Learning quantum models from quantum or classical data

H.J. Kappen

Donders Institute, Department of Biophysics, Radboud University, the Netherlands

E-mail: `b.kappen@science.ru.nl`

17 January 2020

Abstract. In this paper, we address the problem how to represent a classical data distribution in a quantum system. The proposed method is to learn quantum Hamiltonian that is such that its ground state approximates the given classical distribution. We review previous work on the quantum Boltzmann machine (QBM) [1, 2] and how it can be used to infer quantum Hamiltonians from quantum statistics. We then show how the proposed quantum learning formalism can also be applied to a purely classical data analysis. Representing the data as a rank one density matrix introduces quantum statistics for classical data in addition to the classical statistics. We show that quantum learning yields results that can be significantly more accurate than the classical maximum likelihood approach, both for unsupervised learning and for classification. The data density matrix and the QBM solution show entanglement, quantified by the quantum mutual information I. The classical mutual information in the data $I_c \leq I/2 = C$, with C maximal classical correlations obtained by choosing a suitable orthogonal measurement basis. We suggest that the remaining mutual information $Q = I/2$ is obtained by non orthogonal measurements that may violate the Bell inequality. The excess mutual information $I - I_c$ may potentially be used to improve the performance of quantum implementations of machine learning or other statistical methods.

1. Introduction

Current successes in machine learning [3, 4, 5] has ignited interesting new connections between machine learning and quantum physics, loosely referred to as quantum machine learning. The transfer of ideas goes in both directions. Quantum annealing [6, 7] has been successfully applied to hard optimization problems in machine learning [8, 9]. Machine learning methods also find useful applications in quantum physics, such as characterizing the ground state of a quantum Hamiltonian [10] or to learn different phases of matter [11].

In addition, there are efforts to exploit quantum mechanical features for learning and coding [12]. In particular, attempts have been made to extend probability calculus to quantum density operators [13, 14, 15]. Recently, [1, 2] have proposed learning methods for density matrices called the quantum Boltzmann machine (QBM). One approach is to maximize the classical likelihood $L = \sum_s q(s) \log p(s|w)$, with p the diagonal of the density matrix ρ. As the authors remark, this approach faces difficulties because the gradients of the likelihood are hard to evaluate (see Appendix A). For this reason, they introduce a lower bound on the likelihood using the Golden-Thomson inequality and maximize this bound. But this has the disadvantage that parameters of quantum statistics cannot be learned.

[1] introduce another approach to learning the QBM, which is to minimize the relative entropy between a model density matrix ρ and a target density matrix η ‡. This is equivalent to maximizing a quantum likelihood $L = \text{Tr}\,\eta \log \rho$. This approach does not suffer from the above difficulties. The quantum likelihood generalizes maximum likelihood learning to density matrices: In the case that the model density matrix is diagonal, the quantum likelihood reduces to the classical likelihood.

The aim of this paper is to show how the quantum likelihood method can be used to learn classical machine learning problems such as unsupervised and supervised learning. We restrict ourselves to spin systems without latent variables, ie. there is data statistics on all spins. For extensions to latent variable models see [16]. For unsupervised learning problems, data can be represented as a classical distribution $q(s)$. By generalizing q to a density matrix, the classical statistics (such as expectation values and correlations) are augmented with quantum statistics. The quantum statistics provides features of the data that are not available from low order classical statistics and may be useful for learning.

The paper is organized as follows. In section 2 we review the unsupervised learning problem and the classical Boltzmann machine. In section 3 we review the quantum likelihood and the derivation of QBM learning rule. In section 4 we present numerical results that demonstrate how to apply the QBM to quantum tomography, ie. to recover the quantum Hamiltonian or the ground state wave function from measured quantum statistics. We consider the quantum anti-ferromagnetic Heisenberg model and a fully

‡ This idea was independently proposed in the arXiv preprint arXiv:1803.11278, which is the draft version of the current article.

connected quantum spin glass model. For finite temperature, this problem is strictly convex and has a unique solution. In this case the Hamiltonian can be fully recovered from the quantum statistics. For zero temperature the solution is degenerate. The learned Hamiltonian still reproduces all statistics correctly, but the solution is not unique. For low temperature, the convergence of the learning algorithm slows down dramatically. The results in section 4 confirm previous findings reported in [1].

Sections 5 and 6 contain the main novel material of this paper. In section 5 we apply the QBM to learn classical statistical problems, such as unsupervised learning and classification. We propose to represent the target data distribution by a rank one density matrix η, which we call the data state. For unsupervised learning, we show that the QBM with only pairwise interactions can learn the parity problem exactly, whereas the BM cannot learn this problem. In the case of supervised learning, we learn a density matrix over the joint input-output state space and construct the classifier by conditioning the density matrix on a classical input state. Surprisingly, in this way a large number of hard nonlinear classification problems can be learned that cannot be learned by a classical BM. This approach differs significantly from the recently proposed quantum perceptron [17] which learns a density matrix on the output qubit, whose statistics are conditioned on classical input states.

The data state η is a much more complex object than the classical probability distribution q from which it is derived. It displays full quantum features such as non locality and entanglement. In section 6 we explain what these quantum features mean for classical data. We can measure η in different measurement bases and the classical distribution $q(s)$ corresponds to a particular choice of basis. In general, the quantum mutual information $I(\eta)$ between sub systems is larger than the classical mutual information $I_c(q)$. We conjecture that $I_c(q) \leq I(\eta)/2$ in general, which we support by numerical experiments and show that $I_c(q) = I(\eta)/2$ when the sub systems are deterministically related. By optimizing the orthogonal measurement basis, the mutual information can be maximized to $I_c(\tilde{q}) = I(\eta)/2 = C$ where C denotes the maximal classical mutual information. The information is called classical because the statistics of the measurement outcomes can be described by a classical joint probability distribution $\tilde{q}$ on local variables. The remaining quantum information $Q = I - C$ resides in non-local features of the state η. In section 6.1 we illustrate the excess quantum information Q and the violation of the Bell inequality for the simplest possible example of two fully correlated binary variables.

2. Classical learning

We first briefly review classical learning. Consider a data set of samples $s^\mu, \mu = 1, \ldots N$, where each sample s^μ is a vector of length n. The data set can be written as a so-called empirical probability distribution

$$q(s) = \frac{1}{N} \sum_{\mu=1}^{N} \delta_{s,s^\mu} \tag{1}$$

Classical learning can be defined to find a model distribution $p(s)$ that is as close as possible to q, where the 'distance' is defined as the relative entropy or Kullback-Leibler divergence between the distributions q and p

$$KL(q,p) = \sum_s q(s) \log \frac{q(s)}{p(s)} \tag{2}$$

Minimizing $KL(q,p)$ with respect to p is equivalent to maximizing the classical likelihood

$$L_c = \sum_s q(s) \log p(s) \tag{3}$$

In the case of the classical Boltzmann machine learning problem, the state space consists of all vectors $s = (s_1, \ldots, s_n)$ with $s_i = \pm 1$ binary spin variables and p is a Boltzmann distribution

$$p(s|w) = \frac{e^{H(s|w)}}{Z} \qquad Z(w) = \sum_s e^{H(s|w)} \tag{4}$$

with $H(s|w) = \sum_r H_r(s) w_r$ linear in w_r. $H_r(s)$ are interaction terms involving typically a small subset of the components of s such for instance $s_i, s_i s_j, \ldots$. For p of the form Eq. 4 the likelihood Eq. 3 becomes

$$L_c(w) = \langle H \rangle_q - \log Z \tag{5}$$

where $\langle H \rangle_q$ is the expectation value of H with respect to the empirical distributions q. The maximization can be performed by gradient ascent on L_c:

$$\Delta w_r \propto \frac{\partial L_c}{\partial w_r} = \langle H_r \rangle_q - \langle H_r \rangle_p \tag{6}$$

where we used that $\frac{\partial \log Z}{\partial w_r} = \langle H_r \rangle_p$ and $\langle \ldots \rangle_p$ is expectation with respect to the Boltzmann distribution p. Learning stops when the gradients are zero, ie. when the statistics defined by H_r are equal: $\langle H_r \rangle_q = \langle H_r \rangle_p$. A common example is when $H(s|w) = \sum_i w_i s_i + \sum_{i>j} w_{ij} s_i s_j$, the learning rule becomes

$$\Delta w_i \propto \langle s_i \rangle_q - \langle s_i \rangle_p \qquad \Delta w_{ij} \propto \langle s_i s_j \rangle_q - \langle s_i s_j \rangle_p \tag{7}$$

Eq. 7 is the well known Boltzmann Machine learning rule [18].

3. Quantum learning

In the quantum case, one represents both the data and the model as a density matrix instead of a probability vector. In this section, we review this idea that was first introduced in [1]. A density operator ρ is a Hermitian positive semi definite operator. In the case of binary spins, ρ can be represented on an orthogonal basis $|s = (s_1, \ldots, s_n)\rangle$ as a matrix of size $2^n \times 2^n$ with components $\rho(s, s')$. We also represent the wave function $|\psi\rangle$ of a quantum system on the same basis by its component vector $\psi(s) = \langle s|\psi \rangle$ (see Appendix B). We use this basis dependent representation throughout the paper. In section 6 we discuss measurement in other bases. ρ has trace one $\text{Tr}\,\rho = 1$ and real eigenvalues $\lambda_s \geq 0$ and $\sum_s \lambda_s = 1$. The density matrix is a generalization of a probability distribution and reduces to the latter when it is diagonal.

The von Neumann, or quantum, entropy of a density matrix ρ is defined as

$$h(\rho) = -\operatorname{Tr}(\rho \log \rho) \tag{8}$$

with $\log \rho$ the matrix logarithm of ρ and $\rho \log \rho$ a matrix product. It is easy to show that $h(\rho) = -\sum_s \lambda_s \log \lambda_s$, with $\lambda_s \geq 0$ the eigenvalues of ρ. The entropy is maximal when all λ_s are equal. The minimal entropy $h(\rho) = 0$ when $\lambda_s = \delta_{s,s^*}$ for some state s^*. In this case ρ is a rank one matrix and can be written as $\rho = \psi \psi^\dagger$ (or $\rho(s, s') = \psi(s)\psi^*(s')$) and is called a pure state. When ρ is diagonal: $\rho(s, s') = p(s)\delta_{s,s'}$, the quantum entropy is equal to the classical entropy $h(\rho) = h_c(p) = -\sum_s p(s) \log p(s)$.

The notion of expectation value for probability distributions is generalized for density matrices. The expectation value of a matrix A, called an observable, is defined as $\langle A \rangle_\rho = \operatorname{Tr}(A\rho)$, with $A\rho$ the matrix product. When A is a Hermitian matrix, $\langle A \rangle_\rho$ is real. When A is a diagonal matrix, $\langle A \rangle_\rho = \langle A \rangle_p$ with p the diagonal of ρ and $\langle A \rangle_p$ the classical expectation value. In this case we call $\langle A \rangle_\rho$ a classical statistic (in the basis $|s\rangle$) . When A is a non diagonal matrix, $\langle A \rangle_\rho$ are statistics of ρ that do not have a classical analogue. We call these quantum statistics.

The relative entropy between density matrices η and ρ is defined as [19]

$$S(\eta, \rho) = \operatorname{Tr}(\eta \log \eta) - \operatorname{Tr}(\eta \log \rho) \tag{9}$$

One can show that $S \geq 0$ which follows from Klein's inequality [19]. Eq. 9 generalizes the classical relative entropy Eq. 2 to density matrices. When η is the density matrix of the data and ρ is the model density matrix, the quantum learning problem is to find ρ that minimizes S. This is equivalent to maximizing the quantum likelihood

$$L(\rho) = \operatorname{Tr}(\eta \log \rho) \tag{10}$$

As an immediate generalization of the classical BM case discussed in section 2, we consider model density matrices of the form

$$\rho = \frac{1}{Z}e^H \qquad Z = \operatorname{Tr}\left(e^H\right) \qquad H = \sum_r H_r w_r \tag{11}$$

with H the quantum Hamiltonian and e^H is the matrix exponential of H. H and H_r are Hermitian matrices and $w = \{w_r, r = 1, \ldots\}$ are real parameters. The model Eq. 11 is referred to as the quantum Boltzmann machine (QBM) [1, 2].

