} \| \rho_\theta)$ w.r.t. the optimal QBM $\rho_{\theta^{\text{opt}}}$.

We begin by training the QBM ρ_θ with SGD. Using Theorem 1, we can achieve $|\langle H_i \rangle_\eta - \langle H_i \rangle_{\rho_\theta}| \leq \epsilon$ for all i with polynomial sample complexity. This implies a similar relation w.r.t. the optimal model: $|\langle H_i \rangle_\eta - \langle H_i \rangle_{\rho_{\theta^{\text{opt}}}}| \leq \epsilon$. By

the triangle inequality we have that $|\langle H_i \rangle_{\rho_\theta} - \langle H_i \rangle_{\rho_{\theta^{\text{opt}}}}| \leq 2\epsilon$. Then

$$\begin{aligned}
S(\rho_{\theta^{\text{opt}}} \parallel \rho_\theta) &= \text{Tr}[\rho_{\theta^{\text{opt}}} \log \rho_{\theta^{\text{opt}}}] - \sum_i \theta_i \text{Tr}[\rho_{\theta^{\text{opt}}} H_i] + \log Z_\theta \\
&= \text{Tr}[\rho_{\theta^{\text{opt}}} \log \rho_{\theta^{\text{opt}}}] - \sum_i \theta_i \text{Tr}[\rho_{\theta^{\text{opt}}} H_i] + \log Z_\theta + \sum_i \theta_i \text{Tr}[\rho_\theta H_i] - \sum_i \theta_i \text{Tr}[\rho_\theta H_i] \\
&= \text{Tr}[\rho_{\theta^{\text{opt}}} \log \rho_{\theta^{\text{opt}}}] - \text{Tr}[\rho_\theta \log \rho_\theta] + \sum_i \theta_i (\text{Tr}[\rho_\theta H_i] - \text{Tr}[\rho_{\theta^{\text{opt}}} H_i]).
\end{aligned} \tag{45}$$

Similarly,

$$S(\rho_\theta \parallel \rho_{\theta^{\text{opt}}}) = -\text{Tr}[\rho_{\theta^{\text{opt}}} \log \rho_{\theta^{\text{opt}}}] + \text{Tr}[\rho_\theta \log \rho_\theta] - \sum_i \theta_i^{\text{opt}} (\text{Tr}[\rho_\theta H_i] - \text{Tr}[\rho_{\theta^{\text{opt}}} H_i]). \tag{46}$$

Thus

$$\begin{aligned}
S(\rho_{\theta^{\text{opt}}} \parallel \rho_\theta) &\leq S(\rho_{\theta^{\text{opt}}} \parallel \rho_\theta) + S(\rho_\theta \parallel \rho_{\theta^{\text{opt}}}) \\
&= \sum_i (\text{Tr}[\rho_\theta H_i] - \text{Tr}[\rho_{\theta^{\text{opt}}} H_i]) (\theta_i - \theta_i^{\text{opt}}) \\
&\leq \sum_i |\text{Tr}[\rho_\theta H_i] - \text{Tr}[\rho_{\theta^{\text{opt}}} H_i]| \cdot |\theta_i - \theta_i^{\text{opt}}| \\
&\leq 2\epsilon \|\theta - \theta^{\text{opt}}\|_1.
\end{aligned} \tag{47}$$

To minimize the quantum relative entropy to precision ϵ' , we choose $\epsilon \leq \frac{\epsilon'}{2\|\theta - \theta^{\text{opt}}\|_1}$. This determines the number of SGD iterations via Theorem 1. Note that the number of iterations remains polynomial in the system size n . Finally, we combine this result with Eq. (40) and obtain the implication

$$|\langle H_i \rangle_\eta - \langle H_i \rangle_{\rho_\theta}| \leq \epsilon, \forall i \implies S(\eta \parallel \rho_\theta) - S(\eta \parallel \rho_{\theta^{\text{opt}}}) = S(\rho_{\theta^{\text{opt}}} \parallel \rho_\theta) \leq 2\epsilon \|\theta - \theta^{\text{opt}}\|_1. \tag{48}$$

This proves Eq. (6) in the main text.

