Learning quantum models from quantum or classical data

H.J. Kappen
Donders Institute, Department of Biophysics, Radboud University, the Netherlands
(Dated: September 25, 2018)

We propose to generalise classical maximum likelihood learning to density matrices. As the objective function, we propose a quantum likelihood that is related to the cross entropy between density matrices. We apply this learning criterion to the quantum Boltzmann machine (QBM), previously proposed by [1]. We demonstrate for the first time learning a quantum Hamiltonian from quantum statistics. For the anti-ferromagnetic Heisenberg and XYZ model we recover the true ground state wave function and Hamiltonian. The second contribution is to apply quantum learning to learn from classical data. Quantum learning uses in addition to the classical statistics also quantum statistics for learning. These statistics may violate the Bell inequality, as in the quantum case. Maximizing the quantum likelihood yields results that are significantly more accurate than the classical maximum likelihood approach in several cases. We give an example how the QBM can learn a strongly non-linear problem such as the parity problem. The solution shows entanglement, quantified by the entanglement entropy.

PACS numbers: 03.67.-a,89.70.Cf
Keywords: quantum machine learning; density matrix theory; entanglement; Bell inequality

Current successes in machine learning [2–4] has ignited interesting new connections between machine learning and quantum physics, loosely referred to as quantum machine learning. Quantum annealing [5, 6] has been successfully applied to optimisation problems that arise in machine learning [7, 8]. Machine learning methods also find useful applications in quantum physics. [9] have shown excellent results using a Boltzmann machine neural network to approximate the ground state of a given quantum Hamiltonian.

In addition to the issue of efficient computing, there are quantum versions of machine learning algorithms [10] that exploit quantum properties for learning and coding. In particular, attempts have been made to extend probability calculus to quantum density operators [11–13]. Recently, [1] have proposed a learning method for density matrices called the quantum Boltzmann machine (QBM). Their approach is to maximise 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 supplementary material for details. For this reason, they introduce a lower bound on the likelihood using the Golden-Thomson inequality and maximise this bound. But this has the disadvantage that parameters of quantum statistics cannot be learned.

In this paper, we also consider the problem of learning a density matrix, by generalising the classical likelihood criterion to a quantum likelihood. Classical learning can be viewed as minimizing the cross entropy between distributions. The cross entropy naturally generalises to density matrices. We thus define quantum learning as an optimisation problem between a density matrix ρ, that defines our model, and a density matrix η that represents the data. This approach does not suffer from the difficulties in [1]. In the case that the model is classical (ie. the model density matrix is diagonal) the quantum likelihood reduces to the classical likelihood.

We restrict ourselves to discrete problems and apply this idea to both quantum and classical systems. We obtain two novel results that connect quantum physics with machine learning. For quantum systems, we consider the problem of learning the Hamiltonian of a quantum spin system from observed stationary quantum statistics. In this case, the data enter into the quantum likelihood through quantum (sufficient) statistics $\langle A \rangle_\eta$ with η the density matrix of the unknown physical system. We consider the case that the quantum system is in a stationary ground state, described by the quantum wave function ψ, and $\eta = \psi\psi'$. We demonstrate that through quantum learning one can recover the exact ground state wave function including the entanglement.

For classical systems, the data define a so-called empirical probability distribution $q(s)$ with s the possible states of the system. From the empirical data distribution q we construct a rank one density matrix $\eta = \psi\psi'$ with $\psi(s) = \sqrt{q(s)}$. This allows us to compute quantum statistics from classical data. The quantum statistics provides features of the data that are not available from low order classical statistics. We give an example where the QBM can learn a problem while the classical Boltzmann machine (BM) cannot. We believe that quantum learning may be useful for classical data analysis.

QUANTUM LEARNING

We first briefly review classical learning. Consider a set of data samples $s^\mu, \mu = 1, \ldots P$, where each sample is a vector of values. The data set can be written as a so-called empirical probability distribution $q(s) = \frac{1}{P} \sum_\mu \delta_{s,s^\mu}$. Consider a probability distribution $p(s)$.

Classical learning can be defined to find p that minimize $S(q,p)$, with S the cross entropy between the distributions q and p

$$S(q,p) = \sum_s q(s) \log \frac{q(s)}{p(s)} \tag{1}$$

Minimizing S with respect to p is equivalent to maximizing the classical likelihood

$$L_c = \sum_s q(s) \log p(s) \tag{2}$$

In the quantum case, we represent both the data and the model as a density matrix. The density matrix is a generalisation of the probability distribution. In discrete state problems, a probability distribution is a vector p with components $p(s) \geq 0$ with s labeling the different states and $\sum_s p(s) = 1$. The density matrix ρ is a Hermitian positive definite matrix with components $\rho(s,s')$ and $\mathrm{Tr}\rho = 1$. ρ has real eigenvalues $\lambda_s \geq 0$ and $\sum_s \lambda_s = 1$. When $\rho(s,s') = p(s)\delta_{s,s'}$ the density matrix reduces to a classical probability distribution.

The notion of expectation value for probability distributions is generalised for density matrices. The expectation value of a matrix A is defined as $\langle A \rangle_\rho = \mathrm{Tr}\,(A\rho)$, 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. In this case we call A a classical statistics. 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 cross entropy between density matrices is defined as [15]

$$S(\eta,\rho) = \mathrm{Tr}\,(\eta \log \eta) - \mathrm{Tr}\,(\eta \log \rho) \tag{3}$$

with $\log \rho$ the matrix logarithm of ρ. It can be shown that $S \geq 0$ and $S = 0$ if and only if $\rho = \eta$ (Klein's inequality, see [15]). Eq. 3 generalises the concept of cross entropy between distributions Eq. 1 to density matrices and reduces to the latter for diagonal density matrices.

