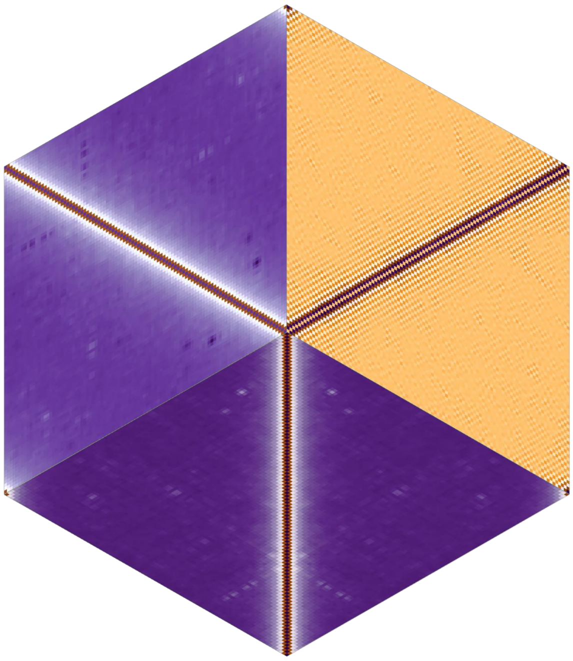
Training large Quantum Boltzmann Machines with Neural Quantum States

Author
Manu Compen

Supervisor
prof. dr. H.J. Kappen

Second Assessor
dr. J.H. Mentink



Radboud University

IN·DEI·NOMINE·FELICITER

A thesis presented for the degree of
Master of Science

December 19, 2019

Preface

During my bachelor's in Physics and Astronomy and subsequently my master's in Physics of Molecules and Materials, I have always enjoyed a fascination for digital solutions to physical problems. Writing a thesis about the combination of intelligent algorithms and quantum physics was a delight. In addition, it has been particularly fulfilling bringing these algorithms to life by the development of my own software.

I would first like to thank my supervisor, prof. dr. H.J. Kappen, for his inexhaustible willingness to help, discuss, explain and philosophize. It has been a joy and a great learning experience trying to find our way through a stubborn subject. Roeland, it has been great having you as a friend and colleague. I will look back fondly on many esoteric discussions about our shared passions ranging from mathematical physics to the automation of software documentation. Onno, during our intensive collaboration we spent hours discussing the QBM. The last word has yet to be spoken.

About the titlepage

The figure on the titlepage is an adaptation of the figures in Appendix B. In the figure, the heatmaps of spin correlations of the Heisenberg model with 100 spins are projected on their respective axis.

Contents

Chapter 1

Introduction

Since the inception of computer technology researchers have been trying to bring their machines to life. This inclination emerges quite naturally from the observation that computer machines are capable of performing many logical operations each second. The ability to harness this power to make creative and thoughtful decisions, similar to how the brain operates, would lead to extremely powerful technology.

The *fantasy* to attribute intelligent characteristics to inanimate objects far precedes information technology. Early examples can be found in Greek myths in the form of the bronze robot Talos, the fire-stealing Prometheus and fembot Pandora. Although mythological figures, it is argued that these ancient robots were not described as being brought to life by the gods; rather, they were created with methods and materials accessible to mankind much like their contemporary nonfiction equivalents [1].

Modern day ideas about artificial intelligence have escaped the realms of myths and fantasy. Applications are widespread, ranging from facial recognition technology and disease prediction in healthcare [2] to automated driving, movement in robotics and language processing [3]. Moreover, contemporary AI research is aimed at coming up with more general solutions that show human performance on a wide range of problems, instead of highly specialized tasks such as the recognition of handwritten digits. These more high-level and generally applicable AI algorithms are called *strong* AI. The transformation from weak to strong AI requires an evolution towards a more abstract form of intelligence technology. A preliminary demonstration of such intelligence is the autonomous development of strategies in games with many degrees of freedom [4].

It may be tempting to interpret the weak to strong AI transition as purely a software problem. However, software not being a match for natural intelligence is not the only problem. Hardware is also not up to par with the brain in terms of parallelism and power consumption. Simulating the brain with a simplified neuronal model has been done with conventional hardware at a rate of 1500s per 1s of brain activity [5]. This required 12GW of power: a gap of twelve orders of magnitude compared to the 20W of a human brain.

This brings up another facet. New hardware solutions are not only needed for powerful AI: they are also urgently needed in order to reduce the global ICT power consumption. Today, the CO_2 carbon footprint of the entire ICT ecosystem is equal to the footprint of the aviation industry. In 2030, ICT power consumption is forecasted to be responsible for $10\% - 20\%$ of global energy use [6].

One approach to energy efficient hardware is parting ways with the prevailing von Neumann architecture, which dictates that information processing and storage are handled separately. Straying away from this information technology paradigm has been popularized by neuromorphic computing research. A recent notable effort in this regard is the development of neuromorphic chips such as Loihi [7]. Interestingly, reducing ICT power consumption is not necessarily purely hardware-driven: machine learning can be employed to e.g. increase cooling efficiency of data centres significantly [8].

In order to make computing technology smarter and more efficient, a new approach to software as well as hardware is clearly needed. Quantum computing is a strong contender to fulfil this role. At variance with the classical world, quantum descriptions of a physical system are inherently probabilistic, famously illustrated by the Schrödinger's cat Gedankenexperiment. Moreover, quantum systems exhibit *entanglement*. Entanglement induces quantum correlations that have no classical equivalent. Naturally, these phenomena open doors to capture, transmit and receive information in different (and sometimes more time- and energy efficient) ways. The power of quantum algorithms is particularly clear when they directly translate intractable classical problems to tractable quantum problems. Shor's prime factorization algorithm [9] is perhaps the most tangible example of such a translation.

Oftentimes, the added value of quantum physics in machine learning is more subtle. In these cases, one is aimed at quantum *usefulness* rather than supremacy. While not directly solving a classical intractibility, the Quantum Boltzmann machine (*QBM*) provides a rich probability model due to its quantum features. In some cases, these quantum characteristics are a necessity in order to accurately learn a (classical) probability distribution. Moreover, it allows one to perform quantum tomography on *mixed* quantum states. The QBM (published by M. Amin in 2018 [10]) is a quantum extension of the classical Boltzmann machine. This thesis is centred around this particular quantum algorithm. A background of the physical theory relevant to this Quantum Machine Learning (*QML*) invention is given in chapter 2. Fundamental to QML in general is the probabilistic *qubit* as replacement for the deterministic bit. Qubits are the basic units of information in quantum computers. They are experimentally implemented by spin-$\frac{1}{2}$ particles and other two-level systems. Some hallmark quantum spin systems are described in chapter 3. The QBM is laid out in chapter 5.

In physics, theory and experiment are often out of sync. This is no different in quantum computing, where the imagination of theoreticians is often at odds with experimental limits. An important factor in this regard is *quantum decoherence* due to imperfect isolation of a quantum system with respect to its environment. For this reason, quantum machine learning is often performed on classical hardware, merely simulating perfect quantum behaviours.

The quasi-quantum approximative approach has its pitfalls, since it inherently relies on classically feasible projections of an exponentially scaling quantum state space. Training the Quantum Boltzmann Machine involves inferring spin expectation values for randomly connected spin systems. Determining exact quantum correlations of a spin system without quantum hardware requires exact diagonalization of the Hamiltonian. For qubit systems larger than 30 spins this becomes intractable on conventional hardware[1]. Due to the lack of capacity to represent the entire state space, sampling becomes an inevitability for systems of this size. To this end, the popular *Markov Chain Monte Carlo* sampling method is described in chapter 4.

Spin expectation values of QBM systems can be approximated by their ground state statistics in the zero-temperature (or large spectral gap) limit. This opens the way for variational ground state methods. Although variational schemes have been implemented succesfully on quantum hardware [12][13], they are currently only implemented for small systems. Therefore, we approach this classical intractability with the Neural Quantum State (*NQS*) algorithm. This variational algorithm (published by G. Carleo and M. Troyer in 2017 [14]) provides an efficient way to learn a neural network embedding of a quantum state, given a Hamiltonian. Subsequently, this representation can be used to efficiently sample spin correlations. This is where the QBM and NQS algorithms meet each other. The NQS is described and tested on a variety of models in chapter 6 with its application to the QBM in mind.

Quantum Machine Learning is a bidirectional field of research. It entails the usage of quantum physics to the benefit of machine learning and vice versa, demonstrated by the QBM and NQS respectively. The QBM has the potential to be an extremely useful and widely applicable resource for scientific research in order to model large real-life data sets with complex structures. Implemented on quantum hardware, it allows learning a probability distribution by direct measurement instead of sampling. For quantum physics in particular the QBM provides a means to determine or model the quantum state from measurements; a solution to one of the greatest challenges in this field. Solving the intractability problem of the QBM with the NQS would therefore be of great scientific relevance. This leads to the research question of this thesis:

"Can the Neural Quantum State algorithm be used to train the quantum probabilistic model of the Quantum Boltzmann Machine?"

1.1 Software

The used software to learn NQS's and train QBMs was written in Julia [15]. The NQS software has been compiled in the open source package *NeuralQuantumState.jl* [16] and can be added from any Julia (v1) REPL.

[1] For systems with large amounts of symmetries, the effective Hilbert space is reduced. Using functional matrix representations in addition to these symmetries allows diagonalization up to 50 spins [11].

Chapter 2

Density Matrices

Probability has many faces in the context of quantum physics. On the one hand we have classical probability descriptions of systems, such as the Boltzmann distribution for classical ensembles, while on the other hand we have probability amplitudes describing quantum superpositions of states. Moreover, there is a classical interpretation of quantum probability according to Born's rule. In order to keep track of both classical and quantum probability information of a system, an all-encompassing description is needed. This description is given by density operator theory [17][18].

2.1 Pure and mixed quantum systems

Consider an ensemble consisting of a statistical mixture of N subsystems. Each subsystem is described by a (normalized) wavefunction $|\psi_i\rangle$, $i = 1, \ldots, N$. The subsystems are not necessarily described by orthogonal states: in general $\langle\psi_i|\psi_j\rangle \neq 0$. A simple example of this is a statistical mixture of three two-level systems described by the pure states $|\uparrow\rangle$, $|\downarrow\rangle$ and $\frac{1}{\sqrt{2}}(|\downarrow\rangle \pm |\uparrow\rangle)$. For an orthonormal set of discrete eigenstates $\{|\phi\rangle\}$, corresponding to a complete set of commuting observables, normalized wavefunctions can be decomposed according to

$$|\psi\rangle = \overbrace{\sum_i |\phi_i\rangle\langle\phi_i|}^{\hat{\mathbb{1}}} |\psi\rangle = \sum_i \langle\phi_i|\psi\rangle |\phi_i\rangle \equiv \sum_i c^{\psi}_{\phi_i} |\phi_i\rangle\,, \tag{2.1}$$

where c denotes a complex-valued scalar. The expectation value of an observable $\hat{A}$ w.r.t. a pure state ψ is denoted

$$\langle\hat{A}\rangle_{\psi} = \langle\psi|\hat{A}|\psi\rangle = \sum_{i,j} \langle\phi_j|\psi\rangle\langle\psi|\phi_i\rangle\,\langle\phi_i|\hat{A}|\phi_j\rangle = \sum_{i,j} (c^{\psi}_{\phi_i})^* c^{\psi}_{\phi_j}\,\langle\phi_i|\hat{A}|\phi_j\rangle\,. \tag{2.2}$$

The notion of statistical uncertainty is incorporated by denoting the average over elements $|\psi_k\rangle$ in the mixture as

$$\langle\hat{A}\rangle = \sum_{k=1}^{N} W_k \langle\hat{A}\rangle_{\psi_k}. \tag{2.3}$$

This could describe e.g. a statistical mixture consisting of (pure) right- and left polarized photons with $W_R = 0.5$ and $W_L = 0.5$. In any case, $W_i \in [0, 1]$ and

$\sum_k W_k = 1$. The ensemble average gives rise to the density operator, identified by

$$\langle \hat{A} \rangle = \sum_{k=1}^{N} \sum_{i,j} W_k \langle \phi_j | \psi_k \rangle \langle \psi_k | \phi_i \rangle \ \langle \phi_i | \hat{A} | \phi_j \rangle \tag{2.4}$$

$$= \sum_{i,j} \langle \phi_j | \overbrace{\Big(\sum_{k=1}^{N} W_k | \psi_k \rangle \langle \psi_k | \Big)}^{\hat{\rho}} | \phi_i \rangle \ \langle \phi_i | \hat{A} | \phi_j \rangle \, . \tag{2.5}$$

The density matrix in the ϕ-basis is given by

$$\rho_{ji} = \langle \phi_j | \hat{\rho} | \phi_i \rangle = \sum_{k=1}^{N} W_k (c_{\phi_i}^{\psi_k})^* c_{\phi_j}^{\psi_k}. \tag{2.6}$$

The density operator identified in equation 2.5 is Hermitian ($\hat{\rho}^* = \hat{\rho}$) since W_k is real. From equation 2.6 it follows that $(c_{\phi_i}^{\psi})^* c_{\phi_j}^{\psi} = c_{\phi_i}^{\psi} (c_{\phi_j}^{\psi})^*$. Having identified the density operator, the ensemble average can be written as a sum over the orthonormal basis set $\{|\phi\rangle\}$, giving rise to the trace property (note the identity operator in equation 2.5)

$$\langle \hat{A} \rangle = \sum_{j} \langle \phi_j | \hat{\rho} \hat{A} | \phi_j \rangle = \mathrm{Tr}(\hat{\rho} \hat{A}). \tag{2.7}$$

This surprisingly clean property is an essential result of density matrix theory. Other important properties are:

- The density operator has unit trace, seeing that

$$\mathrm{Tr}(\hat{\rho}) = \langle \mathbb{1} \rangle = \sum_{k=1}^{N} W_k \langle \psi_k | \psi_k \rangle = \sum_{k=1}^{N} W_k = 1. \tag{2.8}$$

- $\mathrm{Tr}(\hat{\rho}^2) = 1$ iff $\hat{\rho}^2 = \hat{\rho}$ iff $\hat{\rho}$ pure.

- Density matrices with one non-zero diagonal entry ρ_{ii} and zeros elsewhere describe pure ensembles. However, this statement is not reversible (iff) since this matrix is representation-dependent. Purity is only ensured by the previous property.

- D-dimensional density matrices ρ with $\rho_{ii} = 1/D$ for $i \in \{1, 2, \dots, D\}$ and $\rho_{ij} = 0$ for $i \neq j$ describe random ensembles. Diagonality implies representation invariance.

2.2 Entropy

In the Schödinger picture (time-dependent operators) the Liouville equation,

$$i\hbar \frac{d\hat{\rho}(t)}{dt} = [\hat{H}(t), \hat{\rho}(t)], \tag{2.9}$$

describes the equation of motion for density operators. In thermal equilibrium we have $\frac{d\hat{\rho}(t)}{dt} = 0$, so that the Hamiltonian and density operators have a common set of orthonormal eigenfuctions $\{|\eta\rangle\}$ with

$$\hat{H}\,|\eta_k\rangle = E_k\,|\eta_k\rangle \qquad \hat{\rho}\,|\eta_k\rangle = \rho_k\,|\eta_k\rangle\,. \tag{2.10}$$

An additional requirement for thermal equilibrium is maximized entropy. Entropy for density matrices is measured by the *von Neumann entropy* [19],

$$S = -\mathrm{Tr}(\hat{\rho}\ln\hat{\rho}), \tag{2.11}$$

where $0 \leq S \leq \ln(D)$ with the lower bound and upper bounds describing pure and random ensembles respectively, thereby quantifiying uncertainty about the system[1,2].