Additional numerical results

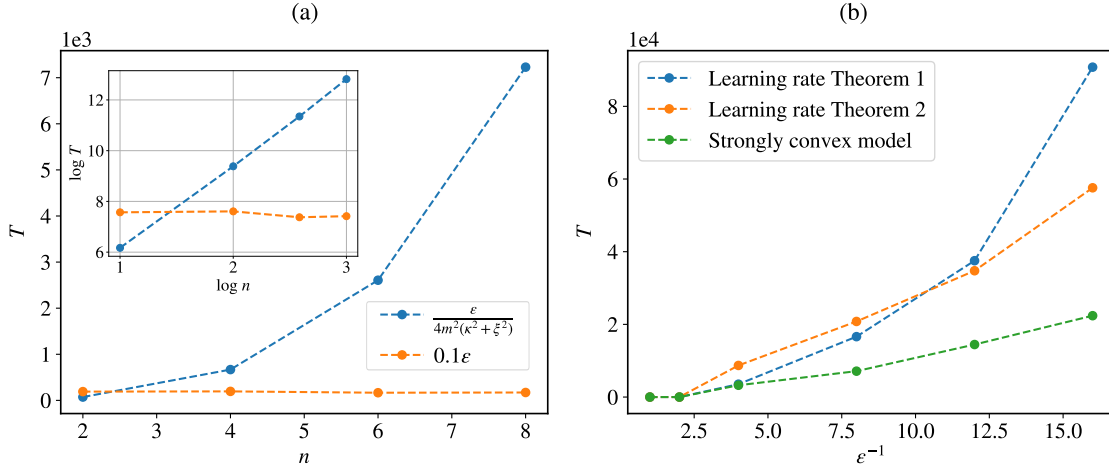
Here we provide additional numerical simulation results which verify the scaling behaviors proven in Theorems 1 and 2. In order to obtain robust estimates one would need to perform many repetitions of SGD training and analyze different settings of the QBM model and target state. However, to keep the computational cost of the simulations manageable, we only consider a few configurations of system size, learning rate, and ansatz. We focus on the target state constructed from the expectation values of the XXZ Heisenberg model (see Supplementary Note 5). Unless stated otherwise, we use a fully-connected QBM model with Hamiltonian ansatz $\mathcal{H}_\theta = \sum_{k=x,y,z} \sum_{i,j>i} \lambda_{ij}^k \sigma_i^k \sigma_j^k + \sum_i \gamma_i^k \sigma_i^k$, where the number of terms is quadratic in the system size, $m \in \mathcal{O}(n^2)$. For each configuration, we run SGD training only once, starting from the maximally mixed state.

Supplementary Figure 2 (a) shows the required number of SGD iterations, T , to reach convergence for different system sizes. Convergence is defined as the first iteration where the QBM matches the target expectation values up to precision $\epsilon = 0.1$, i.e., when the QBM learning problem is solved to precision ϵ . We use a combined model and data expectations precision of $\kappa^2 + \xi^2 = 0.01$ which is larger than $\frac{\epsilon}{2m}$ as required by our theorems. We compare training with the learning rate from Theorem 1 to training with a constant learning rate $\gamma_t = 0.1\epsilon$. We observe that for the learning rate from Theorem 1 (blue line) the number of iterations T grows with the system size. From the log-log plot in the inset we find evidence that the scaling follows a power law with a power (slope) of ≈ 3.3 . This indicates that $T \propto n^{3.3}$, in agreement with the upper bound from Theorem 1 applied to our model, $T \propto m^2 \in \mathcal{O}(n^4)$. When using a constant learning rate we observe a much more rapid convergence (orange line). In particular, T does not seem to increase with the system size. However, there are no proven convergence guarantees for this setting. That is, for very noisy gradients and/or high target precision, a constant learning rate may lead to unstable training. In practice one could opt for an in-between learning rate schedule or use an optimizer with adaptive learning rates, such as the well-known ADAM optimizer.