When η is the density matrix of the data and ρ is the model density matrix, we define quantum learning to find ρ that minimises S. This is equivalent to maximizing the quantum likelihood

$$L(\rho) = \mathrm{Tr}\,(\eta \log \rho) \tag{4}$$

We consider density matrix models of the form

$$\rho = \frac{1}{Z} e^H \qquad Z = \mathrm{Tr}\left(e^H\right) \qquad H = \sum_r H_r w_r \tag{5}$$

with e^H is the matrix exponential of H. H and H_r are Hermitian matrices and $w = \{w_r, r = 1, \ldots\}$ are given real parameters. In physics, H is the Hamiltonian of the quantum system, and H_r describe the various interaction terms in the Hamiltonian. In classical statistics a model of the form Eq. 5 is known as an exponential family model because H depends linearly on the parameters w_r. Exponential family models have the advantage that the parameters w_r can be estimated through sufficient statistics [16]. The model Eq. 5 is referred to as the quantum Boltzmann machine (QBM) [1].

The quantum likelihood Eq. 4 for the QBM Eq. 5 is

$$L(w) = \langle H \rangle_\eta - \log Z \tag{6}$$

Learning is defined as gradient ascent on the quantum likelihood. Since H is linear in the parameters, $\frac{\partial}{\partial w_r} \langle H \rangle_\eta = \langle H_r \rangle_\eta$. The derivative $\frac{\partial}{\partial w_r} \log Z$ is computed through the Trotter expansion $e^H = \lim_{m\to\infty} \left(e^{H/m}\right)^m$. Then

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

Thus $\frac{\partial}{\partial w_r} \mathrm{Tr}\left(e^H\right) = \mathrm{Tr}\left(H_r e^H\right)$ and $\frac{\partial}{\partial w_r} \log Z = \langle H_r \rangle_\rho$. This gives the learning rule for the QBM:

$$\Delta w_r \propto \frac{\partial L}{\partial w_r} = \langle H_r \rangle_\eta - \langle H_r \rangle_\rho \tag{7}$$

For n quantum spin variables σ_i^k we consider the Hamiltonian

$$H = \sum_{i=1}^n \sum_{k=x,y,z} w_i^k \sigma_i^k + \sum_{i=1,j>i}^n \sum_{k=x,y,z} w_{ij}^k \sigma_i^k \sigma_j^k \tag{8}$$

$\sigma_i^{x,y,z}$ are Pauli spin 1/2 operators (see supplementary material). For this Hamiltonian, the QBM learning rule Eq. 7 becomes

$$\Delta w_i^k = \left\langle \sigma_i^k \right\rangle_\eta - \left\langle \sigma_i^k \right\rangle_\rho \quad \Delta w_{ij}^k = \left\langle \sigma_i^k \sigma_j^k \right\rangle_\eta - \left\langle \sigma_i^k \sigma_j^k \right\rangle_\rho \tag{9}$$

with $k = x, y, z$. The QBM learning rule reduces to classical BM learning when k takes only the value $k = z$ in Eqs. 8 and 9.

LEARNING A QUANTUM HAMILTONIAN

We show that we can accurately learn the density matrix of a given quantum system from data. We consider the situation that we can measure quantum statistics of an unknown quantum system. Our goal is to estimate the quantum Hamiltonian that is the source of these quantum statistics. As an example, we consider the anti-ferromagnetic Heisenberg model in 1 dimension. The ground state wave function ψ of this quantum system is a singlet state that is fully entangled. (This is in contrast to the ground state of for instance a classical anti-ferromagnet for which the ground state is degenerate.) From ψ, the quantum statistics $\left\langle \sigma_i^k \right\rangle_\eta$ and $\left\langle \sigma_i^k \sigma_j^k \right\rangle_\eta$ $k = x, y, z$ are computed (fig. 6) and are used to learn the QBM and the BM.

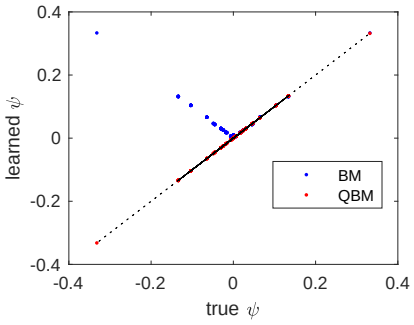

FIG. 1. (Color online) Quantum learning of the 1 dimensional anti-ferromagnetic Heisenberg chain of 10 spins with periodic boundary conditions. The Hamiltonian Eq. 8 has couplings $w_{ij}^{x,y,z} = -1$ for nearest neighbours and $w_{ij}^{x,y,z} = 0$ otherwise ($w^{x,y,z}$ top row figure 6) and external fields $w_i^{x,y,z} = 0$. From H, the ground state wave function ψ and the quantum statistics $\left\langle \sigma_i^k \right\rangle_\eta = 0$ and $\left\langle \sigma_i^k \sigma_j^k \right\rangle_\eta$ are computed (second row figure 6). All statistics are used to learn the quantum Boltzmann machine (Eq. 9) and all classical statistics are used to learn the classical Boltzmann machine (Eq. 15). Scatter plot of the true ground state wave function ψ versus the ground state wave function of the learned Hamiltonian. The QBM correctly reproduces ψ. For comparison $\sqrt{p_{\rm bm}(s)}$, with $p_{\rm bm}(s)$ the solution of the classical BM, is also plotted, which correctly reproduces $|\psi|$ but not the sign. The learned quantum couplings $w_{\rm QBM}^{x,y,z}$ (third row left fig. 6) resemble the original couplings $w^{x,y,z}$ up to an overall scaling and perfectly reproduce the quantum statistics (fourth row left figure 6). The learned classical couplings $w_{\rm BM}$ (third row right fig. 6) do not resemble the original couplings $w^{x,y,z}$ and only reproduce the classical statistics (fourth row right figure 6).