In the canonical ensemble (constant number of particles, volume and temperature), the equilibrium operator can be found by employing the Lagrange-multiplier method for the optimization problem

$$\max S, \quad \text{while} \quad \mathrm{Tr}(\hat{\rho}) = 1 \quad \text{and} \quad \mathrm{Tr}(\hat{\rho}\hat{H}) = E. \tag{2.12}$$

Using that $S = \sum_k \rho_k \ln\rho_k$ and $\mathrm{Tr}(\hat{\rho}\hat{H}) = \sum_k \rho_k E_k$, the Lagrangian can be written as

$$\begin{aligned}\mathcal{L} &= S - \lambda_1(\mathrm{Tr}\hat{\rho} - 1) - \lambda_2(\mathrm{Tr}(\hat{\rho}\hat{H}) - E) &(2.13)\\ &= -\sum_k \rho_k \ln\rho_k - \lambda_1(\sum_k \rho_k - 1) - \lambda_2(-E + \sum_k \rho_k E_k) &(2.14)\\ &= -\sum_k \rho_k(\ln\rho_k + \lambda_2 E_k + \lambda_1) + \lambda_2 E + \lambda_1. &(2.15)\end{aligned}$$

The gradient w.r.t. the variables at interest is $\nabla\mathcal{L} = (\partial_{\rho_k}, \partial_{\lambda_1}, \partial_{\lambda_2})$. From $\partial_{\rho_k}\mathcal{L}$ it follows that $\rho_k = \exp(-1 - \lambda_1 - \lambda_2 E_k)$ for all k. By substituting this expression for ρ_k in the trace constraint,

$$\mathrm{Tr}(\hat{\rho}) = \exp(-1 - \lambda_1)\sum_i \exp(-\lambda_2 E_i) = 1, \tag{2.16}$$

it follows that the density matrix of the canonical ensemble in thermal equilibrium has a decomposition with spectral coefficients

$$\rho_k = \frac{\exp(-\beta E_k)}{\sum\limits_i \exp(-\beta E_i)}, \tag{2.17}$$

where the Lagrange multiplier λ_2 is a degree of freedom identified as the inverse temperature $\beta \equiv \lambda_2 \equiv 1/(k_b T)$. Consequently, the Boltzmann distribution for quantum systems is denoted

$$\hat{\rho} = \frac{\exp(-\beta\hat{H})}{\mathrm{Tr}(\exp(-\beta\hat{H}))}. \tag{2.18}$$

The denominator in equation 2.18 is called the (canonical) *partition function.*

[1] This will be discussed more subtly in 2.3.2.

[2] The rigorous mathematical foundation of density operator theory can be attributed to John von Neumann. To get an idea of the historical impact of his work on quantum physics, admire the foreword of the 2018 republication of his book: *Mathematical Foundations of Quantum Mechanics* [20]. This book includes a motivation for the von Neumann entropy in section *V*.2.

2.3 Entanglement

Entanglement is a correlation property of quantum mechanics that has no classical equivalent. Utilization of this quantum property in machine learning algorithms is therefore of importance in order to outperform their classical equivalents. More generally, in quantum computing it seems entanglement is the main hope for quantum supremacy. On the other hand, quantum adaptations of classical algorithms must be able to capture entanglement appropriately. As will be discussed, the ability of a neural architecture to express entanglement depends on the entanglement scaling law.

Schrödinger coined the term *entanglement* with the following description [21]:

> "When two systems, of which we know the states by their respective representatives, enter into temporary physical interaction due to known forces between them, and when after a time of mutual influence the systems separate again, then they can no longer be described in the same way as before, viz. by endowing each of them with a representative of its own. I would not call that one but rather the characteristic trait of quantum mechanics, the one that enforces its entire departure from classical lines of thought. By the interaction the two representatives [the quantum states] have become entangled."

Summarizing, a system composed of two sub-systems described by $|\psi\rangle_A$ and $|\psi\rangle_B$, with respective Hilbert spaces $\mathcal{H}_A$, $\mathcal{H}_B$, is entangled if it cannot be represented by a tensor product of wavefunctions, i.e. does not have a *separable* representation

$$|\psi\rangle = |\psi\rangle_A \otimes |\psi\rangle_B\,, \tag{2.19}$$

where $\otimes$ denotes the tensor product. Condidering statistical ensembles, separability requires that the density operator $\hat{\rho}$ of the composite system can be represented as

$$\hat{\rho} = \sum_k W_k\, \hat{\rho}_A^k \otimes \hat{\rho}_B^k. \tag{2.20}$$

The most disconcerting consequence of entanglement is that the state of system A, entangled with system B, may be steered by performing measurements exclusively on system B. This curious and hard-to-interpret quantum characteristic has been a thorn in the sides of many 20th century physicists [22].

Most notably Einstein asserted that quantum mechanics gave a correct, but incomplete description. He disagreed with the view that the physical state of an object could be dependent on the history of measurements performed on said object and argued for a 'detached observer' theory of quantum physics instead, using *hidden variables* to model quantum correlations between systems A and B with local descriptions. His critique was shared by other physicists, resulting in the Einstein-Podolski-Rosen argument against, what was in their view, an incomplete theory. In particular they argued that, like the position and momentum of colliding two billiard balls, classical correlations share a common cause, and the Copenhagen interpretation of quantum physics was wrong not to include this concept. The argument was settled by Bohr in 1935 in favor of the Copenhagen interpretation. Nearly three decades later, the proposed EPR interpretation (and any other hidden variable theory) was definitely proven inconsistent with quantum mechanics by Bell's inequality.

2.3.1 Bell's inequality

Imagine two spin-1/2 particles entangled (e.g. due to a decay process), so that they reside in the (non-separable) singlet configuration

$$|\psi\rangle = \frac{1}{\sqrt{2}}(|\uparrow\downarrow\rangle - |\downarrow\uparrow\rangle). \tag{2.21}$$

The average product of spins, repeatedly measured with two detectors placed in series with orientations **a** and **b** is denoted $P(\mathbf{a}, \mathbf{b})$. This automatically implies $P(\mathbf{a}, \mathbf{a}) = 1$, $P(\mathbf{a}, -\mathbf{a}) = -1$ and for arbitrary orientations $P(\mathbf{a}, \mathbf{b}) = -\mathbf{a} \cdot \mathbf{b}$. All (inherently classical) hidden variable theories result in Bell's inequality [23]:

$$|P(\mathbf{a}, \mathbf{b}) - P(\mathbf{a}, \mathbf{c})| \leq 1 + P(\mathbf{b}, \mathbf{c}), \tag{2.22}$$

following from simple probability theory considerations. A quantum mechanical counter-example violating this inequality for the singlet configuration is quite simple: when $\mathbf{a}, \mathbf{b}$ and $\mathbf{b}, \mathbf{c}$ are at 45° angles, so that $P(\mathbf{a}, \mathbf{b}) = 0$ and $P(\mathbf{a}, \mathbf{c}) = P(\mathbf{b}, \mathbf{c}) = -1/\sqrt{2}$, the inequality of equation 2.22 is inconsistent since $1/\sqrt{2} \not\leq 1 - 1/\sqrt{2}$.

In reference to their violation of the Bell inequality, the four maximally-entangled two-qubit states are named the *Bell states* $\{|\psi_{\text{Bell}}\rangle\}$, denoted (omitting normalization) as

$$\left|\psi_{\text{Bell}}^{+1,\pm}\right\rangle = |\uparrow\rangle_A \otimes |\uparrow\rangle_B \pm |\downarrow\rangle_A \otimes |\downarrow\rangle_B \,, \tag{2.23}$$

$$\left|\psi_{\text{Bell}}^{-1,\pm}\right\rangle = |\uparrow\rangle_A \otimes |\downarrow\rangle_B \pm |\downarrow\rangle_A \otimes |\uparrow\rangle_B \,. \tag{2.24}$$

2.3.2 Entanglement and purity

It should be clarified that entanglement and purity are two separate, co-existent properties of a quantum system. Mixed states are very well capable of entanglement, in which case we speak of *noisy entanglement*. While the Bell-states look harmless, the interpretation of entanglement is still subtle, especially when mixed with other systems. A fundamental understanding of these systems is detrimental for experimental quantum computing, especially with reference to *decoherence*; the mixing of a pure state with other states. Quantifying entanglement and purity is an ongoing subject of research. For example, it has been shown that entanglement and purity dissipate at different rates in a two-qubit system in contact with a heat bath [26].

	entangled	non-entangled
pure	$\lvert\psi_{\text{Bell}}\rangle$	$\lvert\psi\rangle = \lvert\ldots\rangle_A \otimes \lvert\ldots\rangle_B$
mixed	$\hat{\rho}_{\text{Werner}} = p \left\lvert\psi_{\text{Bell}}^{-1,-}\right\rangle \left\langle\psi_{\text{Bell}}^{-1,-}\right\rvert + (1-p)I/4$	$\hat{\rho} = \sum_k W_k \, \hat{\rho}_A^k \otimes \hat{\rho}_B^k$

Table 2.1: Some examples of (non-)separable and pure/mixed two-qubit states. The Werner state is a Bell-state plus white noise model, only non-separable for $p > 1/3$ [24]. Moreover, Werner states can be described by hidden variable models for $p \leq 5/12$ [25].

Tracing out subsystems

The von Neumann entropy plays a special role in quantifying entanglement for bipartite systems (divisible in $\mathcal{H}_A$ and $\mathcal{H}_B$), illuminating the relation between purity and entanglement to some extent. In particular, the (pure) non-degenerate zero-temperature ground state of some quantum lattice Hamiltonian $\rho = |\psi\rangle\langle\psi|$ has von Neumann entropy given by equation 2.11. Considering the state of one part A of a bipartite section of the lattice can be done by *tracing out* the B-part. This gives the *reduced* density matrix $\rho_A = \mathrm{Tr}_B\rho$. For example, the matrix of the $(+1,+)$-Bell state is

$$\begin{aligned}\rho &= \left|\psi_{\text{Bell}}^{+1,+}\right\rangle \left\langle\psi_{\text{Bell}}^{+1,+}\right| && (2.25)\\ &= \frac{1}{2}\Big(|\uparrow\rangle_A \otimes |\uparrow\rangle_B + |\downarrow\rangle_A \otimes |\downarrow\rangle_B \Big)\Big(\langle\uparrow|_A \otimes \langle\uparrow|_B + \langle\downarrow|_A \otimes \langle\downarrow|_B \Big) && (2.26)\\ &= \frac{1}{2}\Big(|\uparrow\rangle\langle\uparrow| \otimes |\uparrow\rangle\langle\uparrow| + |\downarrow\rangle\langle\uparrow| \otimes |\downarrow\rangle|\uparrow\rangle + |\downarrow\rangle\langle\uparrow| \otimes |\downarrow\rangle\langle\uparrow| + |\downarrow\rangle\langle\downarrow| \otimes |\downarrow\rangle\langle\downarrow| \Big), && (2.27)\end{aligned}$$

where the system sub-indices are implied A in front of, and B after the $\otimes$-sign respectively. Taking the trace over sub-system B implies a summation over the (orthonormal) basis of B:

$$\begin{aligned}\rho_A &= \sum_{|\psi\rangle=\{|\uparrow\rangle,|\downarrow\rangle\}} \Big(\mathcal{I} \otimes \langle\psi|\Big)\rho\Big(\mathcal{I} \otimes |\psi\rangle\Big) && (2.28)\\ &= \frac{1}{2}\Big(|\uparrow\rangle\langle\uparrow| + |\downarrow\rangle\langle\downarrow| \Big). && (2.29)\end{aligned}$$

This density matrix describes a fully random ensemble, therefore having non-zero entropy. Saliently, the uncertainty quantified by the von Neumann entropy is purely quantum in this case: not due to classical uncertainty, but strictly due to entanglement [27].

2.3.3 Entanglement scaling law

Having established entanglement as a non-negotiable property of quantum theory, the next logical question is: “How entangled is my system?”. Although violation of the Bell inequality means that a system is entangled the converse is not true [24], rendering the ”degree of Bell-violation” useless as a faithful measure. The subject of (measuring) entanglement is still an area of active research, most notably summarized in [28]. Instead of going over different measures of entanglement, we will now look at a more relevant aspect of entanglement with regards to many-body quantum Hamiltonians and machine learning, following Ref. [27]. The reason for this is that besides playing the “game” of quantifying the entanglement of quantum systems with a single number, more general statements can be made about the *scaling* of entanglement. This general classification into *kinds of* entanglement will prove to be of importance since it can be shown that some machine learning architectures can only model certain kinds of entanglement.

While it seems natural that entanglement scales with the size of a bipartite section A of the lattice, many ground states of are said to adhere to an *area law* of entanglement: the von Neumann entropy scales linearly with the boundary size of A. If the entropy does scale with the size of A, as might seem more natural, the state adheres to a *volume law*, which is typical for systems that do not consider only nearest-neighor interactions. Satisfying the area law is crucial for numerical (ground-state) algorithms like *DMRG*, but also of importance to machine learning algorithms such as the neural-net wavefunction (chapter 6).

The area law is closely related to the concept of *quantum criticality*, which on its turn manifests itself in quantum spin lattices in qualitative changes in spin-correlations. For zero-temperature systems, the spin-correlations can be explained solely due to entanglement, hence entanglement undergoes a transition as well at these critical points. In fact, at critical points the entanglement entropy scales logarithmically with volume [29]. Only gapped, local, non-critical Hamiltonians obey the area law of entanglement.

Chapter 3

Spin Lattice Models

In this chapter, a few example quantum spin-1/2 lattice models are exhibited along with their analytical difficulties. The study of the magnetic and electric properties of magnetically ordered solids has been greatly advanced by these models. In particular they add to the understanding of high-T_c superconducting copper-oxides [30]. The lattices described by these models consist of atoms paired with localized electrons, interacting solely due to their spins through the exchange interaction; an effective interaction due to Coulumb repulsion and the Pauli exclusion principle. Most models studied in literature include only nearest-neighbor interactions, most famously the Heisenberg model and the Transverse Field Ising Model. The anti-ferromagnetic Heisenberg model originates from the most basic many-body particle model: the Hubbard model.

3.1 Heisenberg models

Ferromagnetic interactions arise from spatially overlapping wavefunctions, for which aligned spins reduce the Coulomb repulsion. Anti-ferromagnetic interactions arise from spatially separated wavefunctions where anti-alignment reduces the kinetic energy of the electrons [31]. Anti-ferromagnetism is characterized by a positive exchange interaction. The quantum anti-ferromagnetic Heisenberg (AFH) model is denoted in its most general form as a sum over two-local interaction terms in all directions,

$$H = \sum_{ij} w^x_{ij}\sigma^x_i\sigma^x_j + w^y_{ij}\sigma^y_i\sigma^y_j + w^z_{ij}\sigma^z_i\sigma^z_j, \tag{3.1}$$

where σ^k_i denotes the Pauli spin operator on site i and direction k and the sum is over nearest neighboring spin pairs ij. The Pauli operators act locally and are formed by tensor products, making the Hamiltonian H exponential in the system size. The Pauli spin operators -which form a basis for the space of 2 × 2 Hermitian matrices (together with the identity matrix)- are

$$\sigma^x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \sigma^y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \qquad \sigma^z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{3.2}$$

Moreover, they are unitary with eigenvalues ±1. The σ^z operator has corresponding eigenvectors

$$\psi_{+1} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \psi_{-1} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{3.3}$$

The many-body state of a system is conventionally represented in this basis, also known as the *computational basis*. Consequently one can represent such configurations in ket-notation as $|s\rangle = |s_1, s_2, \ldots\rangle$, where s_i denotes the eigenvalue of σ_i^z.