The quantum likelihood Eq. 10 for the QBM Eq. 11 is

$$L(w) = \langle H \rangle_\eta - \log Z \tag{12}$$

Learning is defined as gradient ascent on the quantum likelihood Eq. 12. One can compute $\frac{\partial}{\partial w_r}e^H$ through the Trotter formula $e^H = \lim_{m \to \infty}\left(e^{H/m}\right)^m$:

$$\frac{\partial}{\partial w_r}e^H = \lim_{m \to \infty}\frac{H_r}{m}\underbrace{e^{H/m}\ldots e^{H/m}}_{m \text{ terms}} + \ldots$$

$$+ \underbrace{e^{H/m}\ldots e^{H/m}}_{a \text{ terms}}\frac{H_r}{m}\underbrace{e^{H/m}\ldots e^{H/m}}_{m-a \text{ terms}} + \ldots + \ldots \underbrace{e^{H/m}\ldots e^{H/m}}_{m \text{ terms}}\frac{H_r}{m}$$

$$= \int_0^1 dt\, e^{Ht}H_r e^{H(1-t)}$$

Thus $\frac{\partial}{\partial w_r} \mathrm{Tr}\left(e^H\right) = \mathrm{Tr}\left(H_r e^H\right)$ and $\frac{\partial}{\partial w_r} \log Z \langle H_r\rangle_\rho$. Since $\langle H\rangle_\eta = \sum_r w_r \langle H_r\rangle_\eta$ we get

$$\Delta w_r \propto \frac{\partial}{\partial w_r} L = \langle H_r\rangle_\eta - \langle H_r\rangle_\rho$$

For the rest of the paper we consider binary quantum spin systems with Hamiltonian

$$H = \sum_{i=1}^n \sum_{k=x,y,z} w_i^k \sigma_i^k + \sum_{i=1,j>i}^n \sum_{k=x,y,z} w_{ij}^k \sigma_i^k \sigma_j^k \tag{13}$$

$\sigma_i^{x,y,z}$ are Pauli spin 1/2 operators (see Appendix B). For this Hamiltonian, the learning rule Eq. 13 becomes

$$\Delta w_i^k = \epsilon \left(\left\langle \sigma_i^k \right\rangle_\eta - \left\langle \sigma_i^k \right\rangle_\rho\right) \qquad \Delta w_{ij}^k = \epsilon \left(\left\langle \sigma_i^k \sigma_j^k \right\rangle_\eta - \left\langle \sigma_i^k \sigma_j^k \right\rangle_\rho\right) \tag{14}$$

with $k = x, y, z$ and $\epsilon > 0$ the learning rate. The QBM reduces to the classical BM when k takes only the value $k = z$ in Eqs. 13 and 14.

4. Learning a quantum Hamiltonian

In this section we provide some examples of learning a density matrix of an unknown quantum system from observed quantum statistics. The results in this section confirm previous findings reported in [1]. As a first example, we consider the anti-ferromagnetic Heisenberg model in 1 dimension with true Hamiltonian H given by Eq. 13 with couplings $w_{ij}^{x,y,z} = -1$ for nearest neighbors and $w_{ij}^{x,y,z} = 0$ otherwise (Figure 1a, top row) and external fields $w_i^{x,y,z} = 0$. From H, the data density matrix $\eta_\beta = \frac{1}{Z} e^{\beta H}$ is constructed for $\beta = 1, 2, \infty$ and the quantum statistics $\left\langle \sigma_i^k \right\rangle_{\eta_\beta} = 0$ and $\left\langle \sigma_i^k \sigma_j^k \right\rangle_{\eta_\beta}$ (shown for $\beta = \infty$ in Figure 1a, second row) are computed. These statistics are used to train the QBM using the learning rule Eq. 14 that minimizes the relative entropy $S(\eta, \rho)$ between the data density matrix η and the model density matrix ρ (Figure 1b). Learning stops when the change in S approaches machine precision or when 2000 iterations are reached. For $\beta = 1, 2$, learning converges fast to the optimal solution ($S(\eta_{\beta=1}, \rho) = 2 \times 10^{-13}$ and $S(\eta_{\beta=2}, \rho) = 1 \times 10^{-12}$) and the Hamiltonian parameters are accurately reconstructed (RMS error is 5×10^{-7} and 2×10^{-6}, respectively). For all β, the learned density matrix accurately models the quantum statistics since this is the learning fixed point in Eq. 14 (for $\beta = \infty$ compare Figure 1a, second and fourth row).

For lower temperature, learning becomes increasingly difficult. For the $\beta = \infty$ learning problem, $\eta_{\beta=\infty} = \psi\psi^\dagger$ is a rank one density matrix and $\rho \propto e^{\sum_r w_r H_r} \to \phi\phi^\dagger$ slowly approaches a rank one density matrix as the couplings w_r diverge during learning ($S(\eta_{\beta=\infty}, \rho) = 5 \times 10^{-6}$ after 2000 iterations). Figure 1c shows a scatter plot of ϕ versus ψ which has the same small error. However, the Hamiltonian parameters are inaccurately reconstructed. In order to compare the learned parameters with their true values, we estimate an effective β_{eff} from the learned parameters and define the RMS error as $\min_{\beta_{\mathrm{eff}}} \sum_r (\beta_{\mathrm{eff}} w_r^{\mathrm{true}} - w_r)^2$. We obtain $\beta_{\mathrm{eff}} = 5.6$ and the RMS error is 0.18. This error is large, but still sufficiently small to detect the zero and nonzero couplings (compare Figure 1a, first and third row).
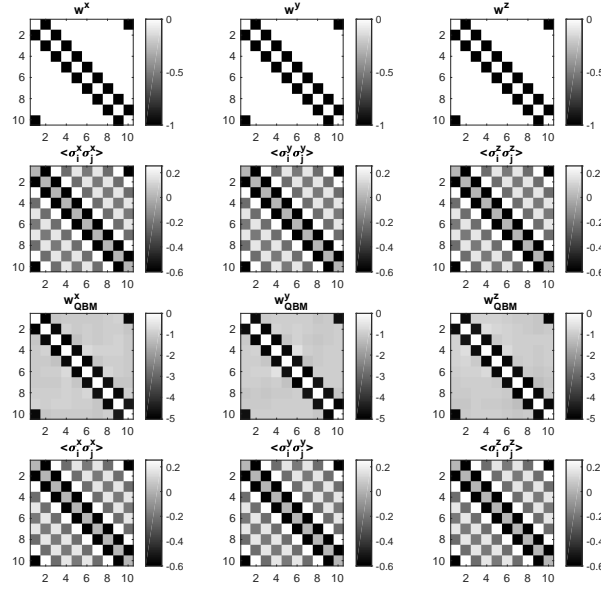
(a) Hamiltonian couplings and pair wise statistics.
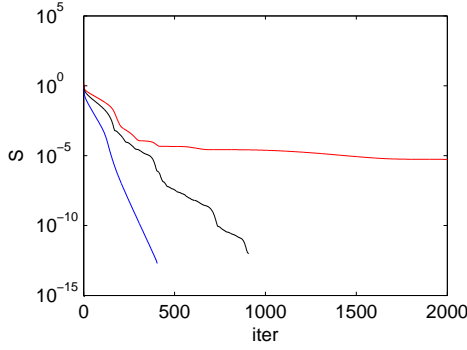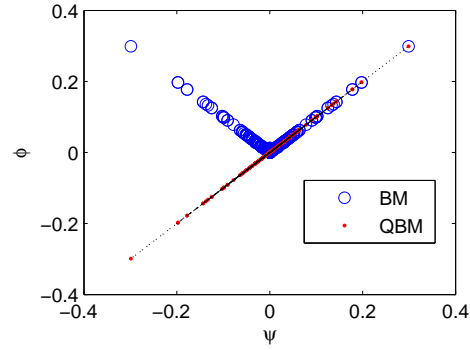

(b) Relative entropy versus learning iteration.

(c) Ground state wave function ϕ of the learned QBM and classical BM distribution.

Figure 1. (Color online) Quantum learning of the one-dimensional anti ferromagnetic Heisenberg chain of $n = 10$ spins with periodic boundary conditions. (a) Top row: The Hamiltonian Eq. 13 has couplings $w_{ij}^{x,y,z} = -1$ for nearest neighbors and $w_{ij}^{x,y,z} = 0$ otherwise and external fields $w_i^{x,y,z} = 0$. Second row: The quantum statistics $\left\langle \sigma_i^k \sigma_j^k \right\rangle_{\eta_\beta}, k = x, y, z$ for $\beta = \infty$. Third row: The QBM couplings $w_{ij}^{x,y,z}$ after converged learning from the $\beta = \infty$ statistics. Fourth row: The QBM statistics $\left\langle \sigma_i^k \sigma_j^k \right\rangle_\rho, k = x, y, z$. (b) Relative entropy $S(\eta, \rho)$ versus learning iteration for $\beta = 1$ (blue), $\beta = 2$ (black) and $\beta = \infty$ (red). (c) Scatter plot of the 2^n components of the ground state wave function $\phi(s)$ of the learned quantum Hamiltonian versus the components of the true ground state wave function $\psi(s)$ (red). For comparison $\sqrt{p_{\text{bm}}(s)}$ versus $\psi(s)$ with p_{bm} the solution of the classical BM (blue).

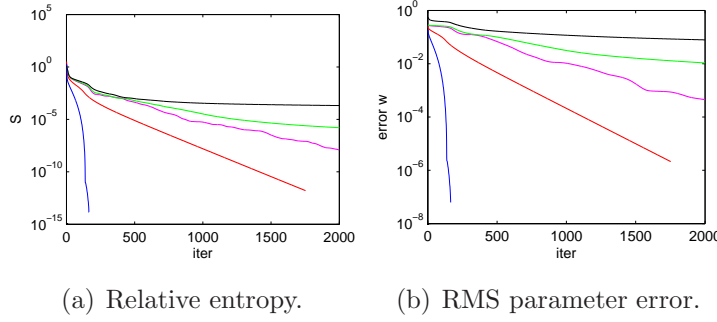(a) Relative entropy.   (b) RMS parameter error.

Figure 2. (Color online) Quantum learning of fully connected quantum spin glass Hamiltonian of $n = 10$ spins with random couplings. (a) Relative entropy $S(\eta, \rho)$ versus learning iteration for $\beta = 0.5$ (blue), $\beta = 1$ (red), $\beta = 3$ (magenta), $\beta = 5$ (green) and $\beta = \infty$ (black). (b) RMS error of the estimated Hamiltonian parameters $w_i^{x,z}, w_{ij}^{x,y,z}$ versus learning iteration for these same β values.

We also train a classical BM on these same problems using the learning rule Eq. 7. The BM does not yield a good solution for any β: $S(\eta_{\beta=1}, \rho_{\mathrm{bm}}) = 3.08, S(\eta_{\beta=2}, \rho_{\mathrm{bm}}) = 3.73, S(\eta_{\beta=\infty}, \rho_{\mathrm{bm}}) = 4.01$ with $\rho_{\mathrm{bm}} = \mathrm{diag}(p_{\mathrm{bm}})$ and p_{bm} the learned BM distribution. The BM correctly reproduces the classical statistics w_{ij}^z, w_i^z for which it is optimized, but not the quantum statistics. Surprisingly for $\beta = \infty$, p_{bm} correctly reproduces the absolute values $\sqrt{p_{\mathrm{bm}}(s)} \approx |\psi(s)|$ for all states s, but of course not the correct signs (Figure 1c).