The results in fig. 1 show that the QBM learns a Hamiltonian that has the correct ground state wave function ψ. The solution is close to perfect (quantum likelihood $L = -0.00009$) and correctly reproduces all quantum statistics (supplementary material fig. 6). The couplings w_i^k, w_{ij}^k diverge during learning so that the model density matrix $\rho \propto e^H \rightarrow \psi\psi'$ converges to a rank one solution and H approaches its true value, up to an overall scale factor. The classical BM does not yield a good solution (quantum likelihood $L = -4.19$). The learned classical couplings do not resemble the original couplings $w^{x,y,z}$ (supplementary material fig. 6). Surprisingly, the classical statistics used by the BM are sufficient to reproduces the absolute values of ψ correctly $\sqrt{p_{\rm bm}} = |\psi|$ (fig. 1), but of course cannot reproduce the correct signs. The BM correctly models the classical statistics but not the quantum statistics. In fig. 7 we show that these results generalise to the one dimensional XYZ model.

The results in fig. 1 illustrate that it is important to recover the correct phase (sign) of the wave function in order to model the quantum statistics. Here, these are obtained through the density matrix formalism and a reconstruction of the quantum Hamiltonian. They cannot be obtained by minimizing the classical cross entropy of $|\psi|^2$ directly. However, [14] show that by using multiple bases representations of the wave function simultaneously, one can also obtain the correct phase information.

QUANTUM STATISTICS IN CLASSICAL DATA

We can also apply the QBM to learn classical data. In this case, we construct a rank one density matrix from the data or from its distribution q as follows

$$\eta = \psi\psi' \qquad \psi(s) = \sqrt{q(s)} \tag{10}$$

Quantum learning minimises the quantum cross entropy $S(\eta, \rho)$ with ρ the model density matrix. When ρ is diagonal, $\rho(s, s') = p(s)\delta_{s,s'}$, we obtain $S(\eta, \rho) = S(q, p)$ and quantum learning reduces to the classical case. Thus, quantum learning is the generalisation of classical learning to density matrices. It can find better solutions because the optimisation is over a larger class of models. In addition, it can learn from quantum statistics that have no classical analogue.

The optimal solution ρ for the quantum learning problem is a density matrix that represents the quantum statistics in the data and has no equivalent in terms of a probability distribution. When ρ is (approximately) a rank one matrix, it can be decomposed as $\rho = \phi\phi'$ and one can define the QBM probability distribution $p_{\rm QBM}(s) = \phi^2(s)$. In addition to $S(\eta, \rho)$, we can then use $S(q, p_{\rm QBM})$ as a measure to quantify how well the QBM represents q in classical distribution sense.

All classical and quantum statistics are directly computable from the classical data distribution q. While the classical statistics is linear in q, the quantum statistics are quadratic in $\sqrt{q}$. See supplementary material Quantum Boltzmann machine for details.

In fig. 2top we show that both the BM and the QBM can perfectly learn a classical spin glass data distribution. In fig. 2bottom we show that when the data distribution $q = \psi^2$, with ψ the ground state of a quantum Hamiltonian, the QBM can perfectly learn this problem and the BM fails.

We compare the BM and QBM on the problem suggested in [1]. The data distribution is generated from a mixture of 8 random patterns with 10 % added noise. A typical result is shown in fig. 3. Average results over 10 instances give classical cross entropies $S(q, p_{\rm BM}) = 0.139 \pm 0.05$ and $S(q, p_{\rm QBM}) = 0.043 \pm 0.04$ and quantum cross entropies $S(\eta, \rho_{\rm BM}) = -3.53 \pm 0.11$ and $S(\eta, \rho_{\rm QBM}) = -0.282 \pm 0.14$. The QBM is significantly more accurate than the BM.

In fig. 4 we demonstrate that the QBM can learn the parity relation. Data are generated from a ferro magnet on $s_{1:9}$ and $s_{10} = \prod_{i=1}^{9} s_i$ is the parity of $s_{1:9}$. The classical BM cannot learn this problem while the QBM learns this problem perfectly. The quantum statistics

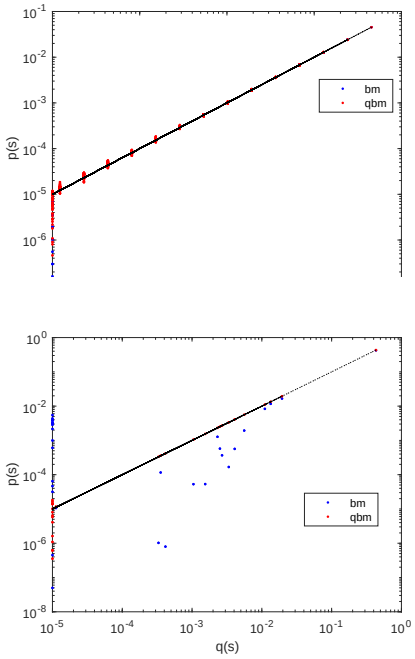

FIG. 2. (Color online). Comparison of QBM and BM solution on a problem of $n = 10$ spins. The figures show scatter plots of $p_{\rm BM}(s)$ and $p_{\rm QBM}(s)$ versus the true probability $q(s)$. Top: Data distribution q is a fully connected spin glass Boltzmann distribution $p(s) \propto \exp\left(\sum_{i>j} w_{ij} s_i s_j\right)$ with $w_{ij} \sim \mathcal{N}\left(0, \frac{1}{\sqrt{n}}\right)$ Gaussian distributed. Classical cross entropies $S(q, p_{\rm BM}) = 5.3e-14$ and $S(q, p_{\rm QBM}) = 1.1e-5$). Quantum cross entropies $S(\eta, \rho_{\rm BM}) = -2.94$ and $S(\eta, \rho_{\rm QBM}) = -0.0015$. Bottom: Data distribution $q = \psi^2$ with ψ the ground state wave function of H (Eq. 8) with $w_{ij}^{x,y,z} \sim \mathcal{N}\left(0, \frac{1}{\sqrt{n}}\right)$ Gaussian distributed and $w_i^{x,y,z} = 0$. Classical cross entropies $S(q, p_{\rm BM}) = 0.125$ and $S(q, p_{\rm QBM}) = 0.0002$). Quantum cross entropies $S(\eta, \rho_{\rm BM}) = -1.50$ and $S(\eta, \rho_{\rm QBM}) = -0.0157$.