Supplementary Figure 2 (b) shows the required number of iterations versus the target precision, ϵ^{-1} . For these simulations we use $\kappa^2 + \xi^2 = 0.1$ and system size $n = 6$. Using a fully-connected QBM, we compare training with

the learning rate from Theorem 1 (blue line) to training with the learning rate from Theorem 2 (orange line). Recall that we provided evidence for the strong convexity of the quantum relative entropy of fully-connected QBMs in Supplementary Note 2. If the loss landscape is indeed strongly convex, we can use the learning rate from Theorem 2 and the number of steps scales as $1/\epsilon^2$ otherwise it is expected to scale as $1/\epsilon^4$. Thus, we expect the learning rate from Theorem 2 to outperform the learning rate from Theorem 1. Indeed we observe convergence in a low degree polynomial number of iterations for the orange line, versus a higher degree polynomial number of iterations for the blue line. Since the strong convexity of fully-connected QBMs is unproven, it is hard to justify the choice of learning rate. We additionally tested the scaling using a 1D nearest-neighbor Hamiltonian ansatz for the QBM (green line), which is known to be strongly convex as discussed in Supplementary Note 2. The green line shows a very low degree polynomial scaling when using the learning rate schedule from Theorem 2. These numerical results match the expected scaling from our theorems.

Lastly, here and in the main text we focused on the error in the expectation values. If one is interested in other performance metrics, they can employ Eq. (48) to convert our expectation value plots to relative entropy bounds.



Supplementary Figure 2. (a) Number of SGD iterations required to solve the QBM learning problem for precision $\epsilon = 0.1$ versus system size. The learning rate from Theorem 1 is compared to a constant learning rate. The inset shows the log-log plot of the data in the main panel. (b) Number of SGD iterations to solve the QBM learning problem for system size $n = 6$ versus the target precision ϵ^{-1} . We compare a fully-connected QBM (blue and orange lines) to a 1D nearest-neighbor QBM (green line).

SUPPLEMENTARY NOTE 4: GUARANTEED PERFORMANCE IMPROVEMENT BY PRE-TRAINING

In this appendix, we first prove Theorem 3 in the main text, and then discuss various pre-training models.

Proof of Theorem 3 in the main text

For completeness, we start by restating Theorem 3 from the main text.

Theorem 3 (QBM pre-training). *Assume a target η and a QBM model $\rho_\theta = e^{\sum_{i=1}^m \theta_i H_i} / Z$ for which we like to minimize the relative entropy $S(\eta \| \rho_\theta)$. Initializing $\theta^0 = 0$ and pre-training $S(\eta \| \rho_\theta)$ with respect to any subset of $\tilde{m} \leq m$ parameters guarantees that*

$$S(\eta \| \rho_{\theta^{\text{pre}}}) \leq S(\eta \| \rho_{\theta^0}), \quad (49)$$

where $\theta^{\text{pre}} = [\chi^{\text{pre}}, 0_{m-\tilde{m}}]$ and the vector χ^{pre} of length $\tilde{m}$ contains the parameters for the terms $\{H_i\}_{i=1}^{\tilde{m}}$ at the end of pre-training. More precisely, starting from $\rho_\chi = e^{\sum_{i=1}^{\tilde{m}} \chi_i H_i} / Z$ and minimizing $S(\eta \| \rho_\chi)$ with respect to χ ensures Eq. (49) for any $S(\eta \| \rho_{\chi^{\text{pre}}}) \leq S(\eta \| \rho_{\chi^0})$.

Proof. First, we relate the difference in relative entropy between two parameter vectors in the full space to the difference in relative entropy of the pre-trained parameter space. In particular, for any real parameter vectors $\theta = [\chi, 0_{m-\tilde{m}}]$