As a second example we repeat the above experiment for a fully connected quantum spin glass. The true Hamiltonian H (Eq. 13) on $n = 10$ spins has random couplings $w_{ij}^k \sim \mathcal{N}\left(0, \frac{1}{\sqrt{n}}\right)$ and random external fields $w_i^{x,z} \sim \mathcal{N}(0, 1)$, with $\mathcal{N}(\mu, \sigma)$ is the Gaussian distribution with mean μ and standard deviation σ. The results are shown in Figure 2 and are qualitatively similar to the AFH model in Figure 1. For small β, learning converges fast to the optimal solution and convergence is slower for increasing β both in terms of S (Figure 2a) and the RMS error in the learned Hamiltonian parameters (Figure 2b). For $\beta = \infty$, convergence is very slow ($S(\eta_{\beta=\infty}, \rho) = 2.1 \times 10^{-4}$ after 2000 iterations) and the RMS error in the learned Hamiltonian parameters is 0.0785 ($\beta_{\mathrm{est}} = 6.4$).

Note, that all these problems have a solution $\rho = \eta_\beta$ with $S = 0$ for any β, but the QBM learning rule has increasing difficulty to find this solution for large β. The reason is the following. When $\eta > 0$ (its eigenvalues are positive), then S is strictly convex in ρ and $S = 0$ if and only if $\rho = \eta$ (see [19]). In this case there is a unique optimum and the parameters of the target Hamiltonian can be recovered exactly from the quantum statistics. Examples are the finite temperature quantum Hamiltonian systems discussed above. Instead, for $\beta = \infty$, the density matrix $\eta_{\beta=\infty} = \psi\psi^\dagger$ is not positive (it has eigenvalues zero), S is not strictly convex [19] and the reconstructed Hamiltonian is not unique. However, provided that the QBM model Hamiltonian contains all relevant interaction terms, the QBM can accurately reconstruct the density matrix from the

observed statistics. Therefore it can also accurately represent any (non local) quantum feature of such system, such as entanglement.

Quantum state tomography as discussed in this section cannot be achieved by optimizing a classical likelihood. However, [20] show that by using multiple bases representations of the wave function simultaneously, one can accurately reconstruct the ground state wave function using a maximum likelihood approach. They do not reconstruct the Hamiltonian.

It should be noted that the data statistics that are required for learning are prescribed by the model Hamiltonian that is assumed. Each term H_r in Eq. 11 introduces a parameter w_r that requires statistics $\langle H_r \rangle$ for learning. This has the advantage that the accuracy and complexity of the learning problem can be adjusted by changing the Hamiltonian. This is in contrast with [21, 22] who present methods that can be used to reconstruct a pure state [21] or mixed state [22] without assuming a particular model (such as the exponential represention in terms of a Hamiltonian Eq. 11). Such 'model free' approach clearly requires much more data for learning and does not reconstruct the Hamiltonian of the system.

5. Quantum statistics in classical data

We can also apply the QBM to learn classical statistical learning problems. We first consider unsupervised learning. In this case, the data is represented as a classical distribution Eq. 1. We define a rank one density matrix

$$\eta = |\psi\rangle \langle\psi| \qquad \psi(s) = \langle s|\psi\rangle = \sqrt{q(s)}e^{i\alpha(s)} \tag{15}$$

which we refer to as the data state. $\alpha(s)$ is an arbitrary s dependent phase. The definition Eq. 15 implies classical and quantum statistics for the classical data distribution q. Note, that the expectation of an observable A is

$$\langle A \rangle_\eta = \sum_{s,s'} A(s,s')\sqrt{q(s)q(s')}e^{i\alpha(s)-i\alpha(s')} \tag{16}$$

For classical statistics, $A(s,s') = A(s)\delta_{s,s'}$ is diagonal and $\langle A \rangle_\eta = \sum_s A(s)q(s)$ is equal to the classical expectation value. When A is not diagonal, $\langle A \rangle_\eta$ defines new 'quantum' statistics for q. Thus, the quantum statistics defined in Eq. 16 generalize the classical statistics. Note, that the classical statistics are linear in q and the quantum statistics are quadratic in $\sqrt{q}$.

We propose the data state η as the target density matrix for the quantum likelihood Eq. 10. The optimal solution ρ is a density matrix that represents the quantum statistics in the classical data and has no equivalent in terms of a probability distribution. When ρ is diagonal, $\rho(s,s') = p(s)\delta_{s,s'}$, the quantum likelihood Eq. 10 reduces to the classical likelihood Eq. 3. Thus, quantum learning is the generalization of classical learning to density matrices. It can find better solutions because the optimization is over a larger class of models. In addition, it learns from quantum statistics that have no classical analogue.

The quantum likelihood depends on the phase $\alpha(s)$ in Eq. 15 which can take infinitely many values. Therefore, we need to define $\alpha(s)$ in order to properly define the learning objective. However, we can absorb $\alpha(s)$ in the model ρ in the following way. Denote η_0 as the density matrix with $\alpha(s) = 0$. Then Eq. 15 states that $\eta = U\eta_0 U^\dagger$ with $U(s', s) = \delta_{s,s'}e^{i\alpha(s)}$ a diagonal unitary matrix. Then

$$L = \text{Tr}\left(\eta \log \rho\right) = \text{Tr}\left(\eta_0 \log(U^\dagger \rho U)\right)$$

Thus, a non-zero $\alpha(s)$ is equivalent to a change in model density matrix from ρ to $U^\dagger \rho U$. The phases α can in principle be learned together with the other parameters that define ρ. The resulting learning rules are similar (but different) from Eq. 14 and their evaluation is of similar computational complexity.

We will not consider this generalization here and choose $\alpha(s) = 0$ for the numerical experiments in this section. We will leave the issue of optimizing α as a topic for future research. The data statistics in the QBM learning rule Eq. 14 are then

$$
\begin{aligned}
&\langle\sigma_i^x\rangle_\eta = \sum_s \sqrt{q(F_i s)q(s)} &\qquad &\langle\sigma_i^x\sigma_j^x\rangle_\eta = \sum_s \sqrt{q(F_i F_j s)q(s)} \\
&\langle\sigma_i^y\rangle_\eta = 0 &\qquad &\langle\sigma_i^y\sigma_j^y\rangle_\eta = -\sum_s s_i s_j \sqrt{q(F_i F_j s)q(s)} \quad (17) \\
&\langle\sigma_i^z\rangle_\eta = \sum_s s_i q(s) &\qquad &\langle\sigma_i^z\sigma_j^z\rangle_\eta = \sum_s s_i s_j q(s)
\end{aligned}
$$

Since η is real symmetric, the expectation of complex Hermitian observables such as $\langle\sigma_i^y\rangle_\eta$ are zero. In general, computation of these data statistics requires 2^n operations. However, when q is given in terms of a data set of N patterns, as in Eq. 1, the classical statistics can be computed linear in N. For the quantum statistics we compute the set of uniquely occurring patterns $\{s^a, a = 1, \ldots N_{\text{unique}}\}$ in the data set with $N_{\text{unique}} \leq N$ and the number of times $N(s^a) > 0$ that each of these patterns occurs . This computation requires $\mathcal{O}(N \log N)$ operations. Subsequently, we can compute each quantum statistics quadratic in N_{unique}§

The QBM can learn classical data problems problems significantly more accurate than the BM, for two reasons. The obvious reason is that the QBM has about 3 times as many parameters than the BM when both include first and second order interactions. Therefore, the QBM should fit the data always at least as good as the BM. However, this does not necessary generalize to new data. The second reason is that the QBM not only has more parameters, but these parameters are constrained by additional statistics (the quantum statistics). So the QBM sees more information about the distribution q than the BM. This also improves learning. We illustrate these ideas with the following two numerical examples.

The first data set was collected from preprocessed multi-electrode array recording from 160 salamander retinal ganglion cells responding to 297 repeats of a 19 s natural

---

§ For instance, for the computation of $\langle\sigma_i^x\rangle = \frac{1}{N}\sum_{a=1}^{N_{\text{unique}}} \sqrt{N(s^a)N(F_i s^a)}$. For each s^a, we compute $N(F_i s^a)$ by scanning the list of all nonzero $N(s^a)$ for $N(F_i s^a)$. This requires less than N_{unique} steps. Possibly, this computation can be accelerated.

movie [23]‖. See Figure 3(a) for data from one repeat. We selected the 5 neurons with highest average firing rates. We used data from 10 repeats to train the BM and the QBM. The QBM solution is significantly better than the BM solution ($S(\eta, \rho_{\mathrm{bm}}) = 1.79$ and $S(\eta, \rho_{\mathrm{qbm}}) = 1.06 \times 10^{-2}$ with $\rho_{\mathrm{bm}} = \mathrm{diag}(p_{\mathrm{bm}})$ and p_{bm} the learned BM distribution), indicating that the QBM represents the (diagonal and non diagonal) statistics of η better.

We can also assess how well the QBM and BM capture the classical statistics. ρ_{qbm} is close to rank one: its largest eigenvalue is 0.9987. ρ_{qbm} is thus well approximated by its rank one approximation: $\rho_{\mathrm{qbm}} \approx \psi\psi^{\dagger}$ with ψ the extreme eigenvector of ρ_{qbm}. From Eq. 15, this implies that ρ_{qbm} approximately represents the classical distribution $p_{\mathrm{qbm}}(s) = |\psi(s)|^2$. The KL divergences are $KL(q|p_{\mathrm{bm}}) = 9.57 \times 10^{-3}$ and $KL(q|p_{\mathrm{qbm}}) = 4.65 \times 10^{-4}$, which shows that the QBM also represents the classical statistics better than the BM. A scatter plot of p_{bm} and p_{qbm} versus the original data distribution q is shown in Figure 3(b). We tested the generalization performance of the BM and QBM solution on 28 test sets, each consisting of 10 repeats, by computing the KL divergences $KL(q_i|p_{\mathrm{(q)bm}})$ between the (Q)BM solution and these 28 empirical distributions q_i. A histogram of these KL divergences are shown in Figure 3(c). We also compute the relative entropies $S(\eta_i|\rho_{\mathrm{(q)bm}})$ with η_i the rank one density matrix from q_i and find $S(\eta_i|\rho_{\mathrm{bm}}) = 1.81 \pm 0.02$ and $S(\eta_i|\rho_{\mathrm{qbm}}) = 0.120 \pm 0.057$. The conclusion is that the QBM solution is significantly more accurate than the BM, both in terms of classical and quantum statistics, and this accuracy generalizes to unseen data.