signal entanglement in classical data. See supplementary material Entanglement. In particular, the entanglement of a single variable s_i with other variables is maximal when the quantum statistics $m_i^x = \langle \sigma_i^x \rangle_\eta = 0$. In this case s_i is a *arbitrary* deterministic function of (a sub set of) the other variables. In fig. 4 the QBM solution $\langle \sigma_i^x \rangle = 0$ for all spins and is realised in part through the quantum couplings $w_{ij}^{x,y}$ that allow spin flips between pairs of spins in the data.

The quantum statistics can violate the Bell inequalities [17]. Consider three observables a, b, c (classical or quantum). If we assume that their interrelation can be captured by a classical probability distribution, then [18]

$$|\langle ab \rangle - \langle ac \rangle| + \langle bc \rangle \leq 1 \qquad (11)$$

where $\langle \ldots \rangle$ denotes the expectation value (classical or quantum). Therefore, if we can construct a, b, c that violate this inequality, the reverse is true and a, b, c cannot

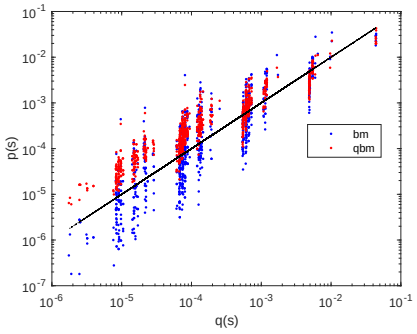

FIG. 3. (Color online). Comparison of QBM and BM solution on a problem of $n = 10$ spins. The data are generated from a mixture of 8 random patterns with 10 % added noise as in [1]. Scatter plot of $p_{\rm BM}(s)$ and $p_{\rm QBM}(s)$ versus the true probability $q(s)$. The quality of the fit is quantified by the classical cross entropy ($S(q, p_{\rm BM}) = 0.125$ and$S(q, p_{\rm QBM}) = 0.0245$). The quantum cross entropy of the BM solution is $S(\eta, \rho_{\rm BM}) = -3.52$ and of the QBM is $S(\eta, \rho_{\rm QBM}) = -0.225$.

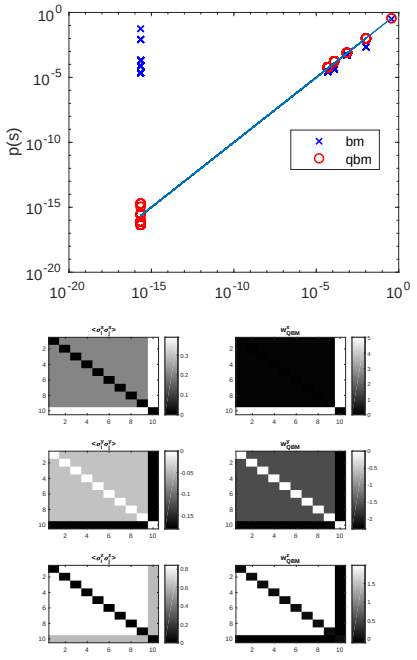

FIG. 4. (Color online). Comparison of QBM and BM solution on a problem of $n = 10$ spins. Data are generated from a fully connected ferro magnet on $s_{1:9}$ with coupling $w = 1/9$ and no external fields, and $s_{10} = \prod_{i=1}^{9} s_i$ is the parity of $s_{1:9}$. Top: scatter plot of $p_{\rm BM}(s)$ and $p_{\rm QBM}(s)$ versus the true probability $q(s)$. The quality of the fit is quantified by the classical cross entropy ($S(q, p_{\rm BM}) = 0.459$ and $S(q, p_{\rm QBM}) = 0.0021$). The quantum cross entropy of the BM solution is $S(\eta, \rho_{\rm BM}) = -2.735$ and of the QBM is $S(\eta, \rho_{\rm QBM}) = -0.404$. Bottom: Quantum statistics and couplings of the QBM.

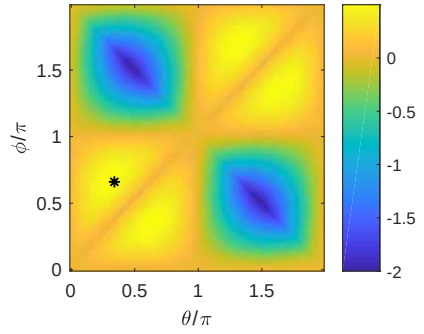

FIG. 5. Violation of the Bell inequality Eq. 11 in classical data. The data are two fully correlated binary variables s_1, s_2. The observables are $a = \sigma_1^z, b = \sigma_1^x \sin\theta + \sigma_1^z \cos\theta, c = \sigma_2^x \sin\phi + \sigma_2^z \cos\phi$. The figure shows $B(\theta, \phi)$ (Eq. 25) as a function of θ, ϕ. When $B > 0$, a, b, c violate the Bell inequality. Maximal violation for $(\theta, \phi) = \left(\frac{\pi}{3}, \frac{2\pi}{3}\right)$ indicated by black star. See supplementary material Bell inequality for details.

be described simultaneously by a classical distribution. When the classical data, as given by q, is represented as a rank one density matrix η, the statistics of η can violate the Bell inequality. In figure 5 we show an adaptation of a well known example, with η constructed from a probability distribution for two fully correlated binary variables. The conclusion is that the three observables a, b, c cannot be modelled by a classical probability distribution.