The interactions w are often taken direction dependent but lattice site independent ($w_{ij}^x = w^x$, $w_{ij}^y = w^y$ and $w_{ij}^z = w^z$). In case the couplings are axis-independent, equation 3.1 is often rewritten as

$$\begin{aligned} H &= \sum_{ij} w_{ij} \vec{\sigma}_{ij} && (3.4) \\ &= J \sum_i \vec{\sigma}_i \vec{\sigma}_{i+1} + \alpha J \sum_i \vec{\sigma}_i \vec{\sigma}_{i+2}, && (3.5) \end{aligned}$$

where the last equality holds for the next-nearest neighbour model.

The AFH model with fully isotropic positive interactions ($w_{ij}^x = w_{ij}^y = w_{ij}^z > 0$) is also referred to as the XXX AFH model. In addition, one could consider XXZ ($w_{ij}^x = w_{ij}^y > 0, w_{ij}^z > 0$) and XYZ ($w_{ij}^x > 0, w_{ij}^y > 0, w_{ij}^z > 0$) AFH models. Another common assumption is periodicity, implying that the lattice forms a ring by connection of the boundary spins[1]

As an example, the two qubit XXX Hamiltonian with $w = 1$ is given by

$$H = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 2 & 0 \\ 0 & 2 & -1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \tag{3.6}$$

with corresponding ground state eigenvector $\psi_0 = (1/\sqrt{2}) \cdot (0, 1, -1, 0)^{\mathrm{T}}$: the singlet state.

Elements $\langle s|H|s'\rangle$ of the Hamiltonian H of equation 3.1 can be evaluated locally by

$$\langle s|H|s'\rangle = \sum_{ij,i<j} (w_{ij}^x - w_{ij}^y s_i s_j) \langle s|F_i F_j s'\rangle + \sum_{ij,i<j} w_{ij}^z s_i s_j \langle s|s'\rangle , \tag{3.7}$$

where $F_i |s\rangle = F_i |s_1, s_2, \ldots, s_i, \ldots\rangle = |s_1, s_2, \ldots, -s_i, \ldots\rangle$, so that $\langle s|F_i s'\rangle$ denotes a Kronecker delta due to the orthonormality of the computational basis.

3.1.1 Marshall-Peierls Theorem

The Marshall-Peierls Theorem reveals that the ground state $|\psi_0\rangle$ of a spin-1/2 XXZ AFH Hamiltonian on an arbitrarily sized bipartite[2] graph can be unitarily transformed to determine its sign structure [32]. In addition,

$$\langle s|\psi_0\rangle = 0 \qquad \langle s| \notin \mathcal{S}, \tag{3.8}$$

where $\mathcal{S} = \{\langle s| \mid \sum_i s_i = 0\}$. The unitary transformation can also be interpreted as a transformation of the Hamiltonian around the z-axis for one of the sublattices, so

[1] Note that these nearest-neighbor models are effectively 1D chains.

[2] The concept of bipartiteness is also of importance in the context of Restricted Boltzmann Machines (section 5.2.1). A lattice is bipartite if it can be divided into two subsets A and B, with connections only between A and B but not between two members of the same subset. In particular, nearest neighbor lattices are bipartite.

that $\sigma^x \to -\sigma^x$ and $\sigma^y \to -\sigma^y$, effectively introducing a minus-sign in $\sigma_i^x \sigma_j^x$ and $\sigma_i^y \sigma_j^y$ terms for nearest-neighbor Hamiltonians. For the non-transformed ground state (with sign structure) $\left|\psi_0^{\pm}\right\rangle$, this implies

$$\left\langle s\middle|\psi_0^{\pm}\right\rangle = (-1)^{N_\uparrow(s)} \left\langle s\middle|\psi_0^{+}\right\rangle , \tag{3.9}$$

where $N_\uparrow(s)$ counts the number of spins with eigenvalue $+1$ on one of the sublattices and $\| \left\langle s\middle|\psi_0^{+}\right\rangle \| = \| \left\langle s\middle|\psi_0^{\pm}\right\rangle \| \quad \forall \left\langle s\right| \in \mathcal{S}$.

Element-wise positivity of the transformed ground state wavefunction can also be derived directly from the action of the Marshall transformation, which ensures all off-diagonal elements of the XXZ Hamiltonian are negative in the σ^z-basis. Consequently, $-H$ has all off-diagonal elements positive, so that the largest positive eigenvalue E_{extreme} of $-H$ is equal to the spectral radius $\rho(-H) = E_{\text{extreme}}$, corresponding to an eigenvector ψ_{extreme} which is element-wise positive by Perron-Frobenius theorem [33]. For $+H$ the eigenvalue spectrum is inverted, so that we denote $-E_{\text{extreme}} = E_0$ with $\rho(H) = -E_0$ and $\psi_{\text{extreme}} = \psi_0$. The off-diagonal negativity property (by physicists known as *stoquasticity*), and consequently the element-wise positivity of ψ_0, will be of importance for numerical ground state methods (section 6.2.2).

3.1.2 Solving the Heisenberg model

Since the size of the Hamiltonian scales exponentially in the system size, exact diagonalization is possible for up to only $\approx$ 30 spins, depending on the available hardware. For highly symmetric models, this limit has been stretched up to fifty spins [11]. Diagonalization can be done exactly through Gaussian elimination in $\mathcal{O}(n^3)$ time or by usage of more efficient iterative methods such as the Lanczos and Arnoldi algorithms. In order to reduce the amount of required memory it is possible to use sparse or even functional (matrix-free) representations, with the setback that this might slow down required matrix operations.

The XYZ AFH models constitute a special class of Hamiltonians solvable by iterative methods different from conventional diagonalization[3]. Two notable methods are *DMRG* [34] and the *Bethe Ansatz*[35][36]. These methods provide benchmark results to test the accuracy of other numerical solvers.

A general method to solve the ground state of quantum spin Hamiltonians is explored in section 6.

[3] More precisely, these Hamiltonians are *integrable.* Integrability in the classical sense refers to the ability to describe dynamics of a system. The three body problem is a classical example of a non-integrable problem.

Chapter 4

Markov Chain Monte Carlo

Monte Carlo methods use stochasticity in order to estimate expectation values w.r.t some probability distribution. While officially developed alongside with the atomic bomb at the *Los Alamos National Laboratory* by Ulam, von Neumann and Metropolis in 1949 [37], the earliest usage of Monte Carlo is accredited to Fermi [38]. From its inception, its applications have spread from neutron fission experiments and nuclear cascades to stock market predictions and artificial intelligence. The theoretical development of the method was accompanied by the experimental development of the first electrical computers: the *ENIAC*, *FERMIAC* and *MANIAC*. Its widespread applications nowadays would have been unthinkable without the tremendous progress made in computer engineering, moving from mechanical calculators and *punching cards* for random numbers to logical units on a nanometer scale and being able to generate millions of random numbers per second with consumer electronics.

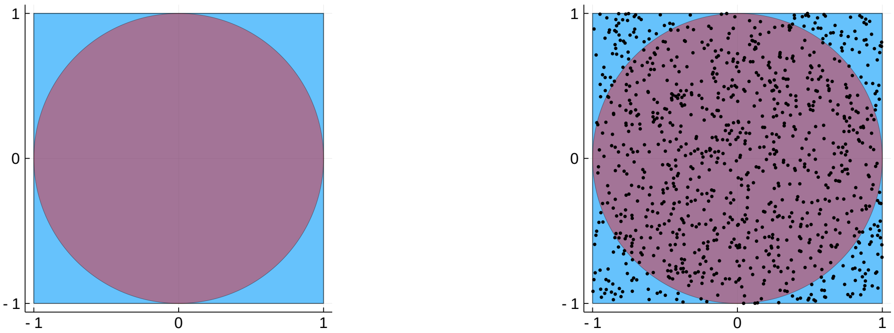
An illustrative example to get acquainted with the concepts of Monte Carlo is the estimation of π. A value proportional to π can be written as the relative surface area of a unit circle inscribed in a unit square (see figure 4.1):

$$\frac{O(\text{circle})}{O(\text{square})} = \frac{\pi r^2}{l \cdot w} = \frac{\pi}{4}. \tag{4.1}$$

Imagine now that a dart is thrown at this figure. The dart will either land in the large zone encompassed by the circumference of the circle, or in one of the four zones encompassed by parts of the circumference of the circle and the square (see figure 4.1). Assuming the dart is thrown without any bias towards certain areas, the probability for the dart to land in the circle is precisely the ratio of the surface areas in equation 4.1. Throwing a dart once gives only limited information: it either landed within the circle or not. However, throwing more darts (or the same dart multiple times) gives a direct estimate of π by counting the number of darts within the circle and dividing it by the total amount of darts thrown. Throwing more darts decreases the variance of our estimate of π according to the law of large numbers. The decrease in the relative error w.r.t. the amount of independent darts thrown is illustrated empirically in table 4.1.

4.1 Introduction to Monte Carlo methods

Monte Carlo (MC) methods are used to generate samples and/or calculate expectation values from probability distributions in cases where the complete probability



(a) A unit circle inscribed in a unit square. The ratio of their relative surface areas is proportional to π.

(b) 1000 Random darts dropped on the surface. From this distribution of darts, $\pi_{\text{estim}} = 3.164$.

Figure 4.1: Monte Carlo estimation demonstrated.

darts $(10^{\cdots})$	π estimate	rel. error
1	3.2	$1.850 \cdot 10^{-2}$
2	3.32	$5.679 \cdot 10^{-2}$
3	3.164	$7.132 \cdot 10^{-3}$
4	3.1504	$2.803 \cdot 10^{-3}$
10	3.14152004	$2.311 \cdot 10^{-6}$

Table 4.1: The Monte Carlo error of the estimated π decreases on average with the amount of darts thrown.

distribution is not known [39]. Samples are single instances $\mathbf{x}$ from the probability distribution $\gamma(\mathbf{x})$. That is to say, taking many samples from the distribution and tallying them in a histogram will resemble $\gamma(\mathbf{x})$. The samples come from a state space $\mathbf{x} \in \mathcal{X}$ which is to be further specified later. MC is particularly applicable in the case of unnormalized probability distributions, which arise naturally from thermodynamics in the form of Boltzmann distributions with intractable normalization constants. Intractable normalization constants are also ubiquitous in Bayesian statistics.

However, even sampling from normalized distributions could still be problematic, since there is not one straightforward way to collect the representative set, without evaluating $\gamma(\mathbf{x})$ everywhere. In order to properly account for the relevant probability masses of a complex distribution, one needs smart strategies to traverse the probability space. One such strategy is discussed in section 4.2.

An important note should be made about the dimensionality scaling of the Monte Carlo method. The expectation value Φ of a function $\phi(\mathbf{x})$ is defined as

$$\Phi \equiv \langle \phi(\mathbf{x}) \rangle \equiv \int d\mathbf{x} \gamma(\mathbf{x}) \phi(\mathbf{x}). \tag{4.2}$$

With a set of N samples, $\{\mathbf{x}^i\}_{i=1}^N$ sampled from $\gamma(\mathbf{x})$, the exact expectation value is approximated by the estimator[1]

$$\hat{\Phi} \equiv \frac{1}{N} \sum_i \phi(\mathbf{x}^i). \tag{4.3}$$

[1] In other words: the *population* mean is replaced by the *sample* mean.

With an increasing amount of samples, the variance of the estimator decreases. This follows simply from the fact that if i.i.d random variables $\mathbf{x}$ are distributed with variance σ^2, then $\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}^i$ has variance σ^2/N. In particular, the variance is independent of the dimensionality of $\mathbf{x}$.

4.1.1 Back to the dartboard

The most naive way to traverse the entire state space $\mathcal{X}$ is by drawing random samples with uniform probability (and approximating the normalization if needed empirically by summing over all encountered probabilities). This is effectively what was done for the estimation of π at the beginning of this chapter. Samples from the square were collected directly with uniform probability and the amount of 'hits' was kept track off. More formally, π was treated as the expectation value of an indicator function w.r.t. γ:

$$\pi \propto \int\limits_{\text{square}} dxdy\phi(x,y)\gamma(x,y), \tag{4.4}$$

with

$$\gamma(x,y) = \frac{1}{\int_{\text{square}} dxdy} \tag{4.5}$$

and

$$\phi(x,y) = \begin{cases} 1 & \text{if } (x,y) \text{ in the circle,} \\ 0 & \text{if } (x,y) \text{ outside the circle.} \end{cases}$$

This integral is replaced by the estimator

$$\hat{\pi} \propto \frac{1}{N}\sum_{i=1}^{N}\phi(x_i,y_i) \tag{4.6}$$

This approach is less effective for distributions $\phi(\mathbf{x})$ with very unevenly divided probability mass, as many uniformly drawn points will have negligible contribution to the integral[2]. While this is not prohibitively problematic in the dartboard example, this typically becomes a problem with high-dimensional probability distributions.

More advanced sampling strategies are needed to account for more complex, scattered probability masses. Two textbook sampling strategies that use a second, known probability distribution $Q(\mathbf{x})$ to sample from the target $\gamma(\mathbf{x})$ are *importance* and *rejection sampling*. Whilst insightful, these methods are not of practical use in this thesis since they perform poorly in high-dimensional problems and rely somewhat heavily on prior knowledge of the target distribution.

4.2 Metropolis-Hastings sampling

A robust method suitable for high-dimensional distributions is the Metropolis-Hastings (MH) method. Instead of using a static proposal probability distribution, MH uses

[2] For example, if we lived in a universe where the unit circle was replaced by a shape with much smaller relative surface area.

a state-dependent proposal distribution $Q(\mathbf{x}'|\mathbf{x})$ to collect samples in a sequential fashion. The proposal state $\mathbf{x}'$ is accepted as the new state with probability

$$A(\mathbf{x}'|\mathbf{x}) = \min\left\{\frac{\gamma(\mathbf{x}')}{\gamma(\mathbf{x})}\frac{Q(\mathbf{x}|\mathbf{x}')}{Q(\mathbf{x}'|\mathbf{x})}\right\}. \tag{4.7}$$

Note that upon rejection of the new sample $\mathbf{x}'$, the old sample is not discarded; rather, it is repeated in the sequence. More importantly, normalization constants disappear in the ratio $\gamma(\mathbf{x}')/\gamma(\mathbf{x})$, making the method suitable for sampling from unnormalized distributions. The proposal distribution could also be taken to be symmetric, reducing the acceptance probability to

$$A(\mathbf{x}'|\mathbf{x}) = \frac{\gamma(\mathbf{x}')}{\gamma(\mathbf{x})}. \tag{4.8}$$

Since the acceptance probability of a proposed state depends only on the current state, the sequential gathering of samples is a *Markovian* process, meaning that the random variable $\mathbf{x}^t$ does not depend explicitly on any previous sample $\mathbf{x}^{t-j}$, j steps back in the sequence [40]:

$$P(\mathbf{x}^{t+1}|\mathbf{x}^t, \mathbf{x}^{t-1}, \mathbf{x}^{t-2}, \ldots) = P(\mathbf{x}^{t+1}|\mathbf{x}^t), \tag{4.9}$$

making the sequence memoryless.