and $\theta' = [\chi', 0_{m-\tilde{m}}]$ we have

$$\begin{aligned}
S(\eta\|\rho_\theta) - S(\eta\|\rho_{\theta'}) &= \text{Tr}[\eta \log \rho_{\theta'}] - \text{Tr}[\eta \log \rho_\theta] \\
&= \sum_{i=1}^m (\theta'_i - \theta_i) \text{Tr}[\eta H_i] - \log \text{Tr}[e^{\sum_{i=1}^m \theta'_i H_i}] + \log \text{Tr}[e^{\sum_{i=1}^m \theta_i H_i}] \\
&= \sum_{i=1}^{\tilde{m}} (\chi'_i - \chi_i) \text{Tr}[\eta H_i] - \log \text{Tr}[e^{\sum_{i=1}^{\tilde{m}} \chi'_i H_i}] + \log \text{Tr}[e^{\sum_{i=1}^{\tilde{m}} \chi_i H_i}] \\
&= \text{Tr}[\eta \log \rho_{\chi'}] - \text{Tr}[\eta \log \rho_\chi] \\
&= S(\eta\|\rho_\chi) - S(\eta\|\rho_{\chi'})
\end{aligned} \tag{50}$$

Now using pre-training vectors $\theta^{\text{pre}} = [\chi^{\text{pre}}, 0_{m-\tilde{m}}]$ and $\theta^0 = [\chi^0, 0_{m-\tilde{m}}] = 0$ we see that $S(\eta\|\rho_{\chi^{\text{pre}}}) \leq S(\eta\|\rho_{\chi^0})$ implies $S(\eta\|\rho_{\theta^{\text{pre}}}) \leq S(\eta\|\rho_{\theta^0})$. Thus, any method that finds such a χ^{pre} guarantees Eq. (49). $\square$

While conclusive, the above proof does not provide us with a method to find such a χ^{pre} , i.e., it is agnostic to the specific pre-training method. As a constructive example, let us consider minimizing χ^{pre} with noiseless gradient descent on a subset of $\tilde{m}$ parameters. This means we update the subset parameters as $\chi^t = \chi^{t-1} - \gamma \tilde{\nabla} S(\eta\|\rho_{\chi^{t-1}})$, where $\tilde{\nabla} S(\eta\|\rho_{\chi^{t-1}})$ is the gradient of the subset of parameters, and γ the learning rate. Since S is L -smooth, we can use the descent Lemma 4 to bound the difference in relative entropy of the subset

$$\begin{aligned}
S(\eta\|\rho_{\chi^t}) - S(\eta\|\rho_{\chi^{t-1}}) &\leq \nabla S(\eta\|\rho_{\chi^{t-1}})^T (-\gamma \tilde{\nabla} S(\eta\|\rho_{\chi^{t-1}})) + \frac{L}{2} \|\tilde{\nabla} S(\eta\|\rho_{\chi^{t-1}})\|^2 \\
&= -\gamma \left(1 - \frac{\gamma L}{2}\right) \|\tilde{\nabla} S(\eta\|\rho_{\chi^{t-1}})\|^2.
\end{aligned} \tag{51}$$

Setting $\gamma \leq \frac{2}{L}$ we obtain $S(\eta\|\rho_{\chi^t}) \leq S(\eta\|\rho_{\chi^{t-1}})$. By recursively applying this inequality we obtain a χ^{pre} with $S(\eta\|\rho_{\chi^{\text{pre}}}) \leq S(\eta\|\rho_{\chi^0})$, which by our theorem above ensures Eq. (49). Note that the smoothness L here is the smoothness on the subset of parameters, which can be bounded by $L \leq 2\tilde{m} \max_i \|H_i\|_2^2$.

Pre-training models

Here we discuss possible pre-training models and strategies to optimize them. We focus on the models discussed in the main text: 1) a mean-field model, 2) a Gaussian Fermionic model, and 3) nearest-neighbor quantum spin models. The advantage of the first two models is that they can be trained analytically. While for the nearest-neighbor models, this is not possible, they satisfy the locality assumptions in [4, 6], and hence have a strongly convex relative entropy.