For large problems, 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 fig. 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. But, clearly, this does not scale to large problem instances. As in classical BM learning, one can apply various approximate inference methods to estimate $\langle H_r \rangle_\rho$. When estimating a quantum Hamiltonian from zero temperature statistics, one can use zero temperature methods that minimize the Raleigh quotient to estimate the ground state wave function with a particular variational Anzats for the wave function [19, 20]. In particular, [9] use a classical Boltzmann machine as the variational ansatz for the wave function that shows great potential. In general, when the density matrix has finite temperature and is not a rank one matrix, one may need to estimate the statistics using diffusion Monte Carlo [21, 22] or mean field methods [23, 24].

This research was funded in part by ONR Grant N00014-17-1-2569. I wish to thank Marta Bela for support and good discussions.

[1] M. H. Amin, E. Andriyash, J. Rolfe, B. Kulchytskyy, and R. Melko, arXiv preprint arXiv:1601.02036 (2016).
[2] Y. LeCun, Y. Bengio, and G. Hinton, Nature **521**, 436 (2015).
[3] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, Nature **518**, 529 (2015).
[4] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, Nature **529**, 484 (2016).
[5] T. Kadowaki and H. Nishimori, Physical Review E **58**, 5355 (1998).
[6] B. Heim, T. F. Rønnow, S. V. Isakov, and M. Troyer, Science **348**, 215 (2015).
[7] S. H. Adachi and M. P. Henderson, arXiv preprint arXiv:1510.06356 (2015).
[8] M. Benedetti, J. Realpe-Gómez, R. Biswas, and A. Perdomo-Ortiz, Physical Review A **94**, 022308 (2016).
[9] G. Carleo and M. Troyer, Science **355**, 602 (2017).
[10] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, Nature **549**, 195 (2017).
[11] N. J. Cerf and C. Adami, Physical Review Letters **79**, 5194 (1997).
[12] N. J. Cerf and C. Adami, Physical Review A **60**, 893 (1999).
[13] R. Schack, T. A. Brun, and C. M. Caves, Physical Review A **64**, 014305 (2001).
[14] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko, and G. Carleo, arXiv preprint arXiv:1703.05334 (2017).
[15] E. Carlen, Entropy and the quantum **529**, 73 (2010).
[16] E. B. Andersen, Journal of the American Statistical Association **65**, 1248 (1970).
[17] J. S. Bell, in *John S Bell On The Foundations Of Quantum Mechanics* (World Scientific, 2001) pp. 1–6.
[18] N. Cerf and C. Adami, Physical Review A **55**, 3371 (1997).
[19] W. L. McMillan, Physical Review **138**, A442 (1965).
[20] J. Carlson, S. Gandolfi, F. Pederiva, S. C. Pieper, R. Schiavilla, K. Schmidt, and R. B. Wiringa, Reviews of Modern Physics **87**, 1067 (2015).
[21] D. Ceperley and B. Alder, Science **231**, 555 (1986).
[22] D. M. Ceperley, Reviews of Modern Physics **67**, 279 (1995).
[23] J. Matera, R. Rossignoli, and N. Canosa, Physical Review A **82**, 052332 (2010).
[24] E. Dominguez and R. Mulet, Physical Review B **97**, 064202 (2018).
[25] D. Ackley, G. Hinton, and T. Sejnowski, Cognitive Science **9**, 147 (1985).
[26] H. Araki and E. H. Lieb, in *Inequalities* (Springer, 2002) pp. 47–57.
[27] J. A. Miszczak, International Journal of Modern Physics C **22**, 897 (2011).

Boltzmann machine

For convenience we review the maximum likelihood learning for the classical Boltzmann machine. The state space consists of all vectors $s = (s_1, \ldots, s_n)$ with $s_i = \pm 1$ binary spin variables. The Boltzmann machine derives its name from the fact that the probability distribution $p(s|w)$ that is learned is a Boltzmann distribution

$$p(s|w) = \frac{e^{H(s|w)}}{Z} \qquad Z(w) = \sum_s e^{H(s|w)} \tag{12}$$

with $H(s|w) = \sum_r H_r(s) w_r$ linear in w_r. $H_r(s)$ are interaction terms involving a 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. 12 the likelihood Eq. 2 becomes

$$L_c(w) = \langle H \rangle_q - \log Z \tag{13}$$

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

$$\Delta w_k \propto \frac{\partial L}{\partial w_r} = \langle H_r \rangle_q - \langle H_r \rangle_p \tag{14}$$

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 probability distribution p. Learning stops when the gradients are zero, ie. when the statistics defined by H_r are equal: $\langle H_r \rangle_q = \langle H_r \rangle_p$. When $H(s|w) = \sum_i w_i s_i + \sum_{i>j} w_{ij} s_i s_j$, the learning rule becomes

$$\Delta w_i \propto \langle s_i \rangle_q - \langle s_i \rangle_p \qquad \Delta w_{ij} \propto \langle s_i s_j \rangle_q - \langle s_i s_j \rangle_p \tag{15}$$

Eq. 15 is the well-known Boltzmann Machine learning rule [25].

Quantum Boltzmann machine details

In the Hamiltonian Eq. 8, $\sigma_i^{x,y,z}$ are 2×2 Pauli spin matrices. Since $\sigma_i^z = \mathrm{diag}(1, -1)$, 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 we use as basis the tensor product of the two bases vectors $|s_i = \pm 1\rangle$ of each spin. The 2^n basis vectors are denoted as $|s\rangle = |s_1, \ldots, s_n\rangle$. On this basis, H has matrix elements

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

with $F_i s$ the state s with spin i flipped to $-s_i$ and all other spins unchanged. For this Hamiltonian, the QBM learning rules Eq. 7 become Eqs. 9