The second data set shows that the QBM can learn problems that cannot be learned by the classical BM. As an extreme example, we consider the parity problem on n spins. The data distribution is $q(s) \propto q_0(s) \exp(\theta \sum_i s_i)$, where $q_0(s) = 1, 0$ when the parity of s, ie. $\prod_i s_i$, is even or odd, respectively and θ biases the spins to a nonzero mean value. This problem cannot be learned by a classical BM, because it can only effectively model up to second order classical statistics and the parity problem requires knowledge of nth order classical statistics. The QBM can learn this problem perfectly. For $\theta = 0$, the optimal solution $\rho_{\mathrm{qbm}} = \frac{1}{2}\left(\psi_1\psi_1^{\dagger} + \psi_2\psi_2^{\dagger}\right)$ is a rank two density matrix with $\psi_{1,2}$ orthonormal vectors that are linear combinations of the even parity ($\sqrt{q_0}$) and odd parity ($\sqrt{1 - q_0}$) solutions. The reason is that the data statistics $\langle \ldots \rangle_\eta$ are identical for the even and odd parity problem. The θ dependent term biases the solution to even ($\theta > 0$) or odd ($\theta < 0$) parity. For $\theta = 1$ and $n = 10$ the solution ρ_{qbm} is close to a rank one (largest eigenvalue is 0.9973) and $S(\eta, \rho_{\mathrm{qbm}}) = 4.69 \times 10^{-3}$. The BM solution has $S(\eta, \rho_{\mathrm{bm}}) = 2.37$. As above, we approximate $\rho_{\mathrm{qbm}} \approx \psi\psi^{\dagger}$ with ψ the extreme eigenvector of ρ_{qbm} and define $p_{\mathrm{qbm}}(s) = |\psi(s)|^2$. The KL divergences are $KL(q|p_{\mathrm{bm}}) = 0.451$ and $KL(q|p_{\mathrm{qbm}}) = 1.31 \times 10^{-5}$. In Figure 4 we show that p_{qbm} correctly represents the biased parity problem while p_{bm} does not capture any structure in the data.
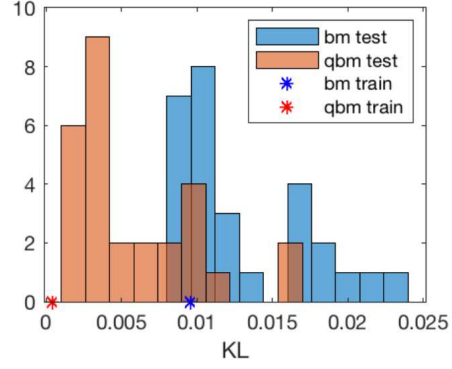
‖ See `https://datarep.app.ist.ac.at/61/2/bint_fishmovie32_100.zip` for further details and the data.
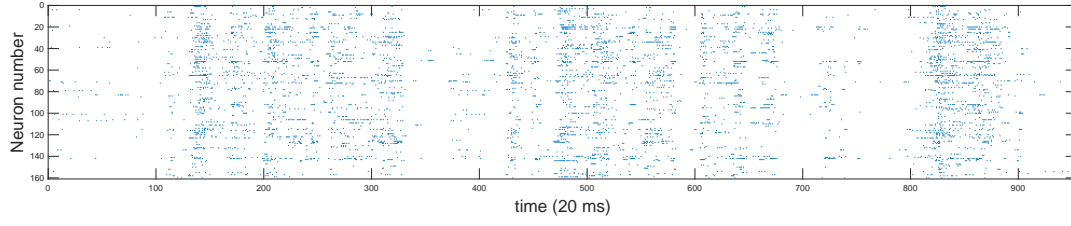
(a) One repeat of neural activity of 160 salamander retinal ganglion cells.
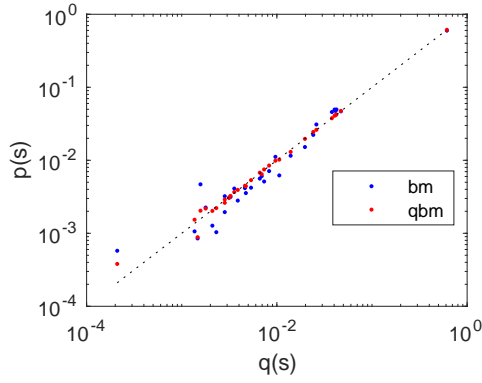
(b) p_{bm} and p_{qbm} versus q.

(c) Histogram of KL divergences of BM and QBM on 28 independent test sets.

Figure 3. (Color online). QBM and BM learning on neural spike data. Data consists of 297 repeats of 19 second simultaneous recordings of 160 neurons. Time is divided into 20 ms time bins. The data is binary and indicates whether a given cell elicited at least one spike in a given bin. One repeats is shown in (a). (b) Scatter plot of BM solution p_{bm} and rank one approximation p_{qbm} of QBM (see text) versus the training data distribution q. (c) Histogram of KL divergences of BM and QBM on 28 independent test sets. KL divergence on the training set is indicated by *.
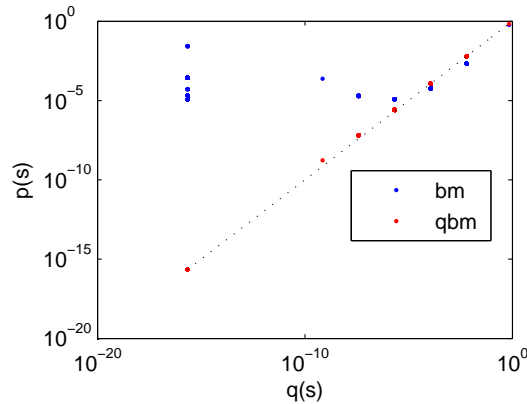

Figure 4. (Color online). Scatter plot of $p_{\mathrm{bm}}(s)$ and $p_{\mathrm{qbm}}(s)$ versus the true probability $q(s)$ for the biased parity problem ($\theta = 1$) on $n = 10$ spins.

5.1. Classification

We can use the QBM for binary classification. Suppose that input vectors are $x = s_{1:n-1}$ and the class labels are $y = s_n = \pm 1$. Denote $y = f(x)$ a deterministic mapping that maps each input pattern x onto a class label y. We define the data distribution as $q(s) = q(x)q(y|x)$ with $q(y|x) = \delta_{y,f(x)}$ to train the BM and QBM.

For the QBM we define the conditional (classical) probability of y given x as

$$p_{\text{qbm}}(y|x) = \frac{|\psi(x,y)|^2}{\sum_y |\psi(x,y)|^2}$$

with ψ the ground state wave function of the learned Hamiltonian. ¶

Similarly, for the BM we define

$$p_{\text{bm}}(y|x) = \frac{p_{\text{bm}}(x,y)}{\sum_y p_{\text{bm}}(x,y)}$$

Since the BM has pairwise interactions, $p_{\text{bm}}(y|x)$ can be written as $p_{\text{bm}}(y|x) = \sigma\left(y \sum_i w_i x_i\right)$ with $\sigma(x)$ is a sigmoid function. Therefore, the BM yields a linear classifier, and can thus only correctly classify linearly separable problems. For the QBM this is not true and $p_{\text{qbm}}(y|x)$ can also correctly classify nonlinear classification problems. For both models, we define the classifier by selecting the most likely class label: $y(x) = \text{argmax}_{y=\pm 1} \, p_{\text{(q)bm}}(y|x)$.

To assess the quality of the QBM as a classifier we use $n = 4$ and consider all $2^{2^{n-1}} = 256$ binary functions on 3 inputs. We train the BM and QBM on all these problems with $q(x) \propto \exp(\theta \sum_i x_i)$ with different values of θ. The effect of $\theta \neq 0$ is to give certain input patterns more weight than others during the training of the BM and QBM. The BM can correctly classify 104 of the 256 problems (these are the linearly separable problems) for a large range of θ values (see Figure 5). For $\theta = 0$, the QBM can correctly classify 160 of the 256 problems. For θ in the range 0.1–0.8 the QBM can correctly classify at least 254 problems and for $\theta \approx 0.6$–0.7, the QBM can correctly classify all 256 problems. Details of the classifications for $\theta = 0$ and $\theta = 0.7$ are shown in table C1.

To better understand the strong dependence of the QBM classification accuracy on θ we list in table C1 also the entropy of the eigenvalues of ρ_{qbm} and the KL divergence between the ground state probability and q. We see that the effect of $\theta > 0$ is that the harder problems become closer to a rank one solution and the ground state becomes closer to the target distribution q. Why this is the case is not clear and requires more investigation. There seems no principled way to choose θ other than exploring a range of values, but one can clearly think of adaptive schemes such as boosting [24]. The

¶ Alternatively, we could have defined a classifier based on the entire density matrix ρ as

$$p(y|x) = \frac{\rho_{\text{qbm}}((x,y),(x,y))}{\sum_y \rho_{\text{qbm}}((x,y),(x,y))}$$

However, numerically we found that using ρ_{qbm} rather than the ground state wave function does not improve classification. The reason may be that successful classification requires a solution $\rho \approx \eta$, which is rank one.
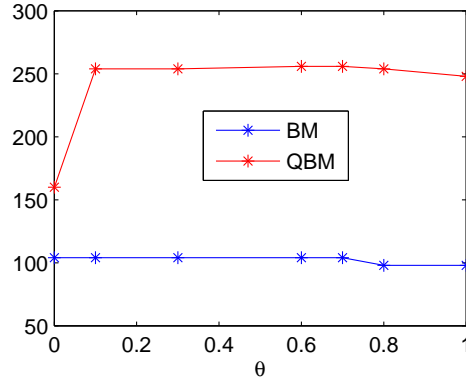
Figure 5. Number of correctly classified problems for the BM and QBM as a function of θ. For the BM, 104 problems are correctly classified for $0 \leq \theta \leq 0.7$. For the QBM, 254 or more problems are correctly classified for $0.1 \leq \theta \leq 0.8$ and all 256 problems are correctly classified for $\theta \approx 0.6$–0.7.

important conclusion is that with a proper θ, the QBM can correctly classify *all* 256 problems.

It is clear that the QBM cannot correctly classify all binary functions for larger problems, because the number of parameters of the QBM scales polynomial as $\mathcal{O}(n^2)$ while the number of training samples that sets the number of constraints scales exponentially as 2^{n-1}. Nevertheless, the present study suggests that the QBM with up to second order interaction can learn many non linearly separable problems than the BM and it is likely that this also holds for larger problems.

6. Entanglement in classical data

The data state Eq. 15 $\eta = |\psi\rangle \langle\psi|$ (with arbitrary phase $\alpha(s)$) is a much more complex object than the classical probability distribution q from which it is derived. It displays full quantum features such as non locality and entanglement. In this section we explain what these quantum features mean for classical data.

η is a basis independent object that allows measurements in any basis. Measurement in the basis where σ_i^z is diagonal yields the classical measurement outcomes. For instance, measuring σ_i^z yields outcomes $s_i = \pm 1$ and measuring $\sigma_i^z \sigma_j^z$ yields outcomes $s_i s_j = \pm 1$. The measurement outcome $s_i = \pm 1$ means that the quantum system is in the state $|s_i = \pm 1\rangle$. Repeated measurement yields statistics whose expected values, such as $\langle\sigma_i^z\rangle_\eta = \langle s_i\rangle_q$ or $\langle\sigma_i^z\sigma_j^z\rangle_\eta = \langle s_i s_j\rangle_q$ etc., are predicted by the joint probability distribution $q(s_1, \ldots, s_n)$.

But η also allows measurement in any other basis, for instance in the basis where σ_i^x is diagonal. The measurement σ_i^x yields outcomes $t_i = \pm 1$ meaning that the quantum system is in the state $|t_i = \pm 1\rangle = \frac{1}{\sqrt{2}}(|s_i = 1\rangle \pm |s_i = -1\rangle)$, which is a superposition of two 'classical' states $|s_i = \pm 1\rangle$. Repeated measurement yields statistics

whose expected values such as $\langle \sigma_i^x \rangle_\eta$ are given by Eq. 17. Equivalently, one can define a probability distribution $\tilde{q}$ in the new basis $|t\rangle = |t_1, \ldots, t_n\rangle$ and $\langle \sigma_i^x \rangle_\eta = \sum_{t_i} t_i \tilde{q}(t_i)$. In general, measurements that are in an orthogonal basis $|t\rangle = \sum_s U_{ts} |s\rangle$, with U a unitary transformation are described by the classical distribution $\tilde{q}(t) = |\langle t|\psi \rangle|^2$ (see Appendix D).