Mean-Field QBM

We define the mean-field QBM by the parameterized Hamiltonian

$$\mathcal{H}_\theta = \sum_i^n \theta_i^x \sigma_i^x + \theta_i^y \sigma_i^y + \theta_i^z \sigma_i^z. \tag{52}$$

Since this Hamiltonian has a simple structure, in which many terms commute, we can find the optimal parameters analytically. First, recall that the QBM expectation values are given by

$$\langle H_i \rangle_{\rho_\theta} = \frac{\partial}{\partial \theta_i} \log \text{Tr}[e^{\mathcal{H}_\theta}] = \frac{\partial}{\partial \theta_i} \log Z_\theta. \tag{53}$$

For the mean-field Hamiltonian, we find

$$Z_\theta = \text{Tr}[e^{\sum_{i=1}^n \theta_i^x \sigma_i^x + \theta_i^y \sigma_i^y + \theta_i^z \sigma_i^z}] = \prod_{i=1}^n \text{Tr}[e^{\theta_i^x \sigma_i^x + \theta_i^y \sigma_i^y + \theta_i^z \sigma_i^z}] = \prod_{i=1}^n 2 \cosh \|\theta_i\|_2, \tag{54}$$

where we have defined $\|\theta_i\|_2 = \sqrt{\theta_i^{x^2} + \theta_i^{y^2} + \theta_i^{z^2}}$. Here we use the commutativity of single qubit operators in the first equality and expand the exponent for the second equality. We therefore get

$$\log Z_\theta = \sum_{i=1}^n \log 2 \cosh \|\theta_i\|_2. \quad (55)$$

From which the derivative follows as

$$\frac{\partial}{\partial \theta_i^{x,y,z}} \log Z_\theta = \frac{\theta_i^{x,y,z}}{\|\theta_i\|_2} \tanh \|\theta_i\|_2. \quad (56)$$

In order to find the optimal QBM parameters for each qubit, i , we then solve the three coupled equations,

$$\frac{\theta_i^{x,y,z}}{\|\theta_i\|_2} \tanh \|\theta_i\|_2 = \langle \sigma_i^{x,y,z} \rangle_\eta, \quad (57)$$

which corresponds to setting the QBM derivative in Eq. (4) in the main text to zero. From the strict convexity of the relative entropy, we know this has one unique solution provided the target expectation values, $\langle \sigma_i^{x,y,z} \rangle_\eta$ form a consistent set, i.e., it comes from a density matrix. We can find the solution by squaring the three equations, and adding them together, giving

$$\|\theta_i\|_2 = \tanh^{-1} \left(\sqrt{\langle \sigma_i^x \rangle_\eta^2 + \langle \sigma_i^y \rangle_\eta^2 + \langle \sigma_i^z \rangle_\eta^2} \right). \quad (58)$$

Here we used that the argument of the tanh is always positive. Substituting this into Eq. (57) we then find the closed-form solution of the QBM parameters

$$\theta_i^{x,y,z} = \langle \sigma_i^{x,y,z} \rangle_\eta \frac{\tanh^{-1} \left(\sqrt{\langle \sigma_i^x \rangle_\eta^2 + \langle \sigma_i^y \rangle_\eta^2 + \langle \sigma_i^z \rangle_\eta^2} \right)}{\sqrt{\langle \sigma_i^x \rangle_\eta^2 + \langle \sigma_i^y \rangle_\eta^2 + \langle \sigma_i^z \rangle_\eta^2}}. \quad (59)$$

In practice, the optimal parameters for an arbitrary mean-field QBM can be obtained by numerically evaluating this expression for the given target expectation values.

Gaussian Fermionic QBM

The Gaussian Fermionic QBM has a parameterized, quadratic, Fermionic Hamiltonian

$$\mathcal{H}_\theta = \vec{C}^\dagger \tilde{\Theta} \vec{C} \equiv \sum_{i,j} \tilde{\Theta}_{ij} \vec{C}_i^\dagger \vec{C}_j. \quad (60)$$

Here, $\vec{C}^\dagger = [c_1^\dagger, \dots, c_n^\dagger, c_1, \dots, c_n]$ is a vector containing n Fermionic mode creation and annihilation operators, which satisfy the Fermionic commutation relations $\{c_i, c_j^\dagger\} = \delta_{i,j}$ and $\{c_i, c_j\} = 0$. These Fermionic operators can be expressed as strings of Pauli operators by the Jordan-Wigner transformation. $\tilde{\Theta}$ is the $2n \times 2n$ dimensional matrix containing the QBM model parameters θ , which can be identified as a Fermionic single-particle Hamiltonian. Note that this matrix needs to be Hermitian, and since terms like $c_1^\dagger c_1^\dagger$ are zero, it has in total n^2 free parameters.