The expectations $\langle \sigma_i^k \rangle_\eta$ and $\langle \sigma_i^k \sigma_j^k \rangle_\eta$ only depend on the data and can be easily computed. When $k = z$, these are the classical statistics in the data $\langle \sigma_i^z \rangle_\eta = \langle s_i \rangle_q$ and $\langle \sigma_i^z \sigma_j^z \rangle = \langle s_i s_j \rangle_q$. Using the definition of expectations for density matrices, the other statistics are computed as

$$\langle \sigma_i^x \rangle_\eta = \sum_s \sqrt{q(F_i s) q(s)} \qquad \langle \sigma_i^y \rangle = 0 \tag{16}$$

$$\langle \sigma_i^x \sigma_j^x \rangle_\eta = \sum_s \sqrt{q(F_i F_j s) q(s)} \qquad \langle \sigma_i^y \sigma_j^y \rangle_\eta = -\sum_s s_i s_j \sqrt{q(F_i F_j s) q(s)} \tag{17}$$

Since η is real symmetric, the expectation of complex Hermitian observables such as $\langle \sigma_i^y \rangle_\eta$ are zero (but not $\langle \sigma_i^y \sigma_j^y \rangle$).

The classical statistics can be computed linear in the size of the data N. For the quantum statistics we compute the unique occurrence $N(s^\mu) > 0$ of each sample s^μ in the data set and $q(s^\mu) = \frac{N(s^\mu)}{N}$. This computation is $\mathcal{O}(N \log N)$. Subsequently, we can compute each quantum statistics linear in N.

QBM learning proposed by [1]

[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 $p(s) = \mathrm{Tr}\left(\Lambda_s\rho\right)$ and

$$L = \sum_s q(s)\log \mathrm{Tr}\left(\Lambda_s\rho\right) = \sum_s q(s)\log \mathrm{Tr}\left(\Lambda_s e^H\right) - \log Z \tag{18}$$

Note, that this expression differs from the quantum likelihood Eq. 4 in the first term $\langle H\rangle_\eta$ only. Since H is linear in w_r, the gradient of $\langle H\rangle_\eta$ is easy and gives the clamped statistics $\langle H_r\rangle_\eta$. As a result, the learning rule Eq. 7 is the generalisation of the classical BM learning rule Eq. 14 to quantum statistics. Instead, the gradient of the first term in Eq. 18 is given by

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

Because of Λ_s the time integration remains and cannot be easily evaluated.

Entanglement

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

$$S(\rho) = -\mathrm{Tr}\left(\rho\log\rho\right) \tag{19}$$

It is easy to show that $S(\rho) = -\sum_s \lambda_s\log\lambda_s$, with $\lambda_s \geq 0$ the eigenvalues of ρ. The entropy is maximal when all λ_s are equal. The minimal entropy $S(\rho) = 0$ when $\lambda_s = \delta_{s,s^*}$ for some state s^*. In this case ρ is a rank one matrix and can be written as $\rho = \psi\psi'$ with $'$ Hermitian conjugate. ψ 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 $S(\rho) = -\sum_s p(s)\log p(s)$.

Entanglement is a feature of a quantum system that has no classical analogue. Consider a density matrix ρ on n variables and consider a sub set of variables A and its complement B. The entanglement between A and B is defined as the quantum mutual information

$$I_{AB} = S(\rho_A) + S(\rho_B) - S(\rho)$$

with $S(\rho)$ the quantum entropy of ρ and $S(\rho_A), S(\rho_B)$ the quantum entropies of the marginal density matrices

$$\rho_A = \mathrm{Tr}_B\left(\rho\right) \qquad \rho_B = \mathrm{Tr}_A\left(\rho\right) \tag{20}$$

on the sub sets A, B, respectively [11]. (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 $\rho_A(s_A, s'_A) = \sum_{s_B}\rho(s_A, s_B, s'_A, s_B)$.)

The quantum mutual information is the generalisation to density matrices of the classical mutual information

$$I^c_{AB} = H(p_A) + H(p_B) - H(p)$$

with p a probability distribution on all variables and $p_{A,B}$ the marginal probability distributions on variables A and B, respectively and $H(p)$ the classical entropy of distribution p. The classical mutual information I^c_{AB} satisfies

$$0 \leq I^c_{AB} \leq \min(H(p_A), H(p_B)) \tag{21}$$

The upper bound on I^c_{AB} can be easily derived by noting that the conditional entropy

$$-\sum_{s_A} p(s_A)\sum_{s_B} p(s_B|s_A)\log p(s_B|s_A) = H(p) - H(p_A) \geq 0$$

The lower bound follows when because I^c_{AB} is the KL divergence between p and $p_A p_B$, which is non-negative.

The quantum mutual information may violate the bound Eq. 21. Instead, we have

$$0 \leq I_{AB} \leq 2\min(S(\rho_A), S(\rho_B))$$

The lower bound follows from the fact that $I_{AB} = S(\rho, \rho_A\otimes\rho_B) \geq 0$ with S the cross entropy Eq. 3. The upper bound follows from the Akari-Lieb inequality [26]

$$S(\rho) \geq |S(\rho_A) - S(\rho_B)|$$

Note that this inequality allows cases where $S(\rho_A) = S(\rho_B) > 0$ and $S(\rho) = 0$. This would be classically forbidden. An example is the fully entangled two spin system, where $S(\rho) = 0$ and $S(\rho_A), S(\rho_B) > 0$.

In the case of quantum learning, $\eta = \psi\psi'$ with $\psi(s) = \sqrt{q(s)}$ is a rank 1 density matrix and $S(\eta) = 0$. Since ψ is a pure state, it allows a Schmidt or singular value decomposition [27] as

$$\psi = \sum_i \sqrt{\lambda_i} v_i w_i'$$

with v_i, w_i orthogonal vectors in A and B, respectively: $v_i' \cdot v_j = w_i' \cdot w_j = \delta_{ij}$ and $\lambda_i > 0$. Therefore

$$\eta_A = \psi\psi' = \sum_i \lambda_i v_i v_i' \qquad \eta_B = \psi'\psi = \sum_i \lambda_i w_i w_i'$$

and $S(\eta_A) = S(\eta_B) = -\sum_i \lambda_i \log \lambda_i$. Thus, the quantum entropies of the two subsystems are identical and the entanglement is $I_{AB} = 2S(\eta_A)$.