==The difference between q and η is reflected in the difference between the classical mutual information $I_c(q)$ and the quantum mutual information $I(\eta)$ between two sub systems A and B (see Appendix E).== For a pure state, such as the data state η, the total entropy is zero $h(\eta) = 0$ and the entropies of the sub systems are equal: $h(\eta_A) = h(\eta_B)$ (see Appendix F). $^+$ Thus, the quantum mutual information Eq. E.3 is

$$I(\eta) = h(\eta_A) + h(\eta_B) - h(\eta) = 2h(\eta_A) \tag{18}$$

For pure states, the quantum mutual information is equal to the entanglement [25].

The quantum mutual information $I(\eta)$ in the data state η is never less than the classical mutual information $I_c(q)$ of the distribution q from which it is composed, ie. $I_c(q) \leq I(\eta)$. The proof is as follows. The diagonal of η is the classical distribution q. The operation that maps a density matrix to its diagonal, is known as the incoherence operator $\Pi(\eta) = \eta^{\text{diag}} = q$. The incoherence operator is a completely positive trace preserving (CPTP) operation. For any CPTP operation Φ [26]

$$S(\Phi(\rho), \Phi(\sigma)) \leq S(\rho, \sigma) \tag{19}$$

Thus we establish the bound:

$$I_c(q) = S(q, q_A q_B) = S(\Pi(\eta), \Pi(\eta_A)\Pi(\eta_B)) \leq S(\eta, \eta_A \eta_B) = I(\eta)$$

which completes the proof.

In fact, we conjecture the stronger statement that $I_c(q) \leq I(\eta)/2$ but we have not been able to prove this. We illustrate this with a numerical example. Consider three spins s_1, s_2, s_3 and define sub system $A = (s_1, s_2)$ and $B = s_3$. We generate random probability distributions q on these three spins and compute $I_c(q)$ and $I(\eta)$ using Eq. 15 with random phases $\alpha(s)$ (Figure 6(a) blue dots) as well as with $\alpha(s) = 0$ (fig 6(b) blue dots). We observe that in all cases $I_c(q) \leq I(\eta)/2$. When the states of subsystem B are a deterministic function of the states of A we can show $I_c(q) = I(\eta)/2$ (see Appendix G for the proof and illustrated in Figure 6 red dots).

One can write for a pure state that $I(\eta) = C + Q$ with $C = Q = I(\eta)/2$ [27, 25, 28]. C is called the (maximal) classical correlation (see Appendix G) and Q is called the quantum correlation. Measuring in an orthogonal basis in which the reduced density matrices of the two sub systems are diagonal defines a distribution $\tilde{q}$ between measurement outcomes for which $I_c(\tilde{q}) = C \geq I_c(q)$. See Appendix G for details. We illustrate this for $\alpha(s) = 0$ with the distributions q_1 and q_2, defined as

$$q_1(s_2, s_3) = \begin{pmatrix} 0.0652 & 0 \\ 0 & 0.9348 \end{pmatrix} \qquad q_2(s_2, s_3) = \begin{pmatrix} 0.25 & 0.375 \\ 0.375 & 0 \end{pmatrix}$$

$^+$ Note, that this identity holds independent of the sizes of sub systems A and B: A can be a single bit and B the rest of the universe.

Le

(a) $\alpha(s)$ is random.
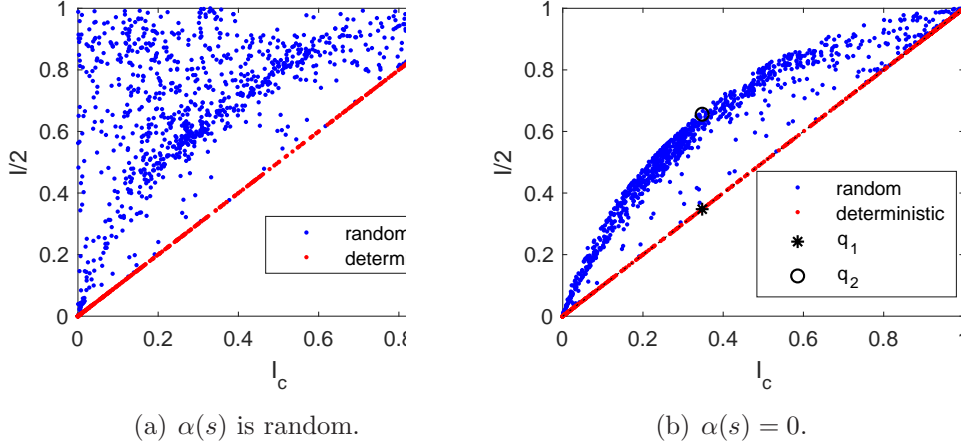(b) $\alpha(s) = 0$.

Figure 6. Scatter plot of $C = I(\eta)/2$ versus $I_c(q)$ between sub systems A and B for three binary variables $A = (s_1, s_2)$ and $B = s_3$ with random probability distribution $q(s_1, s_2, s_3)$. Blue: 1000 random distributions are generated, with each component $q(s_1, s_2)$ generated independently from a Beta distribution with parameters $\alpha = \beta = 0.1$. Subsequently, q is normalized. Red: The same, but with a deterministic relation $s_3 = s_2$. (a) η is constructed with random phases $\alpha(s)$; (b) η is constructed with phases $\alpha(s) = 0$.

independent of s_1, that are chosen such that the classical mutual information between A and B is identical: $I_c(q_1) = I_c(q_2)$. In q_1, s_B depends deterministically on s_A and $I_c(q_1) = I(\eta_1)/2$ (* in Figure 6(b)). In q_2, the relation is not deterministic and $I_c(q_2) < I(\eta_2)/2$ (o in Figure 6). Measurement in the orthogonal basis defined by the Schmidt decomposition yields

$$\tilde{q}_2(t_2, t_3) = \begin{pmatrix} 0.8307 & 0 \\ 0 & 0.1693 \end{pmatrix} \qquad U_A(t_2, s_2) = \begin{pmatrix} 0.83 & 0.5577 \\ 0.5577 & -0.83 \end{pmatrix}$$

and $U_B = U_A$ and $I_c(\tilde{q}_2) = I(\eta_2)/2 > I_c(q_2)$.

Fig. 6 suggests that the quantum implementation of the classical distribution Eq. 15 using non-zero phases $\alpha(s)$ can be superior than using $\alpha(s) = 0$ because for the same classical mutual information $I_c(q)$ it yields larger quantum mutual information $I(\eta)$. Therefore, an optimized orthogonal measurement can yield more classical mutual information $I_c(\tilde{q}) = I(\eta)/2$ for properly chosen non-zero $\alpha(s)$ than for $\alpha(s) = 0$.

6.1. Non orthogonal measurement and the Bell inequality

The statistics of local orthogonal quantum measurements are described by a classical joint probability model $\tilde{q}$ and the mutual information is restricted to $I_c(\tilde{q}) \leq C = I(\eta)/2$. The implication is that the remaining quantum information Q resides in non-local features of the state η [25]. This relates to the fact that the statistics of non-orthogonal measurements violate the Bell inequality, meaning that they cannot be described by a classical probability model involving local variables. Instead such measurement statistics require a description in terms of non-local variables (see Appendix H).

We illustrate the excess quantum information Q and the violation of the Bell inequality for the simplest possible example of two fully correlated binary variables $s = (s_1, s_2)$ with joint probability distribution

$$q(s_1, s_2) = \frac{1}{2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \psi(s_1, s_2) = \frac{1}{\sqrt{2}} \begin{pmatrix} e^{i\alpha_1} & 0 \\ 0 & e^{i\alpha_2} \end{pmatrix} \qquad (20)$$

where $\psi(s)$ is given by Eq. 15.

Note, that the classical mutual information between the two bits is $I_c(q) = 1$ and the quantum mutual information is $I(\eta) = 2$ (Eq. 18). Half of the quantum mutual information $C = I(\eta)/2 = I_c(q)$ is the classical correlation. The other half Q is the quantum correlation. [27] provides an intuitive explanation of this mysterious extra bit of information using the argument of bit erasure. Assume $\alpha_1 = \alpha_2 = 0$. Suppose that Alice has access to spin 1 and Bob has access to spin 2. Alice can apply to her spin one of two unitary transformations 1 or σ^z with equal probability. It is easy to show that this operation transforms $\eta = |\psi\rangle \langle\psi|$ to a diagonal density matrix $\eta' = \mathrm{diag}(q)$. The mutual information $I(\eta') = 1$ since the bits are fully correlated in the classical sense. Thus, by erasing one bit, which is forgetting the information which unitary Alice has applied, one bit of mutual information is lost. Alice can apply to her spin again one of two unitary transformations 1 or σ^x with equal probability. This operation either flips or not flips her spin. The final state $\eta'' = \eta_1 \otimes \eta_2$ is independent and the mutual information $I(\eta'') = 0$. Thus, the explanation why $I(\eta) = 2$ is that it takes the erasure of two bits of information to remove the mutual information between A and B.

We now show how η violates the Bell inequality. We consider 4 measurements $M_{ij} = A_i \otimes B_j, i, j = 1, 2$

$$A_1 = \sigma_1^z, \qquad A_2 = \sigma_1^x \sin\theta + \sigma_1^z \cos\theta$$
$$B_1 = \sigma_2^z \qquad B_2 = \sigma_2^x \sin\phi + \sigma_2^z \cos\phi$$

Denote the measurement outcomes as $m_{ij} = (a_{ij}, b_{ij}), i, j = 1, 2$ with $a_{ij}, b_{ij} = \pm 1$. The locality assumption states that for instance, the outcome of measurement A_1 does not depend on whether it is combined with the measurement B_1 or B_2. In other words, $m_{ij} = (a_i, b_j)$. This implies that the measurement outcomes can be described by a probability distribution $p(a_1, b_1, a_2, b_2)$ on 4 binary variables.

Note, that in the state $|\psi\rangle$ the measurement outcomes $a_1 = b_1$ are fully correlated. Thus, we can discard b_1 since it adds no additional information. Since the measurement outcomes $a_1, a_2, b_2 = \pm 1$ they trivially satisfy $a_1 a_2 + a_1 b_2 - a_2 b_2 \leq 1$ as well as any of the permutations of a_1, a_2, b_2 [29]. Let us assume that there exist a classical probability distribution $p(a_1, a_2, b_2)$ that describes the outcome statistics. Taking the expectation with respect to p we obtain the Bell inequality

$$B = |\langle a_1 a_2 \rangle - \langle a_1 b_2 \rangle| + \langle a_2 b_2 \rangle - 1 \leq 0 \qquad (21)$$
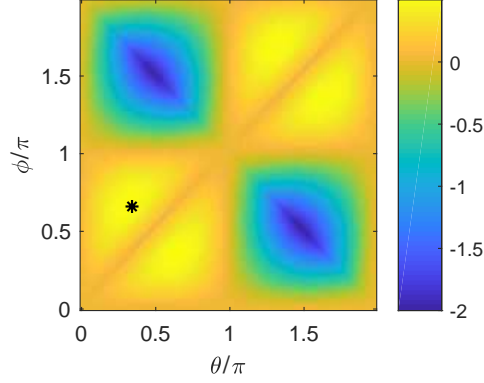
Figure 7. Violation of the Bell inequality Eq. 21 in classical data. The quantum state is built from classical data of two fully correlated binary variables Eq. 20. The figure shows $B(\theta, \phi)$ (Eq. 21) as a function of θ, ϕ. These measurements maximally violate the locality assumption for $(\theta, \phi) = \left(\frac{\pi}{3}, \frac{2\pi}{3}\right)$ indicated by $*$.