In order to find the optimal parameters, we use the fact that the single-particle correlation matrix with entries $[\Gamma_{\rho_\theta}]_{ij} = \langle \vec{C}_i^\dagger \vec{C}_j \rangle_{\rho_\theta}$ contains sufficient information to compute all possible properties of the Gaussian quantum system. This includes all possible observables (via Wick's theorem), entanglement measures, and also sampling from ρ_θ [10]. In particular, the Gaussian Fermionic QBM gradient reduces to the difference in the correlation matrices of the target and the model

$$\frac{\partial S}{\partial \tilde{\Theta}_{ij}} = \langle \vec{C}_i^\dagger \vec{C}_j \rangle_{\rho_\theta} - \langle \vec{C}_i^\dagger \vec{C}_j \rangle_\eta. \quad (61)$$

We can solve this by first determining the target expectation values $\langle \vec{C}_i^\dagger \vec{C}_j \rangle_\eta$ and setting $\langle \vec{C}_i^\dagger \vec{C}_j \rangle_{\rho_\theta^*} = \langle \vec{C}_i^\dagger \vec{C}_j \rangle_\eta$. Then we use the fact that the Hamiltonian of a Gaussian Fermionic system can be written in the eigenbasis of the correlation matrix as

$$H_\eta = \frac{1}{2} W_\eta \sigma^{-1} (\Lambda_\eta) W_\eta^\dagger, \quad (62)$$

where W_η and Λ_η is given by the eigendecomposition $\Gamma_\eta = W_\eta \Lambda_\eta W_\eta^\dagger$, and $\sigma^{-1}(X)$ the inverse sigmoid function. Thus, we (numerically) diagonalize Γ_η and set the optimal Gaussian Fermionic QBM Hamiltonian equal to $H_{\theta^*} = \frac{1}{2} W_\eta \sigma^{-1}(\Lambda_\eta) W_\eta^\dagger$. Since the eigen decomposition of a Hermitian matrix is unique, we find one unique solution. This is in agreement with the strict convexity of the quantum relative entropy.

Geometrically-Local QBM

The last type of restricted QBM model we discuss is the geometrically-local QBM. We focus on nearest-neighbor models on some d -dimensional lattice, for example a one-dimensional chain where each Pauli operator only acts on two neighboring qubits. In full generality, the geometrically-local QBM Hamiltonian is given by

$$\mathcal{H}_\theta = \sum_{k=x,y,z} \sum_{\langle i,j \rangle} \lambda_{ij}^k \sigma_i^k \sigma_j^k + \sum_i^n \gamma_i^k \sigma_i^k, \quad (63)$$

where the sum is over nearest-neighbor sites $\langle i, j \rangle$ of the lattice with periodic boundary conditions. In the main text, we consider for example a $d = 1$ lattice (a ring), and a $d = 2$ square lattice.

In order to use these models for pre-training, we train them with exact gradient descent on the relative entropy until a fixed precision is reached. Importantly, as these Hamiltonians only have $m \in \mathcal{O}(n)$ terms and a finite interaction range, Refs. [4, 6] show that the quantum relative entropy is strongly convex. Therefore, the optimization is guaranteed to converge quickly to the global optimum, recall Theorem 2. However, this requires obtaining Gibbs state expectation values of geometrically local Hamiltonians. This can be done with a quantum computer or, potentially, with classical tensor networks [11, 12].

SUPPLEMENTARY NOTE 5: CONSTRUCTION OF THE TARGET STATE EXPECTATION VALUES

In this appendix, we review how to embed classical data into a target density matrix η . We will follow the approach in [13] for quantum spin models. We also show how to extend this formalism to the Fermionic quantum models needed for the pre-training of our Gaussian Fermionic QBM. Lastly, we describe the two different targets used in our numerical simulations.