4.3 Markov chains

The sequence of samples generated by functions with the Markov property are called *Markov chains*. The goal with these kinds of experiments, known under the collective name *Markov Chain Monte Carlo* (MCMC), is to generate samples from the target distribution. Rather intuitively, the aforementioned MH acceptance rate biases the chain towards regions of the distribution with higher probability mass. However, the success of the rather arbitrary looking MH algorithm should be demystified through more careful investigation.

The first mystery is the unavoidable initialization effect. Upon initialization of the chain, the first sample is strictly sampled from the initialization distribution instead of the target distribution. How many steps are needed before samples can be said to be from the target distribution? What is to say the MH ratio makes the chain converge to the target distribution at all?

Secondly, the proposal distribution Q has not been specified. Remarkably, the MH method leaves the choice of Q up to the imagination of the user. What makes the MH acceptance ratio so special in order for it to be so generally applicable?

Fortunately, the distribution of samples generated with the MH acceptance probability of equation 4.7 can be proven to converge to the target distribution under certain conditions. If these conditions are met, the target distribution is called the *stationary distribution* of the Markov chain and the chain will converge to the stationary distribution starting from any initialization. The convergence conditions are described in the following subsections.

4.3.1 Detailed Balance

The conditional distribution $P(\mathbf{x}^{t+1}|\mathbf{x}^t)$, is called the *transition kernel.* The transition kernel for the MH sampling method is [40]:

$$\begin{aligned} P(\mathbf{x}^{t+1}|\mathbf{x}^t) = &\ P(\mathbf{x}^{t+1} \neq \mathbf{x}^t|\mathbf{x}^t) &+&\ P(\mathbf{x}^{t+1} = \mathbf{x}^t|\mathbf{x}^t) \qquad (4.10)\\ = &\ Q(\mathbf{x}^{t+1}|\mathbf{x}^t)A(\mathbf{x}^{t+1}|\mathbf{x}^t) &+&\ I(\mathbf{x}^{t+1} = \mathbf{x}^t)\Big[1 - \int d\mathbf{y}Q(\mathbf{y}|\mathbf{x}^t)A(\mathbf{y}|\mathbf{x}^t)\Big]. \qquad (4.11)\end{aligned}$$

The first term in this expression describes the acceptance of the candidate state $\mathbf{x}'$, the second term describes its rejection, where I denotes the indicator function. Note that the first term is only defined for $\mathbf{x}^{t+1} \neq \mathbf{x}^t$.

By multiplying either side by $\gamma(\mathbf{x}^t)Q(\mathbf{x}^{t+1}|\mathbf{x}^t)$, one gets

$$\begin{aligned} \gamma(\mathbf{x}^t)Q(\mathbf{x}^{t+1}|\mathbf{x}^t)A(\mathbf{x}^{t+1}|\mathbf{x}^t) &= \min\Big\{\gamma(\mathbf{x}^t)Q(\mathbf{x}^{t+1}|\mathbf{x}^t), \gamma(\mathbf{x}^{t+1})Q(\mathbf{x}^t|\mathbf{x}^{t+1})\Big\} \qquad (4.12)\\ &= \gamma(\mathbf{x}^{t+1})Q(\mathbf{x}^t|\mathbf{x}^{t+1})A(\mathbf{x}^{t+1}|\mathbf{x}^t), \qquad (4.13)\end{aligned}$$

obtaining the *detailed balance* or *time reversal* equation

$$\gamma(\mathbf{x}^t)P(\mathbf{x}^{t+1} \neq \mathbf{x}^t|\mathbf{x}^t) = \gamma(\mathbf{x}^{t+1})P(\mathbf{x}^t \neq \mathbf{x}^{t+1}|\mathbf{x}^{t+1}). \qquad (4.14)$$

Through straightforward evaluation it follows that the detailed balance property also holds for the rejection term.

The link between detailed balance and the stationarity of γ can be demonstrated by integrating the complete detailed balance equation w.r.t. $\mathbf{x}^t$

$$\int \gamma(\mathbf{x}^t)P(\mathbf{x}^{t+1}|\mathbf{x}^t)d\mathbf{x}^t = \gamma(\mathbf{x}^{t+1}), \qquad (4.15)$$

proving that γ is the invariant distribution w.r.t the transition kernel P; in other words, γ is invariant under P in the sense that[3] $\gamma P = \gamma$. Although the MH transition kernel satisifies detailed balance, transition kernels could theoretically be simultaneously invariant and irreversible; detailed balance is a *sufficient* rather than a *necessary* condition for invariance.

Markov Master equation

Perhaps a more insightful way of interpreting detailed balance is by examination of the Markov master equation. This equation considers the flow of probability in a continuous-time picture rather than discrete, denoting $\gamma(\mathbf{x}, t)$ as the probability of γ to take on the value $\mathbf{x}$ at time t.

This interpretation allows only discrete state spaces $\mathcal{X}$. In line with the distributions of interest in this thesis, we could consider $\mathbf{x}_i$ corresponding to a (discrete) configuration of spins on a lattice. The continuous-time approach allows one to write the probability flow in the differential form [41]

$$\frac{d\gamma(\mathbf{x}_j, t)}{dt} = -\sum_i W_{ji}\gamma(\mathbf{x}_j, t) + \sum_i W_{ij}\gamma(\mathbf{x}_i, t), \qquad (4.16)$$

[3] P is defined right-invariant by convention. Alternatively, P is left-invariant if $P\gamma = \gamma$.

where W_{ij} denotes the conditional probability $P(\mathbf{x}^{t+1} = \mathbf{x}_j|\mathbf{x}^t = \mathbf{x}_i)$. The first term of the r.h.s of equation 4.16 can be read as the probability to move away from the configuration $\mathbf{x}_j$, while the second term describes the probability to move towards $\mathbf{x}_j$.

Substituting detailed balance in this equation results in a zero net-flow of probability, $d\gamma(\mathbf{x}_j, t)/dt = 0$, elucidating the notion of γ as the *stationary* distribution further.

4.3.2 Convergence

Having shown that Q can in fact be chosen freely (limited by regularity conditions [40]), the matter of initialization is still an open question. Equation 4.15 ensures that $\mathbf{x}^{t+1}$ is a sample from γ if $\mathbf{x}^t$ is a sample from γ. However, since the chain starts with a sample $\mathbf{x}^0$ from an initialization distribution (e.g. uniform, gaussian), which is variant[4] under P, the invariance property is not sufficient to ensure convergence.

In the treatment of the following material, a discrete state space is again assumed for the sake of practical relevance. For discrete state spaces, the transition kernel can be represented by a transition matrix, which is stochastic s.t.

$$\sum_j P_{ij} = \sum_j P(\mathbf{x}^{t+1} = \mathbf{x}_i|\mathbf{x}^t = \mathbf{x}_j) = 1. \tag{4.17}$$

Irreducibility and aperiodicity

In order for the initial distribution to converge to the target distribution, the chain is required to be *irreducible*. A chain is irreducible if any part of state space has a non-zero probability to be encountered within a finite amount of steps, starting from any other part of state space: $\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}\ \exists t\colon P^t(\mathbf{x}|\mathbf{y}) > 0$. This condition can be violated when some states are absorbing w.r.t. the transition kernel. States are considered absorbing if they are encountered in the chain with finite probability, but can not be escaped with finite probability. Additionally, one may prove that irreducibility implies uniqueness of the stationary distribution [42].

A requirement complementary to irreducibility is *aperiodicity*. The period of a state $\mathbf{x}$ is the greated common divisior of the set $\mathcal{T}(\mathbf{x}) \equiv \{t \geq 1\colon P^t(\mathbf{x}|\mathbf{x}) > 0\}$. An aperiodic chain will have period 1 for all states. This disallows the state space to be divided into, for example, an even and an odd class such that only transitions in the chain between the two classes are possible. Note that such a chain could still be irreducible, even though it is periodic.

Convergence theorem and ergodicity

The total variation distance between two probability distributions γ and ξ is defined as

$$||\gamma - \xi||_{\mathrm{TV}} = \max_{A \subseteq \mathcal{X}} |\gamma(A) - \xi(A)|, \tag{4.18}$$

which can be interpreted as the set of samples generating the largest possible distance between the two distributions. This measure is used in the *Markov Chain*

[4] Knowing the invariant distribution in advance would inherently solve the problem.

Convergence Theorem: If P is irreducible and aperiodic, with invariant distribution γ, then[5]

$$\max_{\mathbf{x}\in\mathcal{X}} ||P^t(\cdot|\mathbf{x}) - \gamma||_{\mathrm{TV}} \leq C\alpha^t, \tag{4.19}$$

for $0 \leq \alpha \leq 1$ and $C > 0$. While theoretically detrimental for the merit of MCMC, the convergence theorem is of little practical use. Even though a large amount of literature is published on the subject of convergence, the determined limits are necessarily loose and inevitably dependent on the specific sampling method and probability-class [43]. Under the motto 'better some than none', heuristics like autocorrelation time (time for a chain to "forget" where it started) could be employed.

According to the *ergodic theorem* irreducible, aperiodic chains that have (almost) converged can be used to calculate expectation values w.r.t. the target distribution. Ergodicity is a omnipresent idea in physics and mathematics, but its exact definition varies subtly between the fields. In the widest sense it means that "time averages equal space averages"- an idea readily applicable to MCMC. The ergodic theorem states that for any starting distribution μ and real-valued function f,

$$\mathbb{P}\Big\{ \lim_{T\to\infty} \sum_{t=0}^{T-1} f(\mathbf{x}^t) = \mathbb{E}_\gamma(f) \Big\} = 1, \tag{4.20}$$

where γ again is the stationary distribution of the irreducible chain that generates samples $\mathbf{x}^t$. The proof of this theorem relies on the Strong Law of Large Numbers.

[5] $P^t(\mathbf{y}|\mathbf{x})$ denotes the probability to move from $\mathbf{x}$ to $\mathbf{y}$ in t steps.

Chapter 5

Quantum Boltzmann Machines

Having defined the mathematical building blocks and physically relevant models, we now turn to the actual quantum learning algorithm. This is preceded by a historical contextualization of the emergence of this quantum algorithm.

Machine Learning (ML) can be defined as ”a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty” [44]. ML is used in particular for data compression, filtering, dimensionality reduction, regression, feature extraction and classification. Machine learning is divided into reinforcement learning (RL), supervised learning (SL) and unsupervised learning (UL). SL learning methods rely on labelled data in order to find patterns in data, whereas UL methods aim to find these patterns autonomously. In contrast to both these methods, RL relies on the autonomous exploration of a parameter space while given excitatory and inhibitory stimuli, similar to how an infant learns to understand its environment. A thorough overview of the historically relevant papers for the development of Deep Learning is given in [45].

Notable ML methods are Bayesian Networks, (Deep) Feedforward Neural Networks (FFNs) and Recurrent Neural Networks (RNNs). The neural network lends its name to the rudimentary similarities to neuronal structures in the brain. The basic ingredients for every neural net are vertices (neurons) and edges of varying strength represented by weights. While FFN architectures edges and vertices do not form a cycle, RNNs do. This makes their respective learning procedures and interpretations of in- and output quite different. Original models of neural models date back to the 1940’s [46] and were soon repurposed for learning algorithms [47]. The imperative back-propagation method for optimizing weights of neural nets of arbitrary depth was developed in the 60s and 70s and popularized in the 90s.

Graphs with vertices and edges remind natural science researchers of crystal lattices and spin models. Conversely, models from physics remind mathematicians of graph theory and neuroscientists of brain structures; ML acts as a vehicle for development of research in a plethora of scientific branches. The mutualism is particularly apparent in combination with physics, the Boltzmann machine being only one example of this. The Boltzmann Machine (BM) was developed around 1985 [48], inspired by ideas from computational biology and psychology. The predecessor of the BM is the Hopfield network, an early RNN with associative memory capabilities. A 2002 publication by Hinton and Salakhutdinov [49] has popularized the BM significantly with the development of the Contrastive Divergence learning method for a specific BM architecture: the Restricted Boltzmann Machine (RBM). Interest in BMs has persisted in the 21st century, in part due to recent developments (and possible applications) in quantum computing.

This chapter will first discuss the moving parts of the BM and their interpretations. From section 5.3 onward, the discussion will focus on the QBM.

5.1 Hopfield networks

Boltzmann machines are stochastic Hopfield networks. Hopfield networks consist of a set of fully connected binary neurons, which turn on (x=1) and off (x=-1) based on the total input a from other neurons. The connection matrix W is symmetric ($w_{ij} = w_{ji}$) and has zeros on the diagonal. The activity rule of a neuron is based on a threshold function

$$x(a) = \Theta(a) \equiv \begin{cases} 1 & a \geq 0 \\ -1 & a < 0. \end{cases} \tag{5.1}$$

The state of each neuron may be updated in a particular order (asynchronously) or they may be updated all at once (synchronously). In both cases, the updated state of a neuron is determined as

$$x_i = \Theta(a_i) = \Theta\Big(\sum_j w_{ij} x_j\Big). \tag{5.2}$$

In the continuous case, the threshold function is replaced by an activation function that smoothly transforms the input a to the $[-1, 1]$ interval by $x_i = \tanh a_i$.

Hopfield networks can be used for pattern completion and (combinatorial) optimization problems. It can be shown that the asynchronous update rule for continuous Hopfield nets minimizes an energy function $E(x) = -\frac{1}{2}\mathbf{x}^T W \mathbf{x}$. The physical terminology hints to its relation to a (classical) spin system with energy

$$E(x) = -\frac{1}{2}\sum_{i,j} J_{ij} x_i x_j, \tag{5.3}$$

where J_{ij} denotes a coupling strength. Minimizing the free energy of this system in the mean-field approximation yields the same asynchronous update scheme with aforementioned input and activation functions [39].

The capability of the Hopfield network to handle pattern completion comes from its associative learning rule, which increases the connection between neurons based on their correlation. The weight matrix W of the network is initialized to the average correlations in some set of *memories* $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$. This set of weights creates energy minima for the memories, so that a noisy copy of a memory is likely to converge to the original upon updating the neurons.

5.2 Boltzmann machines

A powerful ML method especially relevant for this thesis is the Boltzmann Machine (BM). The BM is often employed in the physical sciences for *inverse statistical problems* [50]: "problems where the statistical bulk features of a physical system are known, but where the microscopic interactions are to be learned".

Boltzmann machines are derived from Hopfield nets, but use probabilistic update rules that allow energetically unfavourable transitions rather than deterministic activation functions. This makes it possible to escape from local minima in the process of associative pattern completion, as described in the previous section. The energy change when neuron x_i is flipped can be denoted as

$$\Delta E_i = \sum_j w_{ij} x_j. \tag{5.4}$$

Reminiscent of the Metropolis-Hastings sampling method, the Boltzmann machine flips neuron x_i with probability

$$p_i = \frac{1}{1 + e^{-\Delta E_i / T}}, \tag{5.5}$$

where T is a temperature-like parameter. The higher the temperature, the more uniform this probability becomes, the more likely it is energetically unfavorable transitions are made. The BM is inextricably linked to statistical physics, in particular since equation 5.5 also describes the occupation probability in a two-level system. From statistical physics it is known that a collection of such two-level particles in thermal equilibrium with a heat bath will be distributed according to the Boltzmann distribution

$$P(\mathbf{x}) = Z^{-1} e^{-E(\mathbf{x})/T}, \qquad\qquad Z = \sum_{\mathbf{x}} e^{-E(\mathbf{x})/T}, \tag{5.6}$$

such that the relative frequency of a configuration $\mathbf{x}^A$ w.r.t $\mathbf{x}^B$ is

$$\frac{P(\mathbf{x}^A)}{P(\mathbf{x}^B)} = e^{-[E(\mathbf{x}^A) - E(\mathbf{x}^B)]/T}. \tag{5.7}$$

As previously discussed in the context of Markov chains (chapter 4), the BM is in thermal equilibrium when the distribution of states it generates is time-invariant.