We compute the quantum statistics in the state ψ: $*$

$$\langle a_1 a_2 \rangle = \sum_{s_1' s_2', s_1, s_2} \psi^*(s_1', s_2') \frac{1}{2} \left(A_1 A_2 + A_2 A_1\right)_{s_1', s_1} \psi(s_1, s_2) = \cos \theta$$

$$\langle a_1 b_2 \rangle = \sum_{s_1' s_2', s_1, s_2} \psi^*(s_1', s_2') \left(A_1\right)_{s_1', s_1} \left(B_2\right)_{s_2', s_2} \psi(s_1, s_2) = \cos \phi$$

$$\langle a_2 b_2 \rangle = \sum_{s_1' s_2', s_1, s_2} \psi^*(s_1', s_2') \left(A_2\right)_{s_1', s_1} \left(B_2\right)_{s_2', s_2} \psi(s_1, s_2)$$

$$= \cos(\theta) \cos(\phi) + \sin(\theta) \sin(\phi) \cos(\alpha_1 - \alpha_2)$$

Substituting these correlations in Eq. 21, we observe that B is maximized when $\alpha_1 - \alpha_2 = 0, \pi$ and $\max_{\theta, \phi, \alpha_1 - \alpha_2} B = \frac{1}{2}$ which violates the Bell inequality. θ, ϕ can take multiple values as is illustrated for $\alpha_1 - \alpha_2 = 0$ in Figure 7 where we plot B as a function of θ, ϕ. Note, that these measurements are non-orthogonal. This example demonstrates that the statistics of non-orthogonal measurements cannot be described by a probability distribution of local outcomes.

This seems to contradict the fact that η is built from a classical distribution q that has only local variables. However, there is no contradiction because the Bell condition assumes that the quantum correlations are linear in the classical distribuion $p(a_1, a_2, b_2)$, while in fact they are quadratic in $\sqrt{q}$ (Eqs. 16 and 17).

7. Discussion

In this paper, we have addressed the problem how to represent a classical data distribution in a quantum system. The proposed method is to learn quantum

---

$*$ Note, that A_1 and A_2 do not commute and $A_1 A_2$ is not Hermitian. Therefore the quantum expectation $\langle A_1 A_2 \rangle$ is not real while the classical correlation $\langle a_1 a_2 \rangle$ are real. We therefore identify the classical correlations with the correlations of the Hermitian operator $\frac{1}{2}(A_1 A_2 + A_2 A_1)$.

Hamiltonian that is such that its ground state approximates the given classical distribution. The essential new ingredient is to represent a classical distribution as a rank one density matrix such that $\text{diag}(\eta) = q$. In particular, we proposed Eq. 15.

This choice is not unique. First of all, one needs to specify the unitary transformation $U(\alpha) = \text{diag}(e^{i\alpha})$ in Eq. 15. In our numerical experiments in section 5 we have assumed $\alpha = 0$. The choice of α affects the learning and raises the question how to best choose the phases α. Part of the answer is to observe that $U(\alpha)$ can be equivalently considered part of a model $U^\dagger \rho U$. Since for an arbitrary data set there is no principled choice of ρ, one can extend this reasoning to (a parametrized model) for U. U and ρ can then be jointly optimized to maximize L. The results on entanglement in Figure 6 suggest that choosing $\alpha(s)$ non-zero may result in a quantum state that has more entanglement than when $\alpha(s) = 0$. The issue of the choice of $\alpha(s)$ has not been further explored and is left for future research.

Secondly, a further generalization is to define $\eta = \mu\eta^* + (1 - \mu)\text{diag}(q)$ with $0 \leq \mu \leq 1$ and η^* the rank one density matrix Eq. 15. Note that when $q(s) > 0$ for all s and $\mu \neq 1$, η is a positive matrix while η^* is not (it has eigenvalues 0). Thus choosing η as a learning target with $\mu \neq 1$ in the quantum likelihood Eq. 10 instead of η^* may improve convergence speed since $S(\eta, \rho)$ is strictly convex while $S(\eta^*, \rho)$ is not [19]. On the other hand, an advantage of the choice η^*, instead of η is that it maximizes the entanglement and quantum mutual information: $I(\eta^*) \geq I(\eta)$ for all μ and any sub division A, B. This follows from Eq. 19 and the fact that the mapping $\Phi : \eta^* \to \eta$ is a CPTP operation and $\Phi(\eta_A \otimes \eta_B) = \Phi(\eta_A) \otimes \Phi(\eta_B)$. Thus, the gain in quantum mutual information is largest for $\mu = 1$. The set of density matrices η that have diagonal q is much larger than what we have considered in this paper and other choices may be considered.

When transforming a probability distribution to a density matrix by Eq. 15, the mutual information between subsets of variables increases because $I(\eta) \geq 2I_c(q)$. $C = I(\eta)/2$ is the so-called (maximal) classical correlations and the extra mutual information $I(\eta)/2 - I_c(q)$ is accessible by optimizing the orthogonal measurement basis. Its outcome statistics are described by a classical probability distribution on local variables. When implementing this idea in a quantum device, these extra correlations can in principle be used to improve performance in machine learning or other applications.

The remaining quantum mutual information $Q = I(\eta)/2$ is 'non local' and cannot be accessed by local orthogonal measurements. This part of the mutual information is the true quantum entanglement. The non-locality is manifested in the violation of the Bell inequality and occurs even for classical distributions q that contain deterministic relations, such as two fully correlated classical spins or the parity problem that we considered in Figure 4. It is an open and interesting question whether these non-local features can be used in a quantum device for the type of learning problems that we considered here.

For large simulations, learning the QBM is intractable. In each learning iteration one must compute statistics $\langle H_r \rangle_\rho$ for the current estimated density matrix ρ. In

principle, it requires $\mathcal{O}\left(2^{2n}\right)$ operations and memory to compute the entire density matrix. To generate the results of Figure 1 we effectively made use of a low rank approximation using $L = 6$ extreme eigenvectors and a sparse representation, requiring $\mathcal{O}\left(L2^{n}\right)$ computation. Clearly, this is still exponential in the number variables and does not scale to large problem instances.

As in classical BM learning, one can apply various approximate inference methods to estimate $\langle H_r \rangle_\rho$. One promising approach is to approximate $\langle H_r \rangle_\rho$ by its ground state statistics. This can be done by computing in each learning iteration an approximation to the ground state wave function by making a variational Anzats and minimize the Raleigh quotient for the current instance of the Hamiltonian [30, 31]. In particular, [10] use a (classical) restricted Boltzmann machine as the variational Anzats for the wave function that shows great potential. This approach is currently under investigation.

Alternatively, one can imagine that the computation of the quantum statistics can be done on physical quantum hardware. This would give a hybrid quantum-classical computing scheme [32], where the learning iterations are executed on a classical computer and the computation of $\langle H_r \rangle_\rho$ for the current Hamiltonian (or its ground state approximation) would be done on the quantum device. Initial encouraging results were obtained to implement the approximate quantum learning rule [2] using quantum annealing on a D-Wave computer [33]. See also [34]. An obvious next step is to use the quantum learning rules proposed here to improve these results.

7.1. Acknowledgments

[1] Mária Kieferová and Nathan Wiebe. Tomography and generative training with quantum boltzmann machines. *Physical Review A*, 96(6):062327, 2017.

[2] Mohammad H Amin, Evgeny Andriyash, Jason Rolfe, Bohdan Kulchytskyy, and Roger Melko. Quantum boltzmann machine. *Physical Review X*, 8(2):021050, 2018.

[3] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[4] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[5] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[6] Tadashi Kadowaki and Hidetoshi Nishimori. Quantum annealing in the transverse ising model. *Physical Review E*, 58(5):5355, 1998.

[7] Bettina Heim, Troels F Rønnow, Sergei V Isakov, and Matthias Troyer. Quantum versus classical annealing of ising spin glasses. *Science*, 348(6231):215–217, 2015.

[8] Steven H Adachi and Maxwell P Henderson. Application of quantum annealing to training of deep neural networks. *arXiv preprint arXiv:1510.06356*, 2015.

[9] Marcello Benedetti, John Realpe-Gómez, Rupak Biswas, and Alejandro Perdomo-Ortiz.

Estimation of effective temperatures in quantum annealers for sampling applications: A case study with possible applications in deep learning. *Physical Review A*, 94(2):022308, 2016.

[10] Giuseppe Carleo and Matthias Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355(6325):602–606, 2017.

[11] Juan Carrasquilla and Roger G Melko. Machine learning phases of matter. *Nature Physics*, 13(5):431, 2017.

[12] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195, 2017.

[13] Nicolas J Cerf and Chris Adami. Negative entropy and information in quantum mechanics. *Physical Review Letters*, 79(26):5194, 1997.

[14] Nicolas J Cerf and Christoph Adami. Quantum extension of conditional probability. *Physical Review A*, 60(2):893, 1999.

[15] Ruediger Schack, Todd A Brun, and Carlton M Caves. Quantum bayes rule. *Physical Review A*, 64(1):014305, 2001.

[16] Nathan Wiebe and Leonard Wossnig. Generative training of quantum boltzmann machines with hidden units. *arXiv preprint arXiv:1905.09902*, 2019.

[17] RC Wiersema and HJ Kappen. Implementing perceptron models with qubits. *Physical Review A*, 2019. In press, arXiv:1905.06728.

[18] D. Ackley, G. Hinton, and T. Sejnowski. A learning algorithm for Boltzmann Machines. *Cognitive Science*, 9:147–169, 1985.

[19] Eric Carlen. *Trace inequalities and quantum entropy: an introductory course*, volume 522, pages 73–140. American Mathematical Society, 2010.

[20] Giacomo Torlai, Guglielmo Mazzola, Juan Carrasquilla, Matthias Troyer, Roger Melko, and Giuseppe Carleo. Many-body quantum state tomography with neural networks. *arXiv preprint arXiv:1703.05334*, 2017.

[21] Christopher Ferrie. Self-guided quantum tomography. *Physical review letters*, 113(19):190404, 2014.

[22] Christopher Granade, Christopher Ferrie, and Steven T Flammia. Practical adaptive quantum tomography. *New Journal of Physics*, 19(11):113017, 2017.

[23] Gašper Tkačik, Olivier Marre, Dario Amodei, Elad Schneidman, William Bialek, and Michael J Berry II. Searching for collective behavior in a large network of sensory neurons. *PLoS computational biology*, 10(1):e1003408, 2014.

[24] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[25] J Maziero, L C₋ Celeri, RM Serra, and V Vedral. Classical and quantum correlations under decoherence. *Physical Review A*, 80(4):044102, 2009.

[26] Göran Lindblad. Completely positive maps and entropy inequalities. *Communications in Mathematical Physics*, 40(2):147–151, 1975.

[27] Berry Groisman, Sandu Popescu, and Andreas Winter. Quantum, classical, and total amount of correlations in a quantum state. *Physical Review A*, 72(3):032317, 2005.

[28] Vlatko Vedral. Foundations of quantum discord. In *Lectures on General Quantum Correlations and their Applications*, pages 3–7. Springer, 2017.

[29] N. J. Cerf and C. Adami. Entropic bell inequalities. *Physical Review A*, 55(5):3371, 1997.

[30] William Lauchlin McMillan. Ground state of liquid he 4. *Physical Review*, 138(2A):A442, 1965.

[31] J Carlson, Stefano Gandolfi, Francesco Pederiva, Steven C Pieper, Rocco Schiavilla, KE Schmidt, and Robert B Wiringa. Quantum monte carlo methods for nuclear physics. *Reviews of Modern Physics*, 87(3):1067, 2015.

[32] John Preskill. Quantum computing in the nisq era and beyond. *Quantum*, 2:79, 2018.

[33] Dmytro Korenkevych, Yanbo Xue, Zhengbing Bian, Fabian Chudak, William G Macready, Jason Rolfe, and Evgeny Andriyash. Benchmarking quantum hardware for training of fully visible boltzmann machines. *arXiv preprint arXiv:1611.04528*, 2016.