Classical data encoding

Following [13], one way to encode a classical data-set consisting of N bit strings $\{\vec{s}^\mu \in \{0, 1\}^n\}_{\mu=1}^M$ into a quantum state is by defining the pure state

$$\eta = |\psi\rangle\langle\psi|, \quad (64)$$

with

$$|\psi\rangle = \sum_{\vec{s} \in \{0,1\}^n} \sqrt{q(\vec{s})} |\vec{s}\rangle. \quad (65)$$

Here $q(\vec{s}) = \frac{1}{N} \sum_{\mu=1}^M \delta_{\vec{s}, \vec{s}^\mu}$ is the classical empirical probability for bitstring $\vec{s}$, and $|\vec{s}\rangle$ is a computational basis state indexed by $\vec{s}$. The $q(\vec{s})$ can be found by counting the bitstrings in the data set $\{\vec{s}^\mu\}$. From $|\psi\rangle$ one can compute expectation values such as

$$\langle\psi| \sigma_i^z |\psi\rangle = \sum_{\vec{s} \in \{0,1\}^n} q(\vec{s}) \vec{s}_i \quad (66)$$

for the Pauli spin operator σ_i^z . This can be efficiently computed classically for a polynomially sized data set, i.e., for polynomially many $\vec{s}^\mu$. Computing such expectation values from η is possible for all 1- and 2-local Pauli operators as shown in Ref. [13].

We now show that we can generalize this encoding to Fermionic QBMs, i.e., the terms in the Hamiltonian ansatz consist of Fermionic creation $c_i^\dagger$ and annihilation operators c_i . We define $|\vec{s}\rangle$ to be equal to the Fermionic Fock basis. This is analogous to the computational basis in the spin-picture (by the Jordan-Wigner transformation), but the

bit-strings $\{\vec{s}^\mu \in \{0,1\}^n\}_{\mu=1}^M$ in the data set should now be interpreted as occupation-number vectors of Fermions. Note that the occupation-number basis is defined by the eigenstates of the Fermionic number operator $\sum_i c_i^\dagger c_i$.

The creation and annihilation operators act on the Fock-basis states as follows

$$c_i^\dagger |\vec{s}\rangle = (1 - \vec{s}_i) |\vec{s} + \vec{\delta}_i\rangle, \quad c_i |\vec{s}\rangle = \vec{s}_i |\vec{s} - \vec{\delta}_i\rangle, \quad (67)$$

where $\vec{\delta}_i$ is the unit bit-string with a 1 at position i and zeros everywhere else. With these relations, we can derive the required expectation values for the target η to train the (Gaussian) Fermionic QBM

$$\begin{aligned} \langle \psi | c_i^\dagger c_i | \psi \rangle &= \sum_{\vec{s} \in \{0,1\}^n} q(\vec{s}) \vec{s}_i, \\ \langle \psi | c_i^\dagger c_j | \psi \rangle &= \sum_{\vec{s} \in \{0,1\}^n} \sqrt{q(\vec{s}) q(F_i F_j \vec{s})} (1 - \vec{s}_i) \vec{s}_j, \\ \langle \psi | c_i^\dagger c_j^\dagger | \psi \rangle &= \sum_{\vec{s} \in \{0,1\}^n} \sqrt{q(\vec{s}) q(F_i F_j \vec{s})} (1 - \vec{s}_i) (1 - \vec{s}_j), \quad i \neq j \\ \langle \psi | c_i c_j | \psi \rangle &= \sum_{\vec{s} \in \{0,1\}^n} \sqrt{q(\vec{s}) q(F_i F_j \vec{s})} \vec{s}_i \vec{s}_j, \quad i \neq j \\ \langle \psi | c_i | \psi \rangle &= \sum_{\vec{s} \in \{0,1\}^n} \sqrt{q(\vec{s}) q(F_i \vec{s})} \vec{s}_i \\ \langle \psi | c_i^\dagger | \psi \rangle &= \sum_{\vec{s} \in \{0,1\}^n} \sqrt{q(\vec{s}) q(F_i \vec{s})} (1 - \vec{s}_i), \end{aligned} \quad (68)$$

where F_i flips the Fermion occupation number (from occupied to unoccupied and vice versa) of index i in the vector $\vec{s}$.