This direct relation between occupation probability and (local) energy terms is the basis for the BM learning algorithm. Training is performed by performing gradient descent on the discrepancy between the probability distributions of the BM and the data, respectively denoted by P and Q. The discrepancy is measured by the Kullback-Leibler divergence

$$K(Q||P) = \sum_j Q(\mathbf{x}_j) \ln \frac{Q(\mathbf{x}_j)}{P(\mathbf{x}_j)}. \tag{5.8}$$

The model-dependent part of the KL divergence can be split off

$$K(Q||P) = \sum_j Q(\mathbf{x}_j) \ln Q(\mathbf{x}_j) + \mathcal{L}, \tag{5.9}$$

where

$$\mathcal{L} \equiv -\sum_j Q(\mathbf{x}_j) \ln P(\mathbf{x}_j) \tag{5.10}$$

denotes the average negative log-likelihood. Minimizing the divergence by shifting model parameters thus boils down to minimizing the negative log-likelihood.

The gradient w.r.t the weights can be derived with the chain rule to be

$$\frac{\partial K}{\partial w_{ij}} = -\frac{1}{T}[\langle \mathbf{x}_i \mathbf{x}_j \rangle_Q - \langle \mathbf{x}_i \mathbf{x}_j \rangle_P], \tag{5.11}$$

where $\langle \ldots \rangle_D$ denotes the expectation value of the argument w.r.t. some thermally equilibrated distribution D. This produces the simple learning rule

$$\Delta w_{ij} = \epsilon[\langle \mathbf{x}_i \mathbf{x}_j \rangle_Q - \langle \mathbf{x}_i \mathbf{x}_j \rangle_P], \tag{5.12}$$

with learning parameter ϵ. Its simplicity is a result of the log probability being linear in the energy and the energy being linear in the parameters. Minimizing the discrepancy between the data distribution and the BM distribution can also be regarded as maximizing the likelihood that the BM generated the data distribution. Without hidden units, the KL divergence is convex so that the gradient descent method is not impaired by local minima.

The Boltzmann Machine could be extended with hidden neurons, allowing more complex patterns to be learned. Consider for example the problem of learning patterns which are constrained by parity ($\prod_i \mathbf{x}_i = 1$). With 3 neurons this constraint cannot be satisfied without somehow capturing the complete state product, needing more than just second order neuron correlations. To this end, the values of the hidden units are entirely determined by the data, and are not changed during sampling of the BM distribution.

Additionally, the BM could be defined to include higher order energy terms [51], but the additional cost of learning this model rarely overcomes the merit of increased expressive power.

The generated model is an approximation of the target distribution. The network consists of $(N_v + N_h - 1)(N_v + N_h)/2 + N_v + N_h$ parameters (quadratic in the number visible/hidden neurons), while the target distribution is inherently 2^{N_v}-dimensional. The quality of the model depends on its ability to embed high-dimensional regularities with a low-dimensional representation.

5.2.1 Restricted Boltzmann Machines

The learning algorithms for connected Boltzmann Machines are notoriously slow. The expectation values under the probability distribution P of the BM, are hard to evaluate. One has to make assumptions about the thermalization time, and subsequently a lot of samples are needed to form trustworthy estimators. This problem can be alleviated by using a special type of BM: the Restricted Boltzmann Machine (RBM). RBMs have no connections between neurons of the same type, meaning that all hidden neurons are connected only to visible neurons and vice versa. The RBM structure is also called a *bipartite graph.* For learning problems RBMs are popular since the probabilities conditioned on the complete state of the visible/hidden units factorize, which significantly reduces the computational cost of the sampling procedure.

In quantum machine learning, RBMs have attracted attention for their great representative power, and their ability to capture quantum phenomena (see chapter chapter 6). Deep Boltzmann Machines are formed by adding more (interconnected) hidden layers, and are proven to be exponentially more efficient at representing quantum many-body states [52].

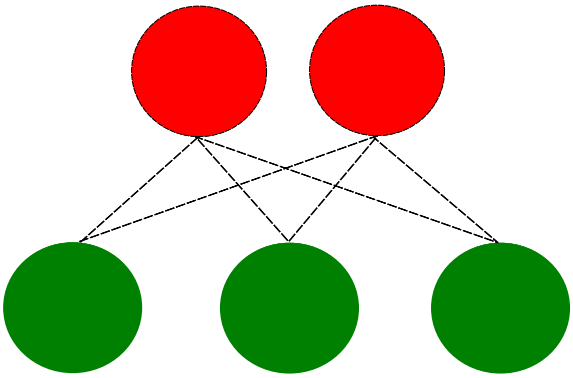


Figure 5.1: A Restricted Boltzmann machine architecture with three visible (green) and two hidden neurons (red).

5.3 Quantum Boltzmann machines

The Quantum Boltzmann Machine (QBM) was published by Mohammad Amin in 2018 [10]. The QBM yields a quantum probabilistic model, based on the quantum Boltzmann distribution. Instead of a scalar energy model based on classical spins, it defines a quantum system with Hamiltonian based on N qubits

$$H = -\sum_i w_i \sigma_i^z - \sum_{i,j} w_{ij} \sigma_i^z \sigma_j^z, \tag{5.13}$$

where w_i and w_{ij} are dimensionless parameters. The Hamiltonian is generated with tensor products of local spin operators,

$$\sigma_i^z \equiv \overbrace{I \otimes \ldots \otimes I}^{i-1} \otimes \sigma^z \otimes \overbrace{I \otimes \ldots \otimes I}^{N-i}, \tag{5.14}$$

where σ^z denotes the Pauli z-matrix. The density matrix, defined as

$$\rho = Z^{-1} e^{-H}, \tag{5.15}$$

is diagonal, since the matrix exponent $e^{-H} = \sum_{k=0}^{\infty} (-H)^k / k!$ is diagonal. Note that inverse temperature β which is typically included in the (quantum) Boltzmann distribution is an elusive parameter when considering parameterized Hamiltonians. Temperature can be absorbed by the weights so that the zero-temperature approximation $\beta \to \infty$ implies $w_r \to \infty$ for all linear parameters w_r.

Let $\mathbf{x} = \{\mathbf{v}, \mathbf{h}\}$ denote a configuration of visible and hidden spin configurations. The marginal distribution is obtained by taking the partial trace over ρ, denoted as

$$P(\mathbf{v}) = \mathrm{Tr}[\Lambda(\mathbf{v})\rho], \tag{5.16}$$

where $\Lambda(\mathbf{v})$ is a diagonal matrix with a 1 on the index of $\mathbf{v}$ and zeros elsewhere. With hidden units, the partial trace can be written as

$$\Lambda(\mathbf{v}) = |\mathbf{v}\rangle \langle \mathbf{v}| \otimes \overbrace{I \otimes \ldots \otimes I}^{N_h}. \tag{5.17}$$

Instead of using only diagonal operators in the σ^z-basis, the non-commutative nature of quantum operators could be explored by including non-diagonal elements in the

Hamiltonian. This is achieved through the addition of σ_i^x, σ_i^y, $\sigma_i^x\sigma_j^x$ and $\sigma_i^y\sigma_j^y$ terms in the Hamiltonian, each weighted accordingly.

The next step is finding a quantum equivalent for the KL divergence in order to minimize the discrepancy between some data distribution of observed spin configurations and the model. This is done by substituting the expression for the quantum marginal of equation 5.16 in the equation for the (classical) negative log-likelihood of equation 5.10, yielding

$$\mathcal{L} = -\sum_{\mathbf{v}} Q(\mathbf{v}) \ln \frac{\mathrm{Tr}[\Lambda(\mathbf{v})e^{-H}]}{\mathrm{Tr}[e^{-H}]}. \tag{5.18}$$

Performing gradient descent on this expression would be possible if the gradient w.r.t. the model parameters would be easily evaluated with the chain rule as was the case for the classical gradient. However, non-commutativity $[e^{-H}, \partial_\theta H] \neq 0$ throws a spanner in the works, introducing integrals that are hard to evaluate in comparison to the classical thermodynamic averages.

The non-vanishing commutator implies that the chain rule cannot be applied trivially. Instead, the matrix exponential is rewritten using the Trotter-Suzuki algorithm [53]: $e^{-H} = (e^{-\delta\tau H})^n$, where $\delta\tau \equiv 1/n$ denotes *imaginary time*, so that the anti-commutator can be assumed $[e^{-\delta\tau H}, \partial_\theta H] = \mathcal{O}(\delta\tau^2)$. This allows the differential to be rewritten by repeated application of the product rule:

$$\begin{aligned}
\partial_\theta e^{-H} &= \partial_\theta \overbrace{\left[e^{-\delta\tau H} \cdots e^{-\delta\tau H}\right]}^{n} && (5.19)\\
&= e^{-\delta\tau H}\partial_\theta e^{-(n-1)\delta\tau H} - e^{-1\delta\tau H}\delta\tau\partial_\theta H e^{-(n-1)\delta\tau H} && (5.20)\\
&= e^{-2\delta\tau H}\partial_\theta e^{-(n-2)\delta\tau H} - e^{-2\delta\tau H}\delta\tau\partial_\theta H e^{-(n-2)\delta\tau H} - e^{-1\delta\tau H}\delta\tau\partial_\theta H e^{-(n-1)\delta\tau H}, && (5.21)
\end{aligned}$$

where $\mathcal{O}(\delta\tau^2)$ errors are neglected. Repeating this exercise for all orders produces the compact representation

$$\partial_\theta e^{-H} = -\sum_{m=1}^{n} e^{-m\delta\tau H}\partial_\theta H \delta\tau e^{-(n-m)\delta\tau H}, \tag{5.22}$$

which can be transformed into integral form by $\delta\tau \to 0$ or $n \to \infty$:

$$\partial_\theta e^{-H} = -\int_0^1 \delta\tau e^{-\tau H}\partial_\theta H e^{(\tau-1)H}. \tag{5.23}$$

Taking the trace over the right side of equation 5.23 makes the integrand τ-independent[1]. By tracing the l.h.s as well, the result is finally

$$\mathrm{Tr}\big[\partial_\theta e^{-H}\big] = -\mathrm{Tr}\big[e^{-H}\partial_\theta H\big]. \tag{5.24}$$

[1] Using the cyclic permutation property Tr[ABC] = Tr[BCA] = Tr[CAB], and the fact that the trace operation ”commutes” with the integral.

By substituting equation 5.24 in equation 5.18, the gradient of the log-likelihood can be written as

$$\partial_\theta \mathcal{L} = \sum_{\mathbf{v}} Q(\mathbf{v}) \Big(\frac{\mathrm{Tr}[\Lambda(\mathbf{v}) \partial_\theta e^{-H}]}{\mathrm{Tr}[\Lambda(\mathbf{v}) e^{-H}]} - \frac{\mathrm{Tr}[\partial_\theta e^{-H}]}{\mathrm{Tr}[e^{-H}]} \Big) \tag{5.25}$$

$$= \sum_{\mathbf{v}} Q(\mathbf{v}) \Big(\frac{\mathrm{Tr}[\Lambda(\mathbf{v}) \partial_\theta e^{-H}]}{\mathrm{Tr}[\Lambda(\mathbf{v}) e^{-H}]} - \langle \partial_\theta H \rangle_\rho \Big), \tag{5.26}$$

where $\langle \cdots \rangle_\rho \equiv \mathrm{Tr}[\rho \cdots]$ denotes the expectation value w.r.t. the model density matrix.

However, the first term in the log-likelihood gradient is hard to evaluate. It produces a number of high-dimensional integrals equal to the number of data vectors, requiring an equal number of MC processes in order to generate estimates. In the original QBM paper [10], this requirement is circumvented by minimizing the upper bound of $\mathcal{L}$, instead of $\mathcal{L}$ itself. The disadvantage of this approach is that the quantum weights are treated as hyperparameters (they are not learnable).

The QBM structure can also be used for discriminative supervised tasks by clamping part of the visible neurons as input. The likelihood function for discriminative learning depends on a distribution conditioned on the output.

5.3.1 Quantum models for classical data

The QBM discussed in the previous section sets the stage for quantum hardware implementations. However, the proposed software implementation based on the bounded likelihood is unable to learn transverse field parameters in e.g. the transverse field Ising model.

Instead of translating density matrices to probability distributions with projection operators, another approach is proposed by Kappen [54]. In this paper, the data is encoded in a *data density matrix* directly, denoted η. The model density matrix ρ describes the quantum probability distribution of the unclamped QBM. Having a quantum description for the model and data allows one to use a quantum metric for the discrepancy between the two. This metric is the *quantum relative entropy*, an extension of the Kullback-Leibler divergence for density matrices. The relative entropy is given by

$$S(\eta|\rho) = \mathrm{Tr}(\eta \log \eta) - \mathrm{Tr}(\eta \log \rho). \tag{5.27}$$

The first term is the *von Neumann entropy* of η. The second term (excluding the minus sign) is the only model-dependent term, and is dubbed the quantum log-likelihood. Optimization of the model density matrix is done by maximizing this term. Gradients of this likelihood still scale exponentially with system size as was the case in the learning algorithm of the previous section. However, the great advantage is that optimizing the quantum metric instead of the classical KL divergence allows all parameters, including w_i^k and w_{ij}^k for $k = x, y$ to be learned by the QBM.

The quantum relative entropy is a thoroughly studied object with many known identities (see e.g. [55][56]), which proves helpful in the analysis of experimental results of the QBM. One important property of the quantum relative entropy is Klein's inequality that states $S(\eta|\rho) \geq 0$ with equality only if $\rho = \sigma$. Of equivalent importance is the joint convexity of S:

$$S((1-t)\eta_0 + t\eta_1 | (1-t)\rho_0 + t\rho_1) \leq (1-t) S(\eta_0|\rho_0) + t S(\eta_1 \rho_1), \tag{5.28}$$

for $\eta_0, \eta_1, \rho_0, \rho_1 \in \mathbf{S}_n$, where $\mathbf{S}_n$ denotes the set of density matrices on $\mathbb{C}^n$. Both boundedness and convexity are crucial properties in order to ensure the reliability of gradient descent methods performed on the likelihood-term.