[34] Kosuke Mitarai, Makoto Negoro, Masahiro Kitagawa, and Keisuke Fujii. Quantum circuit learning. *Physical Review A*, 98(3):032309, 2018.

[35] Huzihiro Araki and Elliott H Lieb. Entropy inequalities. In *Inequalities*, pages 47–57. Springer, 2002.

[36] Jarosław Adam Miszczak. Singular value decomposition and matrix reorderings in quantum information theory. *International Journal of Modern Physics C*, 22(09):897–918, 2011.

[37] Kavan Modi, Aharon Brodutch, Hugo Cable, Tomasz Paterek, and Vlatko Vedral. The classical-quantum boundary for correlations: discord and related measures. *Reviews of Modern Physics*, 84(4):1655, 2012.

[38] Eugene P Wigner. On hidden variables and quantum mechanical probabilities. *American Journal of Physics*, 38(8):1005–1009, 1970.

[39] John S Bell. On the einstein podolsky rosen paradox. *Physics Physique Fizika*, 1(3):195, 1964.

[40] John S Bell. On the problem of hidden variables in quantum mechanics. In *John S Bell On The Foundations Of Quantum Mechanics*, pages 1–6. World Scientific, 2001.

Appendix A. QBM learning using the classical likelihood

[2] and one of the methods discussed in [1] use the classical likelihood for learning the QBM by considering the diagonal of the density matrix. Write $\rho(s, s') = \delta_{s,s'} p(s)$. For each state s, define Λ_s a matrix with components $\Lambda_s(s', s'') = \delta_{s,s'} \delta_{s,s''}$. Then $p(s) = \text{Tr}\left(\Lambda_s \rho\right)$ and the classical log likelihood is

$$L = \sum_s q(s) \log \text{Tr}\left(\Lambda_s \rho\right) = \sum_s q(s) \log \text{Tr}\left(\Lambda_s e^H\right) - \log Z \qquad \text{(A.1)}$$

Note, that this expression differs from the quantum likelihood MT Eq. 4 in the first term only. Its gradient is given by

$$\frac{\partial}{\partial w_r} \log \text{Tr}\left(\Lambda_s e^H\right) = \frac{1}{\text{Tr}\left(\Lambda_s e^H\right)} \int_0^1 dt\, \text{Tr}\left(\Lambda_s e^{Ht} H_r e^{H(1-t)}\right)$$

Because of Λ_s does not commute with e^{Ht}, the time integration remains and the expression cannot be easily evaluated. [2] address this problem by deriving a lower bound on their likelihood using the Golden-Thompson inequality and maximizing this bound. However, as the authors admit, this procedure is clearly sub optimal and is inconsistent when learning transverse field components.

Appendix B. Quantum Boltzmann machine details

In the Hamiltonian Eq. 13, $\sigma_i^{x,y,z}$ are 2×2 Pauli spin matrices in the σ^z basis. Then $\sigma_i^z = \text{diag}(1, -1)$, its eigenvectors are the two component unit vectors $(1, 0)$ and $(0, 1)$, which we denote as $|s_i = \pm 1\rangle$ and eigenvalues $s_i = \pm 1$, respectively. On this basis

$$\sigma_i^z |s_i\rangle = s_i |s_i\rangle, \qquad \sigma_i^x |s_i\rangle = |-s_i\rangle \qquad \sigma_i^y |s_i\rangle = i s_i |-s_i\rangle$$

For n spins, the basis is the tensor product of the basis vectors $|s_i = \pm 1\rangle$ denoted as $|s\rangle = |s_1, \ldots, s_n\rangle$. On this basis, H is a $2^n \times 2^n$ matrix with matrix elements

$$\langle s'|H|s\rangle = \sum_{i=1}^n (w_i^x + i w_i^y s_i)\delta_{s', F_i s} + \sum_{i=1, j>i}^n \left(w_{ij}^x - w_{ij}^y s_i^k s_j^k\right) \delta_{s', F_i F_j s}$$

$$+ \left(\sum_{i=1}^{n} w_i^z s_i + \sum_{i<j}^{n} w_{ij}^z s_i s_j \right) \delta_{s',s}$$

with $F_i s$ the state s with spin i flipped to $-s_i$ and all other spins unchanged.

Appendix C. Details classification performance

Details of classification performance for all 256 classification problems on 3 binary inputs are summarized in table C1.

 (i) A subset of 104 problems is linearly separable (each problem has classification error 0) and can be classified exactly by all methods. The remaining problems are not linearly separable.

 (ii) A subset of 56 problems cannot be classified by BM (each problem has classification error 2-4) and can be exactly classified by QBM.

(iii) A subset of 88 problems. For $\theta = 0$ the QBM solution is approximately rank two, with eigenvalues $\lambda \approx (0.75, 0.25)$ and cannot correctly classify these problems. For $\theta = 0.7$ the QBM solution is closer to rank one and these problems can be exactly classified by QBM.

(iv) A subset of 8 problems that contain the parity and parity-like problems. For these problems, the solution ρ_{qbm} is rank two with two eigenvalues $\lambda = 0.5$. For $\theta = 0$, the ground state of the Hamiltonian is degenerate and there is no solution. For $\theta = 0.7$ the symmetry is broken and these problems can be exactly classified by QBM.

As a sanity check, we also computed the logistic regression solution $p_{\text{lr}}(y|x) = \sigma(y \sum_i w_i x_i)$ that maximizes the conditional classical likelihood $\sum_{x,y} q(x)q(y|x) \log \sigma(y \sum_i w_i x_i)$ instead of the joint classical likelihood that is used for the BM. We found no significant quality differences between the logistic regression and BM solutions (data not shown). We observe that with $\theta = 0.7$, the QBM solution is always closer to a rank one solution than for $\theta = 0$. For $\theta = 0.7$, the quality of the ground state approximation as measured by the KL divergence is excellent for the subsets of 8 problems for which the $\theta = 0$ solution is degenerate, while not much worse for the other problems.

Appendix D. Projective measurements

A projective measurement can be written as a sum $\sum_k \lambda_k E_k$, with $E_k = |\phi_k\rangle \langle\phi_k|$ is a set of Hermitian orthogonal projective operators on a Hilbert space that sum to the identity operator: $\sum_k E_k = I$. The outcome of the measurement is any of its eigenvalues λ_k. The measurement outcome is a stochastic event: repeated measurements E_k on the same quantum system with density matrix ρ may yield different values. The probability

$$p(\text{outcome of measurement is } \lambda_k) = \text{Tr}(E_k \rho) \qquad \sum_k \text{Tr}(E_k \rho) = 1$$

		$\theta = 0$					$\theta = 0.7$	
	BM	QBM	ENT	KL	BM	QBM	ENT	KL
104	0	0	$0.00 - 0.01$	$< 8 \times 10^{-6}$	0	0	$0.00 - 0.06$	$< 1 \times 10^{-3}$
56	2-4	0	$0.20 - 0.24$	$< 6 \times 10^{-4}$	1-4	0	$0.00 - 0.29$	$< 3 \times 10^{-2}$
88	2-4	1-4	$0.56 - 0.63$	$< 4 \times 10^{-1}$	1-3	0	$0.00 - 0.48$	$< 2 \times 10^{-2}$
8	8	8	$0.69 - 0.69$	NA	2-6	0	$0.00 - 0.00$	$< 7 \times 10^{-11}$

Table C1. Classification performance of various methods on all 256 classification problems on 3 binary inputs. Problems are partitioned in 4 subsets. Columns BM and QBM lists the range of classification errors (in the range 0 to 8) for that method for that subset of problems. ENT lists the range of entropies of the QBM solution ρ_{qbm}. KL lists the range of KL divergences $KL(q|p)$ with $p(s) = \frac{1}{Z}|\psi(s)|^2$ and $\psi(s)$ the ground state of H.

In Eq. 15 we defined $|\psi\rangle$ and η in the σ^z basis $|s\rangle$ that we considered throughout the paper. This basis also defines a set of projective measurements $E_s = |s\rangle\langle s|$. The probability of that measurement E_s on this quantum system yields outcome 1 is $\mathrm{Tr}\,(E_s\eta) = q(s)$.

The probability of the measurement outcome depends on the choice of measurement basis. Another valid set of projective measurement is $E_t = |t\rangle\langle t|$ with $|t\rangle$ a complete orthogonal basis. The probability of outcome t is on the same quantum system η is

$$\tilde{q}(t) = |\tilde{\psi}(t)|^2 \qquad \tilde{\psi}(t) = \sum_s \psi(s)U(t,s) \tag{D.1}$$

and $\sum_t \tilde{q}(t) = 1$. $U(t,s) = \langle t|s\rangle$ is a unitary matrix and defines the change of coordinates of $|\psi\rangle$ in the two bases. Define $\tilde{\eta} = U\eta U^\dagger$ as the components of η in the new basis. Then $\tilde{q}(t) = \tilde{\eta}(t,t)$. For any U we get a different set of states $|t\rangle$ and a different classical probability distribution $\tilde{q}(t)$. For $U = I$ we get $\tilde{q} = q$.

Appendix E. Mutual information

Here we review classical and quantum mutual information for bipartite systems.

Suppose that the system of interest is described by variables $s = (s_1, \ldots, s_n)$. We model the interactions between these variables by a probability distribution $q(s)$. We know q, we do not know s and the classical (Shannon) entropy $h_c(q) = -\sum_s q(s)\log q(s)$ quantifies the uncertainty in s given q.

Partition the set of variables $s_1, \ldots, s_n$ into two sub sets A and B, with B the complement of A and write $s = (s_A, s_B)$ with s_A and s_B the vector of variables in A and B, respectively. The uncertainty in s_A is given by the entropy $h_c(q_A)$ with $q_A(s_A) = \sum_{s_B} q(s_A, s_B)$ and similar for s_B.

When the sub systems A and B are correlated, observing the state s_A gives us information on s_B. The uncertainty in s_B is given by the entropy of the conditional distribution $q(s_B|s_A)$ and depends on the observed value of s_A. The conditional entropy is defined as the remaining uncertainty in B when observing s_A, averaged over all values

s_A:

$$h_c(B|A) = -\sum_{s_A} q(s_A) \sum_{s_B} q(s_B|s_A) \log q(s_B|s_A) = h_c(q) - h_c(q_A)$$

The remaining uncertainty is less that the original uncertainty in B: $h_c(B|A) \leq h_c(q_B)$. The mutual information between A and B is the difference:

$$I_c(q) = h_c(q_B) - h_c(B|A) = h_c(q_A) + h_c(q_B) - h_c(q) = KL(q|q_A q_B) \quad \text{(E.1)}$$

where KL is the KL divergence defined in Eq. 2 and $q_A q_B$ is the product of marginal distributions $q_A(s_A) q_B(s_B)$. $I_c(q)$ quantifies how much the uncertainty in sub system B is reduced on average by observing s_A. From the last identity it is clear that this also holds with A and B interchanged. $I_c(q)$ satisfies

$$0 \leq I_c(q) \leq \min(h_c(q_A), h_c(q_B)) \quad \text{(E.2)}$$

Consider a density matrix ρ on n variables. The quantum mutual information is defined in analogy with Eq. E.1 as

$$I(\rho) = S(\rho, \rho_A \otimes \rho_B) = h(\rho_A) + h(\rho_B) - h(\rho) \quad \text{(E.3)}$$

with $h(\rho)$ the von Neumann entropy of ρ as defined in Eq. 8, S the relative entropy defined in Eq. 9 and $h(\rho_A), h(\rho_B)$ the entropies of the reduced density matrices♯

$$\rho_A = \text{Tr}_B(\rho) \qquad \rho_B = \text{Tr}_A(\rho) \quad \text{(E.4)}$$

For the quantum mutual information one can derive the bounds

$$0 \leq I(\rho) \leq 2\min(h(\rho_A), h(\rho_B)) \quad \text{(E.5)}$$

The lower bound follows from Klein's inequality $S(\eta, \rho) \geq 0$ which holds for any two density matrices η, ρ [19]. The upper bound follows from the Araki-Lieb inequality [35]

$$h(\rho) \geq |h(\rho_A) - h(\rho_B)| \quad \text{(E.6)}$$

Von Neumann (quantum) entropy and quantum mutual information are counter intuitive from a classical point of view. For classical systems, when a system A is coupled to B, the entropy of the total system AB cannot decrease: $h_c(q_A) \leq h_c(q)$ with q_A the marginal distribution of q on sub system A. The intuitive explanation for this is to equate entropy with uncertainty. Then the uncertainty of the total system is always at least as large as the uncertainty of the sub system. For quantum systems this is not true. The inequality Eq. E.6 allows cases where the entropy $h(\rho_A) > h(\rho)$, with ρ_A the reduced density matrix of ρ on sub system A. Also, classically, the mutual information that sub system A has about B is always less than the total information in B: $I_c(q) \leq h_c(q_B)$ (Eq. E.2). For quantum systems, since it is possible that $h(\rho_A) > h(\rho)$, Eq. E.3 implies that $I(\rho) > h(\rho_B)$, ie. sub system A has more information about B than the total information in sub system B.