Data used for the numerical simulations

For the numerical simulations, we use two different targets η : a target constructed from a quantum source, and a classical data set embedded into η using the encoding above. For the quantum source we use the XXZ model Hamiltonian

$$\mathcal{H}_{\text{XXZ}} = \sum_{i=1}^{n-1} J (\sigma_i^x \sigma_{i+1}^x + \sigma_i^y \sigma_{i+1}^y) + \Delta \sigma_i^z \sigma_{i+1}^z + \sum_{i=1}^n h_z \sigma_i^z. \quad (69)$$

Here J and Δ are parameters describing the Heisenberg interactions between the quantum spins on a one-dimensional lattice, and h_z is the strength of an external magnetic field. We set $\eta = \frac{e^{\mathcal{H}_{\text{XXZ}}}}{Z}$ with $J = -0.5$, $\Delta = -0.7$, and $h_z = -0.8$, and compute the expectation values $\langle H_i \rangle_\eta$ classically. This is intractable in general, but our aim is to replicate the scenario in which the expectation values are measured experimentally. For example, from a state prepared on a quantum device.

For the classical source, we use the classical salamander retina data set given in Ref. [14]. This data set consists of bit-string data recordings of different features of the response of cells in the salamander retina. We select the first 8 features and trim the data to the first 10 data recordings. We then construct the expectation values $\langle H_i \rangle_\eta$ from the procedure outlined above.

SUPPLEMENTARY REFERENCES

- [1] G. Garrigos and R. M. Gower, Handbook of convergence theorems for (stochastic) gradient methods, arXiv preprint arXiv:2301.11235 [10.48550/arXiv.2301.11235](#) (2023).
- [2] M. B. Hastings, Quantum belief propagation: An algorithm for thermal quantum systems, *Phys. Rev. B* **76**, 201102 (2007).
- [3] H. E. Haber, [Notes on the matrix exponential and logarithm](#) (2018).
- [4] A. Anshu, S. Arunachalam, T. Kuwahara, and M. Soleimanifar, Sample-efficient learning of interacting quantum systems, *Nature Physics* **17**, 931 (2021).
- [5] E. T. Jaynes, Information Theory and Statistical Mechanics, *Physical Review* **106**, 620 (1957).
- [6] J. Haah, R. Kothari, and E. Tang, Learning quantum Hamiltonians from high-temperature Gibbs states and real-time evolutions, *Nature Physics* **20**, 1027–1031 (2024).
- [7] A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind, Automatic differentiation in machine learning: a survey, *Journal of Machine Learning Research* **18**, 1 (2018).
- [8] A. Khaled and P. Richtárik, Better theory for SGD in the nonconvex world, arXiv preprint arXiv:2002.03329 [10.48550/arXiv.2002.03329](#) (2020).
- [9] H.-Y. Huang, R. Kueng, and J. Preskill, Information-theoretic bounds on quantum advantage in machine learning, *Physical Review Letters* **126** (2021).
- [10] J. Surace and L. Tagliacozzo, Fermionic Gaussian states: an introduction to numerical approaches, *SciPost Phys. Lect. Notes* , 54 (2022).
- [11] T. Kuwahara, A. M. Alhambra, and A. Anshu, Improved thermal area law and quasilinear time algorithm for quantum Gibbs states, *Phys. Rev. X* **11**, 011047 (2021).
- [12] A. M. Alhambra and J. I. Cirac, Locally accurate tensor networks for thermal states and time evolution, *PRX Quantum* **2**, 040331 (2021).
- [13] H. J. Kappen, Learning quantum models from quantum or classical data, *Journal of Physics A: Mathematical and Theoretical* **53**, 214001 (2020).
- [14] G. Tkačik, O. Marre, D. Amodei, E. Schneidman, W. Bialek, and M. J. Berry, II, Searching for collective behavior in a large network of sensory neurons, *PLOS Computational Biology* **10**, 1 (2014).