We assume that our data density matrix η can be represented by a 1- and 2-local quantum spin-$\frac{1}{2}$ Hamiltonian. In other words, the model density matrix ρ is parameterized through the weights in the model Hamiltonian by $\rho = e^{-H}/Z$ with

$$H = \sum_{k=x,y,z} \sum_{i=1}^{n} w_i^k \sigma_i^k + \sum_{i=1,j>i}^{n} w_{ij}^k \sigma_i^k \sigma_j^k. \tag{5.29}$$

The likelihood can be rewritten as

$$\begin{aligned}
\mathcal{L} &= \operatorname{Tr}(\eta \log \rho) && (5.30)\\
&= \operatorname{Tr}(\eta(\log e^{-H} - \log Z)) && (5.31)\\
&= -\operatorname{Tr}(\eta H) - \operatorname{Tr}(\eta \log Z) && (5.32)\\
&= -\langle H \rangle_\eta - \log Z. && (5.33)
\end{aligned}$$

Since the Hamiltonian is linear in the connection weights, $H = \sum_r w_r H_r$, we can generalize the gradient w.r.t. the model density matrix as $\partial_{w_r} H_r$. The gradients of the first term of equation 5.33 are simply $\partial_{w_r} \langle H \rangle_\eta = \langle H_r \rangle_\eta$. The second term requires Trotterization as was needed for equation 5.23, so that

$$\partial_{w_r} \operatorname{Tr}(e^{-H}) = -\operatorname{Tr}(e^{-H} \partial_{w_r} H) = -\langle H_r \rangle_\rho, \tag{5.34}$$

providing the simple gradient

$$\partial_{w_r} \mathcal{L} = \langle H_r \rangle_\rho - \langle H_r \rangle_\eta. \tag{5.35}$$

Note that the data statistics $\langle H_r \rangle_\eta$ are constant during the learning process: the gradient descent learning algorithm based on the quantum relative entropy intuitively converges to a model density matrix ρ that reproduces the spin statistics under η, as was the case for the classical Boltzmann machine.

Substituting our specific parameterization in the gradient gives the learning rule $\Delta w_r = -\epsilon \partial_{w_r} \mathcal{L}$, with ϵ a small positive parameter, producing

$$\begin{aligned}
\Delta w_i^k &= \epsilon\Big(\langle \sigma_i^k \rangle_\eta - \langle \sigma_i^k \rangle_\rho\Big), && (5.36)\\
\Delta w_{ij}^k &= \epsilon\Big(\langle \sigma_i^k \sigma_j^k \rangle_\eta - \langle \sigma_i^k \sigma_j^k \rangle_\rho\Big), && (5.37)
\end{aligned}$$

for $k = x, y, z$. When the parameterization is limited to the (classical) z-axis ($k = z$), the classical BM learning rule is immediately retrieved.

This QBM method is able to learn classical and quantum parameters. This allows one to learn a wider range of distributions w.r.t. the classical BM, in particular distributions with inherent quantum correlations (e.g. parity models). Moreover, having a quantum model ρ for a classical data density matrix η allows one to sample efficiently from a classical distribution on quantum hardware.

The challenge lies in obtaining the free statistics $\langle \sigma_r \rangle_\rho$. The Hamiltonian is fully connected and does not adhere to any conventional quantum statistical law, and could demonstrate any of these hard to capture characteristics: glassiness, degeneracy, frustration, volume entanglement and long-range order.

Data density matrix

The data density matrix, η, may originate from experimental expectation values on a quantum lattice with unknown parameterization, or from classical data. In the quantum case, learning can be done directly since the expectation values $\langle H_r \rangle_\eta$ required for the learning rule are known.

In the classical case, a translation is necessary in order to embed the data in a quantum object. This is done by a direct application of Born's rule[2], also referred to as *Qsample* encoding [57]. This method provides an encoding for n-dimensional binary data $\mathbf{x} \in \{0,1\}^{\otimes n}$, distributed according to function $\tilde{\eta}$ and normalized so that $\sum_{\mathbf{x}} \tilde{\eta}(\mathbf{x}) = 1$. The quantum state encoding of this distribution is given by $\sum_{\mathbf{x}} \sqrt{\tilde{\eta}(\mathbf{x})}\, |x\rangle$. The empirical wavefunction (or *qsample*) corresponding to the classical distribution is decomposed as

$$|\phi\rangle = \sum_{i=1}^{N} \sqrt{\tilde{\eta}(\mathbf{x}^i)}\, |i\rangle\,, \tag{5.38}$$

where $2^n \equiv N$. The rank-1 density matrix corresponding to this wavefunction is

$$\eta = |\phi\rangle\langle\phi| = \sum_i \sum_j \sqrt{\tilde{\eta}(\mathbf{x}^i)\tilde{\eta}(\mathbf{x}^j)}\, |i\rangle\langle j|\,. \tag{5.39}$$

Representing the density matrix in the computational basis, states are denoted according to $\mathbf{s}^k = (s_1^k, \ldots, s_n^k)$ where s_i^k denotes the eigenvalue of basis state k and qubit i. Equivalently, $|s_i^k\rangle = |\pm 1\rangle$ with $|{+1}\rangle \equiv |\uparrow\rangle$ and $|{-1}\rangle \equiv |\downarrow\rangle$.

Quantum correlations can be extracted from the classical data distribution by noting

$$\sigma_i^z |s_i\rangle = s_i |s_i\rangle \qquad \sigma_i^x |s_i\rangle = F_i |s_i\rangle \qquad \sigma_i^y |s_i\rangle = i s_i F_i |s_i\rangle\,, \tag{5.40}$$

where spin-flips are denoted with the spin-flip operator $F_i |s_i\rangle \equiv |{-s_i}\rangle$. In particular, the data spin statistics are

$$\begin{aligned}
\langle \sigma_i^x \rangle_\eta &= \sum_{k=1}^{N} \sqrt{\tilde{\eta}(F_i \mathbf{s}^k)\tilde{\eta}(\mathbf{s}^k)} & \langle \sigma_{ij}^x \rangle_\eta &= \sum_{k=1}^{N} \sqrt{\tilde{\eta}(F_i F_j \mathbf{s}^k)\tilde{\eta}(\mathbf{s}^k)} \\
\langle \sigma_i^y \rangle_\eta &= 0 & \langle \sigma_{ij}^y \rangle_\eta &= -\sum_{k=1}^{N} s_i^k s_j^k \sqrt{\tilde{\eta}(F_i F_j \mathbf{s}^k)\tilde{\eta}(\mathbf{s}^k)} \\
\langle \sigma_i^z \rangle_\eta &= \sum_{k=1}^{N} s_i^k \sqrt{\tilde{\eta}(\mathbf{s}^k)} & \langle \sigma_{ij}^z \rangle_\eta &= \sum_{k=1}^{N} s_i^k s_j^k \sqrt{\tilde{\eta}(\mathbf{s}^k)}
\end{aligned} \tag{5.41}$$

The classical distribution $\tilde{\eta}$ is built by counting occurrences of samples $\#(\mathbf{s}^k)$ in the data set X and setting $\tilde{\eta}(\mathbf{s}^k) = \#(\mathbf{s}^k)/|X|$. Consequently, the quantum statistics are computed with a complexity linear in $|X|$.

5.3.2 Rank-1 approximation

Although calculating expectation values w.r.t. the model density matrix is an intractable task for large systems, numerical methods exist (chapter 6) to approximate

[2] Interpreting the absolute square of amplitude coefficients $|\alpha_i|$ as probability of measuring.

the ground state statistics of a Hamiltonian. In particular, zero-temperature Monte Carlo methods such as Variational Monte Carlo (see section 6.2) could be used to retrieve the minimally required correlations in order to train the QBM.

Using only ground state statistics to calculate the expectation values in the learning rule of equation 5.35 constitutes the rank-1 approximation. For classical data and ground state statistics of Hamiltonians that can be mapped to the QBM Hamiltonian, this approximation for ρ is motivated by the fact that the classical density matrix is rank-1 as well.

The expressibility assumption entails that the parameterization of the Hamiltonian possesses the representational power to model the data wavefunction $|\phi\rangle$ with the ground state $|\psi_0\rangle$ so that $\langle\phi|\psi_0\rangle \approx 1$, in which case $|\phi\rangle$ is called *identifiable*. There is no a priori reason to assume that the data wavefunction $|\phi\rangle$ is identifiable. In fact, this is only known when the $|\phi\rangle$ generating the data statistics is known to be a ground state originating from the model parameterization class (equation 5.29). In the rank-1 approximation, the quantum relative entropy is not strictly convex, making quantum state tomography impossible: ρ is merely trained to produce the best parameterization for a generative model. The gradients are given by

$$\begin{aligned}
\langle H_r\rangle_\rho &= \mathrm{Tr}(H_r e^{-H})/Z && (5.42)\\
&= \sum_i \langle\psi_i| \Big[\Big(\sum_k E_k^r |\psi_k\rangle\langle\psi_k|\Big)\Big(\sum_{k'} e^{-E_{k'}} |\psi_{k'}\rangle\langle\psi_{k'}|\Big)\Big] |\psi_i\rangle /Z && (5.43)\\
&= \sum_k E_k^r e^{-E_k}/Z && (5.44)\\
&= \frac{E_0^r + \sum_{k>0} E_k^r e^{\Delta_k}}{1+\sum_{k>0} e^{\Delta_k}} && (5.45)\\
&\approx E_0^r + \mathcal{O}(e^{\Delta_1}), && (5.46)
\end{aligned}$$

where $E_k^r = \langle\psi_k|H_r|\psi_k\rangle$ and the spectral gap $\Delta_k = (E_0 - E_k) < 0$ is assumed sufficiently large. Approximating the gradient with ground state statistics therefore introduces an error of order $\mathcal{O}(e^{\Delta_1})$. Note that even though only ground state statistics are used for training, the model density matrix $\rho = e^{-H}/Z$ is generally not rank-1.

The rank-1 approximation of the likelihood

Effectively taking $\rho = |\psi_0\rangle\langle\psi_0|$ for the calculation of gradients is reminiscent of a zero-temperature approximation. However, the rank-1 approximation of the likelihood is more intricate.

When the temperature of a system consisting of a Boltzmann distributed ensemble of particles goes to absolute zero, the occupation probability of states becomes non-stochastic and the system collapses to the ground state. Denoting $\beta \equiv 1/k_b T$, the zero-temperature limit Gibbs state is given by the

$$\lim_{\beta\to\infty} \rho = \lim_{\beta\to\infty} \frac{e^{-\beta H}}{\mathrm{Tr}(e^{-\beta H})}. \qquad (5.47)$$

By the Spectral Theorem for Hermitian matrices, the Hamiltonian can be decomposed as $H = \sum_i E_i |\psi_i\rangle \langle\psi_i|$, so that $H|\psi_i\rangle = E_i |\psi_i\rangle$. This implies

$$\lim_{\beta\to\infty} \rho = \lim_{\beta\to\infty} \frac{\sum_i e^{-\beta E_i} |\psi_i\rangle \langle\psi_i|}{\mathrm{Tr}(\sum_i e^{-\beta E_i} |\psi_i\rangle \langle\psi_i|)} \tag{5.48}$$

$$= \lim_{\beta\to\infty} \frac{\sum_i e^{-\beta(E_i - E_0)} |\psi_i\rangle \langle\psi_i|}{\sum_i e^{-\beta(E_i-E_0)}} \tag{5.49}$$

$$= |\psi_0\rangle \langle\psi_0| . \tag{5.50}$$

Using equation 5.33, it becomes evident the quantum likelihood can be approximated with only ground state information:

$$\mathcal{L} = -\langle H\rangle_\eta - \log\left(\sum_k e^{-E_k}\right) \tag{5.51}$$

$$= -\sum_r w_r \langle\sigma_r\rangle_\eta - \Big[-E_0 + \log\Big(1 + \sum_{k>0} e^{\Delta_k}\Big)\Big] \tag{5.52}$$

$$\mathcal{L}_1 \approx -\sum_r w_r \langle\sigma_r\rangle_\eta + E_0. \tag{5.53}$$

Note that the first term in equation 5.53 would diverge in the zero-temperature limit, whereas $\lim_{\beta\to\infty} \log Z = E_0$. The rank-1 approximation of the likelihood can therefore be understood as a spectral-gap assumption, rather than zero-temperature.

Furthermore, it follows that $\mathcal{L}_1 \leq 0$ for a rank-1 data density matrix $\eta = |\phi\rangle\langle\phi|$ with $|\phi\rangle = \sum_k c_i |\psi_i\rangle$. This can be demonstrated by substituting the rank-1 density matrix in equation 5.53, s.t.

$$\mathcal{L}_1 = -\sum_{ijk} \langle\psi_i| c_i^* \Big[E_j |\psi_j\rangle\langle\psi_j|\Big] c_k |\psi_k\rangle + E_0 \tag{5.54}$$

$$= -\sum_k |c_k|^2 E_k + E_0 \tag{5.55}$$

$$= (1 - |c_0|^2)E_0 - \sum_{k>0} E_k |c_k|^2 \tag{5.56}$$

$$= \sum_{k>0} |c_k|^2 \Delta_k \leq 0, \qquad \Big(1 - |c_0|^2 = \sum_{k>0} |c_k|^2\Big) \tag{5.57}$$

with equality only when the data vector is perfectly retrieved s.t. $|c_1|^2 = 1$.

Conclusion

The QBM optimization problem can be reformulated in the rank-1 approximation as

$$\max \mathcal{L}_1 = \max\Bigg\{ -\sum_{k=x,y,z} \Big[\sum_{i=1}^n w_i^k \langle\sigma_i^k\rangle_\eta + \sum_{i=1,j>i}^n w_{ij}^k \langle\sigma_i^k \sigma_j^k\rangle_\eta\Big] + E_0 \Bigg\},$$

constrained by

$$H|\psi_0\rangle = E_0 |\psi_0\rangle, \quad H = \sum_{k=x,y,z} \Big[\sum_{i=1}^n w_i^k \sigma_i^k + \sum_{i=1,j>i}^n w_{ij}^k \sigma_i^k \sigma_j^k\Big].$$

Calculating the gradients of $\mathcal{L}$ with ground state statistics induces a $\mathcal{O}(e^{\Delta_1})$ error. Additionally, stochastic errors are induced when the ground state statistics are obtained numerically.

It is possible to approximate the quantum likelihood $\mathcal{L}_1$ with only the weights of the QBM and the ground state energy of the corresponding Hamiltonian. The latter is subject to errors in the estimate of the ground state energy E_0, when obtained numerically. The rank-1 likelihood itself is subject to the spectral gap error of equation 5.53. The numerical error on $\mathcal{L}_1$ slightly impedes the experimental practicality of the QBM, since the efficacy of the QBM update rule (equation 5.37) can be increased by employing update methods adaptive w.r.t. changes in the likelihood.

The quantum relative entropy becomes an ill-defined measure for rank-1 η and ρ: if $\langle\psi|\rho|\psi\rangle$ is zero, but $\langle\psi|\eta|\psi\rangle$ is finite, the quantum relative entropy diverges [58]. Therefore, one should be careful to apply a likelihood-interpretation to $\mathrm{Tr}(\eta\log\rho)$ for rank-1 data and model density matrices. However, the only true assumption for approximating $\mathcal{L}$ and its gradients, is a large spectral gap for H. The (full) model density matrix $\rho = e^{-H}/Z$ is generally not rank-1.

5.4 Entanglement properties

Numerous efforts are made to classify quantum spin Hamiltonians on basis of their ground state entanglement scaling law. A fairly recent overview is given in aforementioned Ref. [27], where results are mentioned of many one-dimensional spin chains, i.a. the (critical) Ising model, the XY model, and disordered chains. However, with regards to the QBM, general statements cannot be made about the entanglement scaling law for a Quantum Boltzmann Hamiltonian (QBH) parameterized by random 1- and 2-spin interactions (equation 5.13). In fact, a QBH is generally non-local, rendering most research on entanglement scaling irrelevant. Long-range interaction Hamiltonians have been examined in Ref. [59], where entanglement is characterized in terms of non-locality of the weights, or equivalently the algebraic $1/r^{\alpha}$ scaling of interaction-weights with distance r and "small" α. Yet, even this description is too restrictive for QBH's, the weights of which do not follow any a priori law.