♯ In components, write $s = (s_A, s_B)$ with s_A, s_B the variables in A and B, respectively. Then $\rho(s, s') = \rho(s_A, s_B, s'_A, s'_B)$ and $\rho_A(s_A, s'_A) = \sum_{s_B} \rho(s_A, s_B, s'_A, s_B)$.

Appendix F. Singular value decomposition

Here we review the Singular value decomposition (SVD), also known as Schmidt decomposition, and show that $h(\rho_A) = h(\rho_B)$ when ρ is a pure state. Write $s = (s_A, s_B)$ with $s_{A,B}$ the states in A and B, respectively. Write $\psi(s_A, s_B)$ as an $n_A \times n_B$ matrix indexed by s_A, s_B with $n_{A,B}$ the number of states in sub system A, B, respectively. Using the SVD [36] we can write

$$\psi = \sum_{k=1}^{d} \sqrt{\lambda_k} v_k w_k^{\dagger} \qquad \psi(s_A, s_B) = \sum_{k=1}^{d} \sqrt{\lambda_k}\, v_k(s_A) w_k^*(s_B) \tag{F.1}$$

with λ_k positive and $v_k, w_k, k = 1, \ldots, d$ complex vectors of dimension $n_{A,B}$, respectively and $d = \min(n_A, n_B)$. When $n_A > d$ additional orthogonal vectors v_k are defined arbitrarily to make the basis of A complete and similar for B. Then $v_k^{\dagger} v_l = w_k^{\dagger} w_l = \delta_{kl}$. From the normalization of ψ it follows that $\sum_k \lambda_k = 1$. The reduced density matrices are

$$\rho_A = \text{Tr}_B(\rho) = \psi\psi^{\dagger} = \sum_{k=1}^{d} \lambda_k v_k v_k^{\dagger} \qquad \rho_B = \text{Tr}_A(\rho) = \psi^{\dagger}\psi = \sum_{k=1}^{d} \lambda_k w_k w_k^{\dagger}$$

where $\psi\psi^{\dagger}$ and $\psi^{\dagger}\psi$ denote matrix products, summing over the inner index. Thus ρ_A and ρ_B have the same eigenvalues and $h(\rho_A) = h(\rho_B) = -\sum_{k=1}^{d} \lambda_k \log \lambda_k$.

Appendix G. Classical correlations

We first define the classical correlations C and show that $C = I(\rho)/2$ for a pure state [37]. A projective measurement $\{E_a = |\psi_a\rangle \langle \psi_a|\}$ on sub system A transforms the density matrix to

$$\rho \to \rho' = \sum_a |\psi_a\rangle \langle \psi_a| \otimes \langle \psi_a|\rho\psi_a\rangle$$

$\langle \psi_a|\rho\psi_a\rangle$ is a density matrix on B with components on the $|s\rangle$ basis $\langle \psi_a|\rho\psi_a\rangle (s_B, s'_B) = \sum_{s_A, s'_A} \psi_a(s_A)\rho(s_A, s_B; s'_A, s'_B)\psi_a(s'_A)$. A observes outcome a with probability

$$p_a = \text{Tr}\left(|\psi_a\rangle \langle \psi_a|\, \rho\right) = \langle \psi_a|\rho_A\psi_a\rangle$$

with $\rho_A = \text{Tr}_B(\rho)$ and the marginal density matrix on system B is

$$\rho_{B|a} = \frac{\langle \psi_a|\rho\psi_a\rangle}{p_a}$$

The classical correlation is defined in analogy with the classical mutual information as the difference $H(B) - H(B|A)$ (first expression in Eq. E.1), maximized over all possible measurements:

$$C = \max_{\{E_a\}}\left(h(\rho_B) - \sum_a p_a h(\rho_{B|a})\right)$$

If ρ is a pure state $\rho = |\psi\rangle \langle \psi|$, we get $\langle \psi_a|\rho\psi_a\rangle = |\tilde{\psi}_a\rangle \langle \tilde{\psi}_a|$ with $\tilde{\psi}_a(s_B) = \sum_{s_A} \psi_a(s_A)\psi(s_A, s_B)$. Thus, $\rho_{B|a}$ is a pure state and $h(\rho_{B|a}) = 0$ and $C = h(\rho_B)$. Since ρ is a pure state, $I(\rho) = 2h(\rho_B)$ (see Eq. 18). Therefore $C = I(\rho)/2$.

The classical mutual information $I_c(\tilde{q})$ depends on the choice of measurement basis. *We now show that $I_c(\tilde{q}) = C$ in the basis where $\tilde{q}$ is diagonal, with $\tilde{q}$ the transformed probability distribution $\tilde{q}$ Eq. D.1.* Consider the Schmidt decomposition Eq. F.1 of $\psi(s_A, s_B)$. We identify the Schmidt components k with the states t_A, t_B and define

$$U_A(t_A, s_A) = v_{t_A}(s_A) \qquad U_B(t_B, s_B) = w_{t_B}(s_B)$$

Since $U_A(t_A, s_A) = \langle t_A | s_A \rangle$ and similar for B, these define the coordinates of $|t_A\rangle$ and $|t_B\rangle$ on the original $|s\rangle$ basis. With this choice we get from Eq. D.1 that

$$\tilde{\psi}(t_A, t_B) = \sqrt{\lambda_{t_A}} \delta_{t_A, t_B} \qquad \tilde{q}(t_A, t_B) = \lambda_{t_A} \delta_{t_A, t_B} \tag{G.1}$$

The basis transformation makes $\tilde{q}$ diagonal. The classical mutual information is

$$h_c(\tilde{q}) = h_c(\tilde{q}_A) = h_c(\tilde{q}_B) = I_c(\tilde{q}) = -\sum_k \lambda_k \log \lambda_k = h(\eta_A) = C$$

which completes the proof.

We show that $I_c(q) = I(\eta)/2$ when the state of sub system B is determined by the state of A or vice versa. When B is determined by A, $h_c(B|A) = 0$ and the classical mutual information is $I_c(q) = h_c(q_B)$. Because s_B depends deterministically on s_A we have $q(s_A, s_B)q(s_A, s'_B) = q^2(s_A, s_B)\delta_{s_B, s'_B}$. Therefore, the reduced density matrix η_B is diagonal: $\eta_B(s_B, s'_B) = \sum_{s_A} \sqrt{q(s_A, s_B)q(s_A, s'_B)}e^{i\alpha(s_A, s_B) - i\alpha(s_A, s'_B)} = \delta_{s_B, s'_B}q_B(s_B)$ with $q_B(s_B) = \sum_{s_A} q(s_A, s_B)$. Therefore, $h(\eta_B) = h_c(q_B)$ and $I_c(q) = I(\eta)/2$.

As an example consider that sub system B is a single spin i. Then η_B can be written in terms of its spin statistics as

$$\eta_B = \frac{1}{2} \begin{pmatrix} 1 + m_i^z & m_i^x - im_i^y \\ m_i^x + im_i^y & 1 - m_i^z \end{pmatrix}$$

with $m_i^{x,y,z} = \langle \sigma_i^{x,y,z} \rangle_\eta$. If spin i depends deterministically on (a subset of) the other spins, η_B is diagonal and thus $m_i^x = m_i^y = 0$. This occurs for the parity problem in Figure 4 for all i, because the value of each spin is fixed once the value of the other spins are given.

Appendix H. The non-locality of quantum physics

[38] gives a particularly clear explanation of the notion of locality in terms of hidden variables. Consider a physical system that is composed of sub systems. For classical systems, it is natural to describe the state of the system in terms of local variables that describe the state of each of the sub systems. Any composite system that is described by local variables satisfies the so-called Bell inequalities [39, 40]. Composite quantum systems can violate these inequalities and thus violate the intuitive notion of locality.

Suppose we do K measurements $M_k, k = 1, \ldots K$ on a quantum system. The hidden variable idea assumes the existence of additional degrees of freedom that cannot be measured directly, but that determine the measurement outcomes m_k. Repeating the same measurement on the same quantum system may yield different outcomes. Therefore m_k can be treated as a stochastic variable. The outcomes of different

measurements on the same quantum system are correlated and the statistical relation between these outcomes can always be described by a joint probability distribution $p(m_1, \ldots, m_K)$.

Now, suppose that the system is composed of two sub systems A and B. We can make simultaneous measurements $M_{ij} = A_i \otimes B_j$ on the two sub systems, with $A_i, i = 1, \ldots, n_A$ and $B_j, j = 1, \ldots, n_B$ the possible measurements on systems A and B, respectively. Suppose that each measurement A_i and B_j has two possible outcomes. Then each measurement M_{ij} has 4 possible outcomes that are determined by the hidden variable $m_{ij} = (a_{ij}, b_{ij})$ with a_{ij} and b_{ij} binary values: for instance measuring $A_1 \otimes B_1$ yields possible outcomes $m_{11} = (+, +), (+, -), (-, +)$ or $(-, -)$. The correlations between all possible measurement outcomes can be fully captured by a joint probability distribution

$$p(a_{11}, b_{11}, a_{12}, b_{12}, \ldots, a_{n_A n_B}, b_{n_A n_B})$$

on $2K$ binary variables with $K = n_A n_B$.

The hidden variable a_{ij} defines the outcome of the measurement A_i on sub system A when at the same time a measurement B_j on sub system B is made. a_{ij} should obviously depend on the measurement A_i, but it is not so clear whether it should depend on the measurement B_j, in particular when the systems A and B are spatially far removed from each other. The Bell assumption of locality is that the measurement outcome a_{ij} does not depend on the measurement that is performed on system B: $a_{ij} = a_i$ and similar $b_{ij} = b_j$. This reduces the $2n_A n_B$ binary variables to $n_A + n_B$ binary variables and

$$p(a_{11}, b_{11}, a_{12}, b_{12}, \ldots, a_{n_A n_B}, b_{n_A n_B}) \to p(a_1, \ldots, a_{n_A}, b_1, \ldots, b_{n_B})$$

The locality assumption implies that the outcomes of quantum measurement can be described by a probability distribution on local variables. The construction by Bell [39] shows that the correlations produced by measurements on a quantum system can violate this assumption. The counter intuitive conclusion is that the statistics of outcomes of measurement A_i on system A depends on what measurement is performed system B. In other words, the correct statistical description requires the non local variables a_{ij}, b_{ij}.