It is useful to compute the entanglement of a single variable s_i, ie. $A = i$ and B are all other variables. In general, the density matrix of a binary variable $s = \pm 1$ can be written as a linear combination of Pauli spin matrices $\sigma^{x,y,z}$ and the identity matrix:

$$\rho = \frac{1}{2}\begin{pmatrix} 1 + m^z & m^x - im^y \\ m^x + im^y & 1 - m^z \end{pmatrix} \tag{22}$$

with $m^{x,y,z}$ the real expected value of $\sigma^{x,y,z}$: $m^k = \mathrm{Tr}(\sigma^k \rho), k = x, y, z$. The eigenvalues of ρ are

$$\lambda_{1,2} = \frac{1}{2}(1 \pm m) \qquad m = \sqrt{(m^x)^2 + (m^y)^2 + (m^z)^2}$$

Thus ρ is positive provided that $m \leq 1$. The entropy of ρ is

$$S(\rho) = -\lambda_1 \log \lambda_1 - \lambda_2 \log \lambda_2 = \log 2 - \frac{1}{2}\log(1 - m^2) - \frac{1}{2} m \log \frac{1+m}{1-m} \tag{23}$$

In the case of classical data, $m_i^y = 0$ and the data density matrix of s_i is

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

with $m_i^z = \langle s_i \rangle_q$ the classical expectation value of s_i in the data distribution q and

$$m_i^x = \langle \sigma_i^x \rangle_\eta = 2 \sum_{s_{\setminus i}} q(s_{\setminus i}) \sqrt{q(s_i|s_{\setminus i}) q(-s_i|s_{\setminus i})}$$

Note, that $0 \leq m_i^x \leq 1$.

From Eq. 23 we infer that $S(\eta_i)$ is maximal when m is minimal. Thus, for given mean value m_i^z, the entanglement is maximal when $m_i^x = 0$. $m_i^x = 0$ iff s_i is a deterministic function of (a sub set of) the other variables. This is because m_i^x (Eq. 16) is a sum of non-negative terms and can only be zero when all terms are zero. This can only be true when for each $s_{\setminus i}$ for which $q(s_{\setminus i}) > 0$, either $q(s_i = 1|s_{\setminus i}) = 0$ or $q(s_i = -1|s_{\setminus i}) = 0$.

Bell inequalities

Consider two fully correlated binary variables $s = (s_1, s_2)$ with joint probability distribution

$$q = \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \psi = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{24}$$

We consider the three observations:

$$a = \sigma_1^z, \qquad b = \sigma_1^x \sin\theta + \sigma_1^z \cos\theta, \qquad c = \sigma_2^x \sin\phi + \sigma_2^z \cos\phi$$

The measurement a is in the z-direction. The measurements b, c make an angle θ, ϕ with the z direction in the $x-z$ plane, respectively. Measurements a, b are on variable s_1 and measurement c is on variable s_2. We compute the statistics in the density matrix $\eta = \psi\psi'$:

$$\begin{aligned}
\langle ab\rangle_\eta &= \sum_{s_1' s_2', s_1, s_2} \psi(s_1', s_2')\,(ab)_{s_1', s_1}\,\psi(s_1, s_2) = \cos\theta \\
\langle ac\rangle_\eta &= \sum_{s_1' s_2', s_1, s_2} \psi(s_1', s_2')\,(a)_{s_1', s_1}\,(c)_{s_2', s_2}\,\psi(s_1, s_2) = \cos\phi \\
\langle bc\rangle_\eta &= \sum_{s_1' s_2', s_1, s_2} \psi(s_1', s_2')\,(b)_{s_1', s_1}\,(c)_{s_2', s_2}\,\psi(s_1, s_2) = \cos(\theta-\phi)
\end{aligned}$$

Eq. 11 becomes

$$B(\theta,\phi) = |\cos\theta - \cos\phi| - 1 + \cos(\theta-\phi) \leq 0 \tag{25}$$

$B(\theta,\phi)$ is plotted in fig. 5 showing maximal violation for instance when $(\theta,\phi) = \left(\frac{\pi}{3}, \frac{2\pi}{3}\right)$. The three operators that maximally violate the Bell inequalities are

$$a = \sigma_1^z \qquad b = \frac{1}{2}\sqrt{3}\sigma_1^x + \frac{1}{2}\sigma_1^z \qquad c = \frac{1}{2}\sqrt{3}\sigma_2^x - \frac{1}{2}\sigma_2^z \tag{26}$$