In conclusion, the ground states of QBH's should be assumed to adhere to volume law, since the QBM learning algorithm generally encounters quantum critical and non-local Hamiltonians.

5.5 Experiments

The Quantum Boltzmann Machine allows one to reconstruct density matrices from arbitrary sources of N spins using the model class defined by $W_{\mathrm{QBM}} = \{w_i^k, w_{ij}^k | i, j = 1 \dots N; k = x, y, z\}$ of equation 5.13 by the update rule of equation 5.37. The update rule relies on expectation values of the operators $\{\sigma_i^k, \sigma_i^k\sigma_j^k | i, j = 1, \dots, N; k = x, y, z\}$. The distance measure between the data and model density matrix, η and ρ respectively, is measured by the likelihood L (see equation 5.27). Quantum state tomography can be performed by using a full rank model density matrix. The rank-1 algorithm should return a generative model that reproduces the statistics.

5.5.1 Full rank

In the case of Quantum State Tomography (QST), the QBM algorithm is demonstrated by algorithm 1. As an example, this algorithm is used to perform QST on the XXX AFH model (see figures 5.2-5.6). The correct weights are found (and consequently the correct statistics generated) for the $N = 10$ AFH model after optimization of a randomly initialized model density matrix ρ. The same experiment was repeated for a fully connected spin-glass (see figures 5.7-5.11). As discussed in [54], the rate of convergence depends on the magnitude of the weights. η becomes rank-1 in the limit $\beta \to \infty$, meaning that S loses its strict convexity property. In experiment, this is expressed by slower convergence for larger weights.

Algorithm 1: The basic QBM algorithm for tomography

```
Result: Optimized QBM model weights $W_{\text{QBM}}$ with $\mathcal{L} \approx 0$
max_iter = 500 ;
momentum = 0.5 ;
$\epsilon = 0.02$ ;
Retrieve the data statistics $\langle H_r \rangle_\eta$ ;
i=1 ;
Initialize model weights $W_{\text{QBM}}$ randomly ;
while i < max_iter do
    Calculate $\langle H_r \rangle_\rho$ ;
    Update $W_{\text{QBM}}$ by $\Delta w_r = \epsilon(\langle H_r \rangle_\eta - \langle H_r \rangle_\rho)$ ;
    Calculate L, ΔL ;
    i+=1 ;
end
```

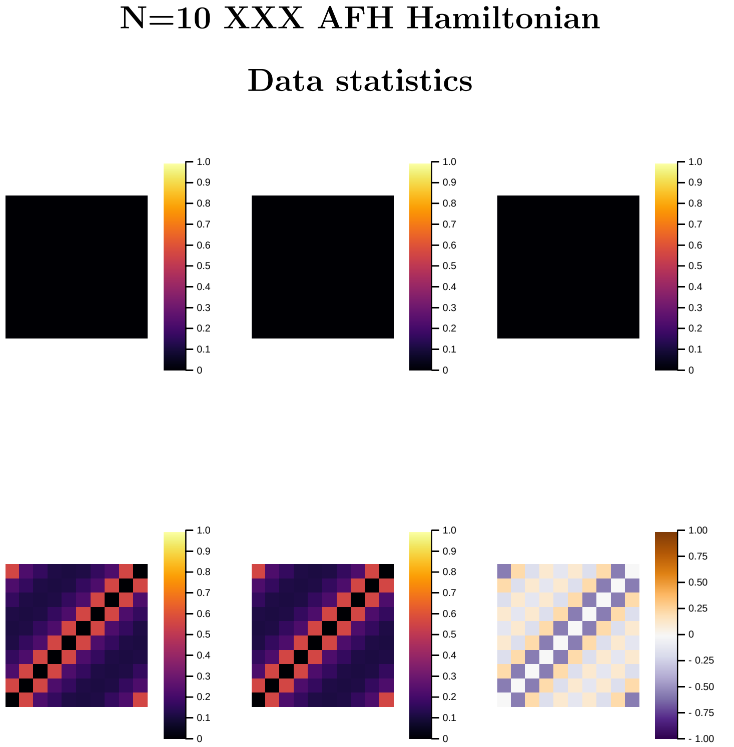


Figure 5.2: The single site (top row) and correlation (bottom row) x, y, z-statistics (respectively in the first to third column) of the data density matrix η in the XXX AFH model.

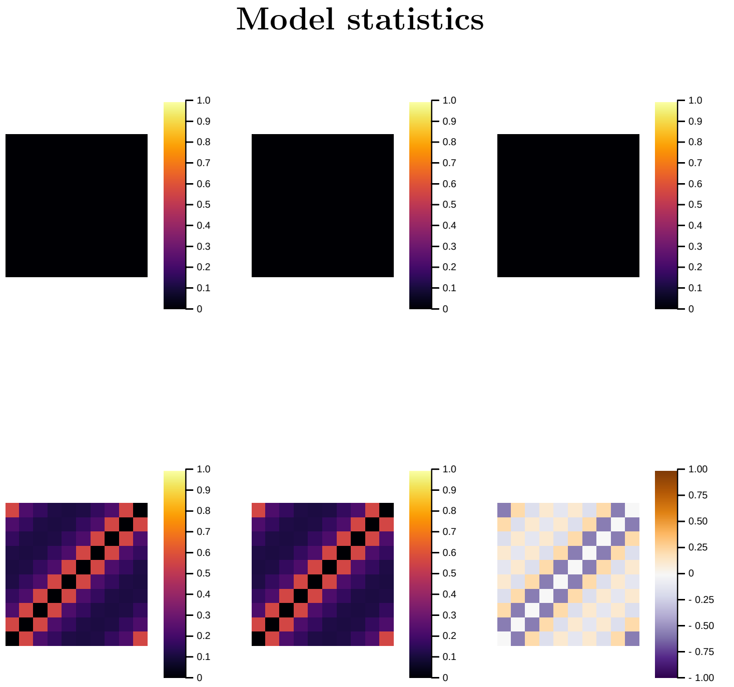


Figure 5.3: The single site (top row) and correlation (bottom row) x, y, z-statistics (respectively in the first to third column) of ρ after optimization. Statistics are reproduced reliably with machine precision for the single site statistics and $\mathcal{O}(10^{-6})$ RMS errors for the correlations.

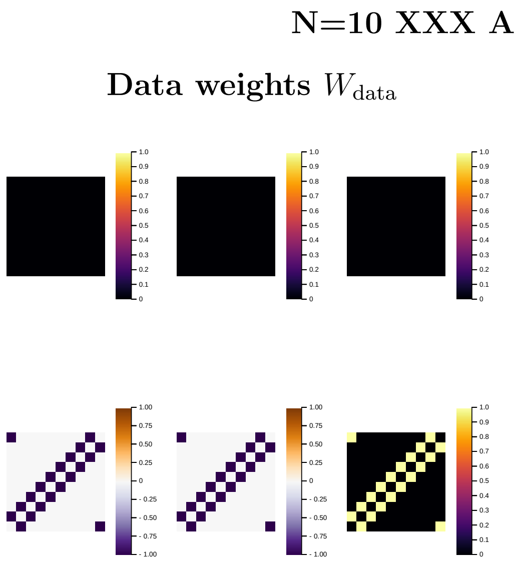


Figure 5.4: The weights of the data Hamiltonian from which the ground state statistics are calculated.

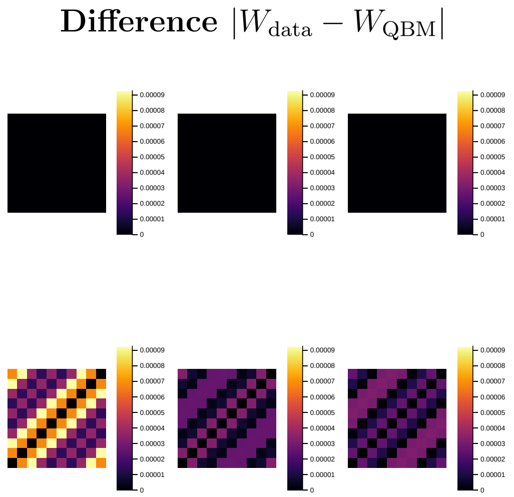


Figure 5.5: Machine precision tomography on single site weights. RMS errors of $\mathcal{O}(10^{-4})$ for w_{ij}^x and $\mathcal{O}(10^{-5})$ for w_{ij}^y and w_{ij}^z.

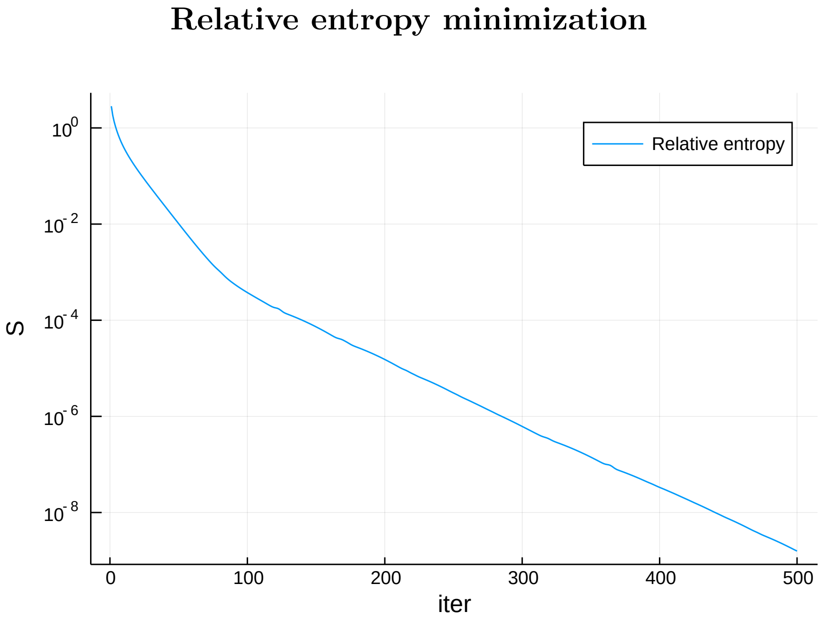


Figure 5.6: Relative entropy converges to zero. Learning stopped forcibly after 500 iterations.

N=10 Spin glass Hamiltonian

Data statistics

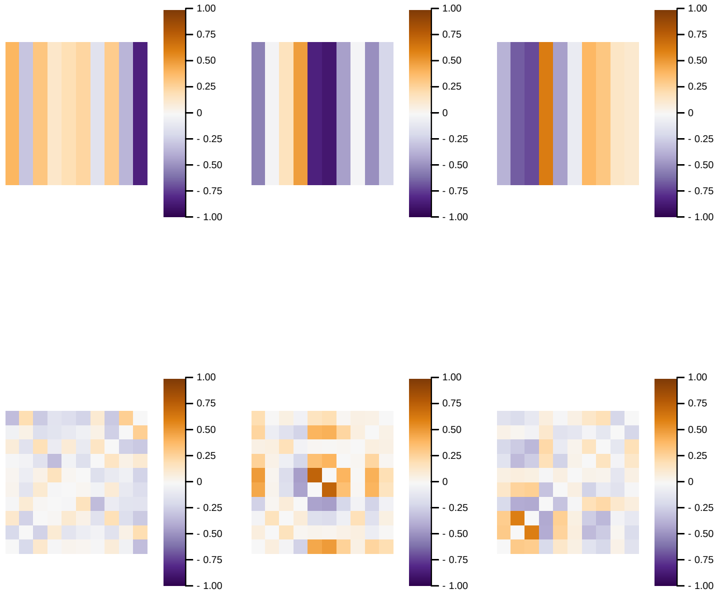


Figure 5.7: The single site (top row) and correlation (bottom row) x, y, z-statistics (respectively in the first to third column) of the data density matrix η of a spin glass Hamiltonian with couplings $w_{ij}^{k} \sim \mathcal{N}(0, 1/\sqrt{N})$ and $w_i^{x,z} \sim \mathcal{N}(0, 1)$.

Model statistics

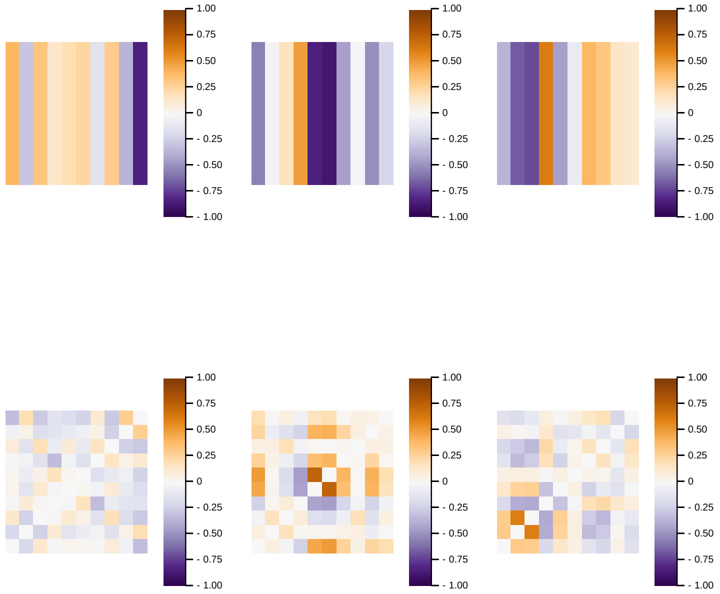
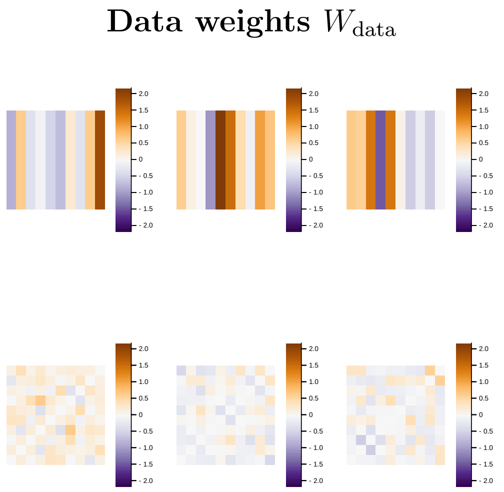


Figure 5.8: The single site (top row) and correlation (bottom row) x, y, z-statistics (respectively in the first to third column) of ρ after optimization. Statistics are reproduced reliably with an RMS error of $\mathcal{O}(10^{-5})$.

N=10 Spin glass Hamiltonian



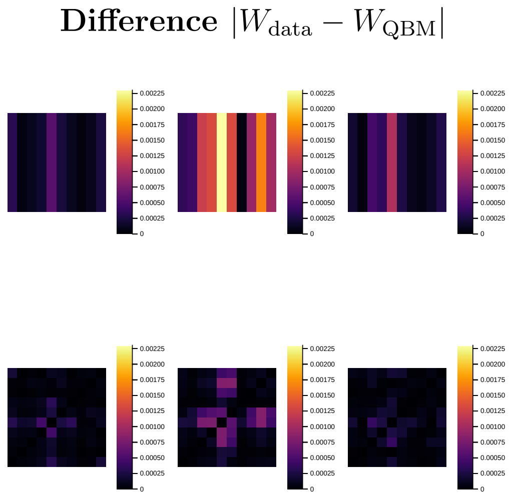


Figure 5.9: The QBM weights that generate Hamiltonian from which the ground state statistics are calculated.