AFH model. Details of fig. 1

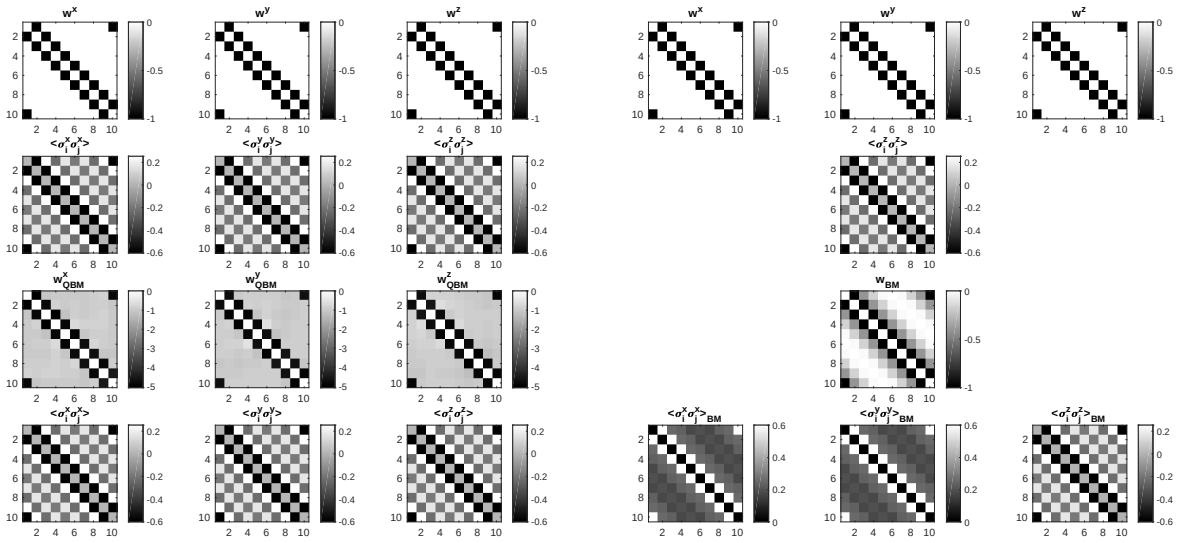

FIG. 6. Details of figure 1. Left figure: QBM. Right figure: BM. Top row: Couplings $w_{ij}^{x,y,z}$ of the anti-ferromagnetic Heisenberg model. Second row: All quantum statistics $\left\langle \sigma_i^k \sigma_j^k \right\rangle, k = x, y, z$ are used to train the QBM. Only the classical statistics $\left\langle \sigma_i^z \sigma_j^z \right\rangle$ are used to train the BM. Third row: The learned couplings $w_{\mathrm{QBM}}^k, k = x, y, z$ resemble the original coupling $w^{x,y,z}$ up to an overall scaling. The learned classical couplings w_{BM} do not resemble the original couplings $w^{x,y,z}$. Fourth row: The statistics computed from the models. The QBM reproduces all quantum statistics correctly. The classical BM only reproduces the classical statistics correctly.

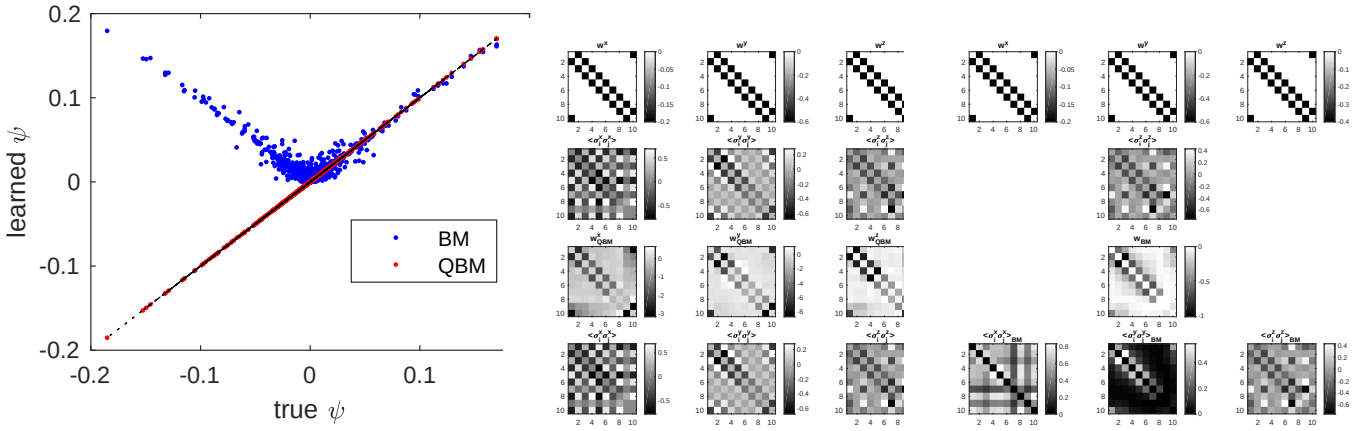

FIG. 7. (Color online). Quantum learning of the 1 dimensional anti-ferromagnetic XYZ chain of 10 spins with periodic boundary conditions. The Hamiltonian has couplings $w_{ij}^x = -0.2, w_{ij}^y = -0.6, w_{ij}^z = -0.4$ for nearest neighbours and $w_{ij}^{x,y,z} = 0$ otherwise ($w^{x,y,z}$ top row middle and right figure) and external fields $w_i^{x,y,z}$ random Gaussian. Left: Scatter plot of the true ground state wave function ψ versus the ground state wave function of the learned Hamiltonian. The QBM correctly reproduces ψ. For comparison $\sqrt{p_{\rm bm}(s)}$, with $p_{\rm bm}(s)$ the solution of the classical BM, which is incorrect. Quantum cross entropies $S(\eta, \rho_{\rm BM}) = -5.04$ and $S(\eta, \rho_{\rm QBM}) = -0.0044$. Middle: QBM solution. Right: BM solution. Top row: Couplings $w_{ij}^{x,y,z}$ of the anti-ferromagnetic Heisenberg model. Second row: All quantum statistics $\left\langle \sigma_i^k \sigma_j^k \right\rangle, k = x, y, z$ are used to train the QBM. Only the classical statistics $\left\langle \sigma_i^z \sigma_j^z \right\rangle$ are used to train the BM. Third row: The learned couplings $w_{\rm QBM}^k, k = x, y, z$ resemble the original coupling w_{XYZ} up to an overall scaling. The learned classical couplings $w_{\rm BM}$ do not resemble the original couplings w_{XYZ}. Fourth row: The statistics computed from the models. The QBM reproduces all quantum statistics correctly. The quantum statistics computed from the classical BM density matrix $\psi\psi'$ with $\psi(s) = \sqrt{p_{\rm bm}(s)}$ correctly capture $\left\langle \sigma_i^z \sigma_j^z \right\rangle$ but not $\left\langle \sigma_i^x \sigma_j^x \right\rangle$ and $\left\langle \sigma_i^y \sigma_j^y \right\rangle$.