Figure 5.10: $\mathcal{O}(10^{-3})$ RMS error for w^y and $\mathcal{O}(10^{-4})$ RMS errors for all other weights.

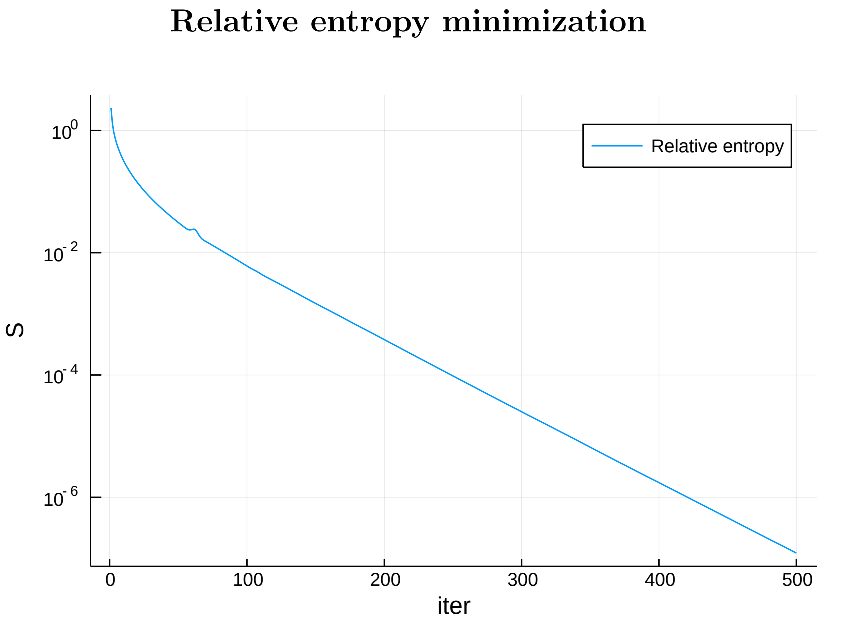


Figure 5.11: Relative entropy converges to zero. Learning stopped forcibly after 500 iterations.

5.5.2 Rank-1

The full-rank experiments are repeated, but now using the rank-1 approximation. The data statistics are generated by calculating the ground state of the Hamiltonian generated by the weights W_{data}. This forces the data statistics to be reproducible by the ground state of a model density matrix ρ. In general, e.g. when the data comes from a classical distribution, this cannot be ensured. The algorithm is summarized in algorithm 2. Note that a smaller learning rate without momentum was used for reliability. With momentum, the likelihood often showed oscillatory behaviour instead of smooth convergence.

The results of the (Marshall-Peierls) AFH model (figures 5.12-5.16) and spin glass Hamiltonian (figures 5.17-5.21) are similar to the full-rank case, with the major distinction that the weights cannot be reproduced. This can be explained by the large but shallow basin of attraction towards the minimum of S: many sets of weights W_{QBM} correspond to a small quantum relative entropy. Moreover, the QBM was trained for a classical data set, consisting of a neuronal recording of a salamander. While this probability vector is not necessarily identifiable (see section 5.3.2), the data statistics were reproduced with low RMS errors (figures 5.22-5.25).

Algorithm 2: The rank-1 QBM algorithm for learning a generative model ρ

```
Result: Optimized QBM model weights $W_{\text{QBM}}$ with $\mathcal{L}_1 \approx 0$
max_iter = 1000 ;
momentum = 0.0 ;
$\epsilon = 0.01$ ;
Retrieve the data statistics $\langle H_r \rangle_\eta$, where $\eta$ is rank-1 ;
i=1 ;
Initialize model weights $W_{\text{QBM}}$ randomly ;
while i < max_iter do
    Calculate $\langle H_r \rangle_\rho$, where $\rho$ is the ground state density matrix of $H$ ;
    Update $W_{\text{QBM}}$ by $\Delta w_r = \epsilon(\langle H_r \rangle_\eta - \langle H_r \rangle_\rho)$ ;
    Calculate L, ΔL ;
    i+=1 ;
end
```

N=10 XXX AFH Hamiltonian

Data statistics

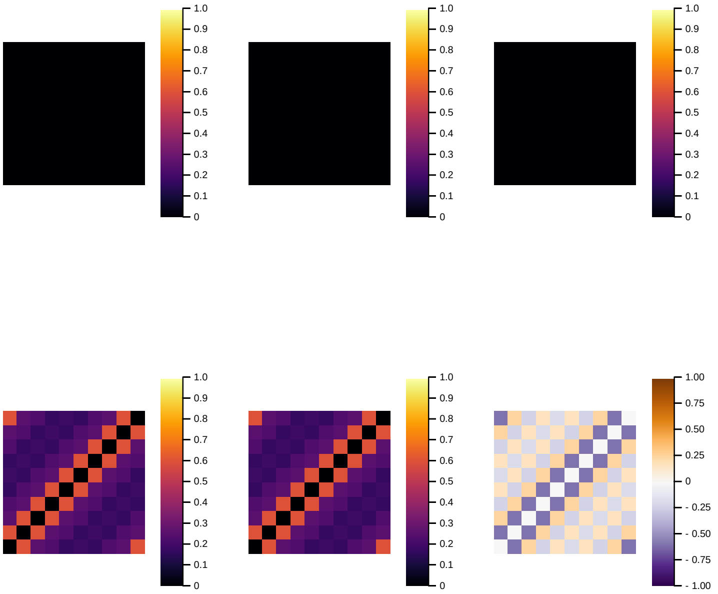


Figure 5.12: The single site (top row) and correlation (bottom row) x, y, z-statistics (respectively in the first to third column) of the rank-1 data density matrix η, constructed from the ground state of a spin glass Hamiltonian with XXX AFH couplings.

Model statistics

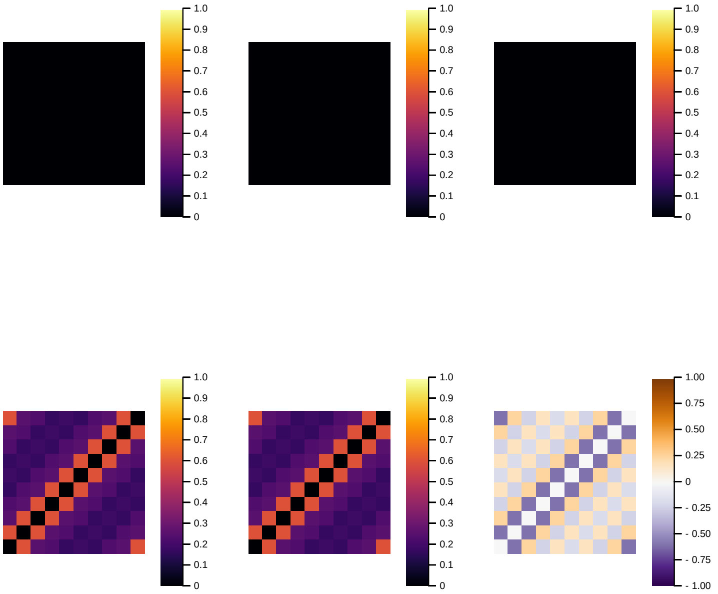
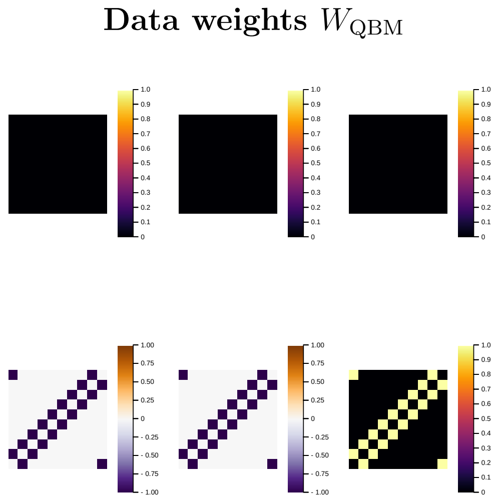


Figure 5.13: The single site (top row) and correlation (bottom row) x, y, z-statistics (respectively in the first to third column) of (rank-1) ρ after optimization. Statistics are reproduced with machine precision for the single site statistics, and RMS errors of $\mathcal{O}(10^{-4})$ for correlations.

N=10 XXX AFH Hamiltonian



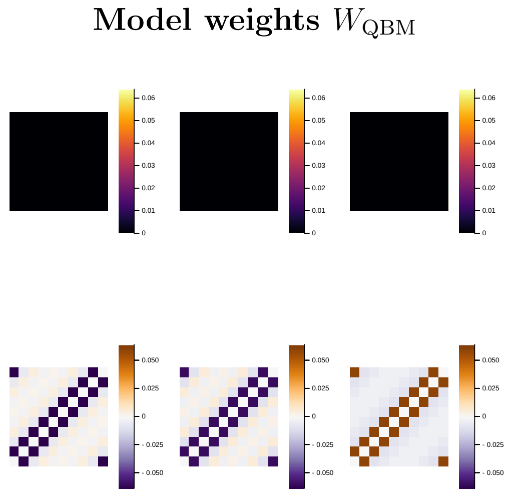


Figure 5.14: The QBM weights that generate Hamiltonian from which the ground state statistics are calculated.

Figure 5.15: The learned weights w_{ij}^k do not correspond to the exact weights of the spin glass Hamiltonian.

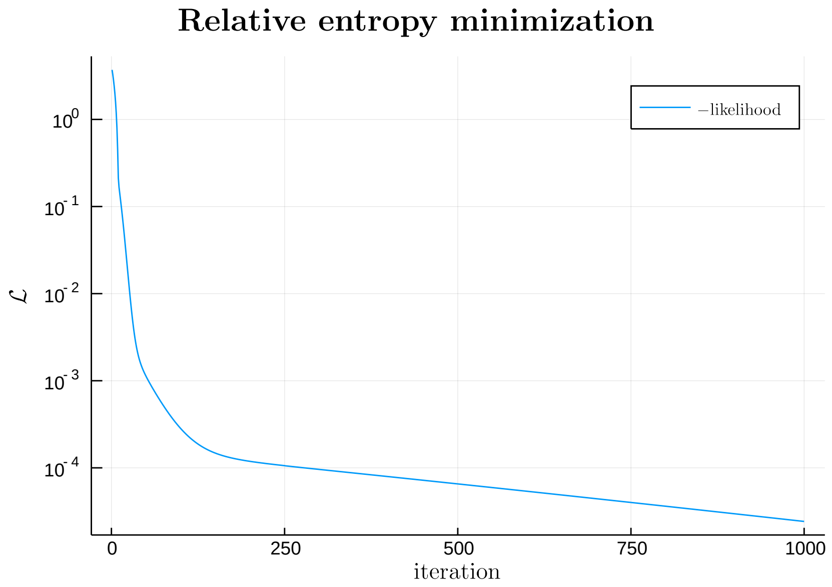


Figure 5.16: The negative likelihood converges smoothly to a low value, indicating that the rank-1 data density matrix η has been retrieved by updating ρ according to the QBM update rule. Learning stopped forcibly after 1000 iterations.

N=10 Spin glass Hamiltonian

Data statistics

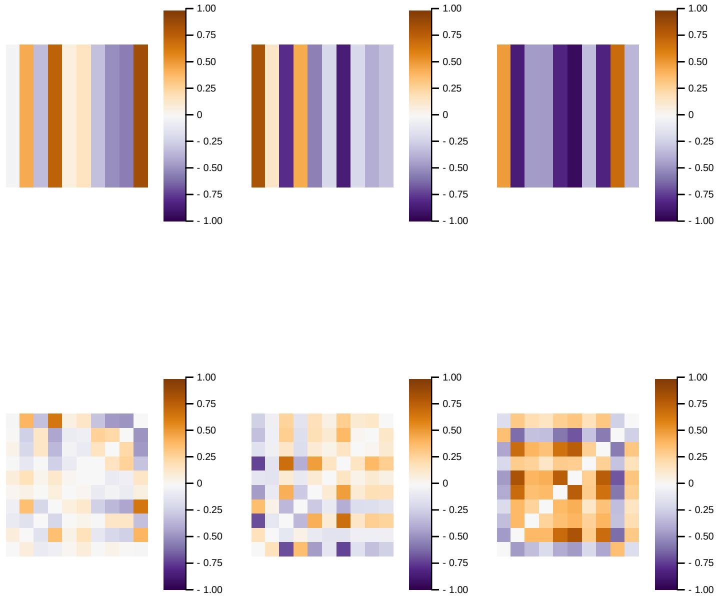


Figure 5.17: The single site (top row) and correlation (bottom row) x, y, z-statistics (respectively in the first to third column) of the rank-1 data density matrix η, constructed from the ground state of a spin glass Hamiltonian with couplings $w_{ij}^k \sim \mathcal{N}(0, 1/\sqrt{N})$ and $w_i^k \sim \mathcal{N}(0, 1)$ and a spectral gap of $|E_0 - E_1| = 0.7$.

Model statistics

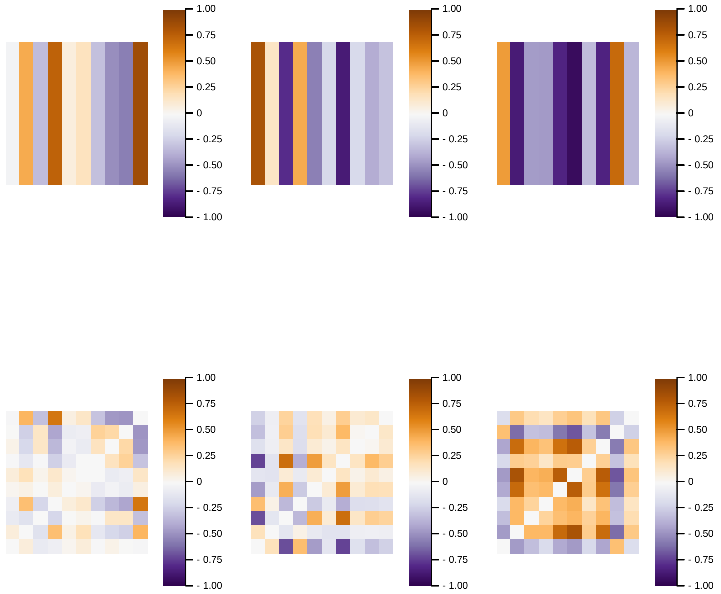
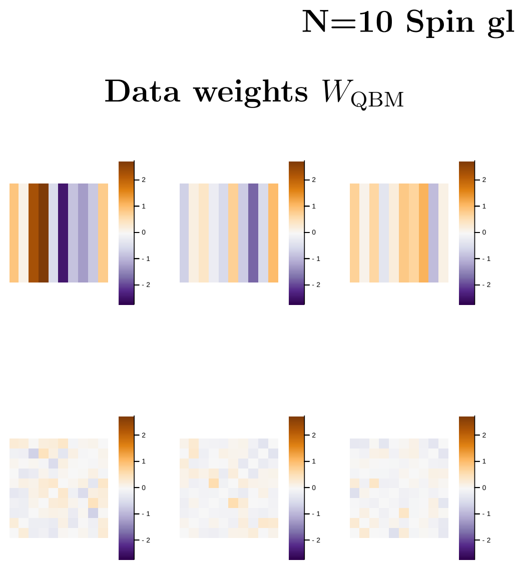


Figure 5.18: The single site (top row) and correlation (bottom row) x, y, z-statistics (respectively in the first to third column) of (rank-1) ρ after optimization. RMS errors of order $\mathcal{O}(10^{-4})$.



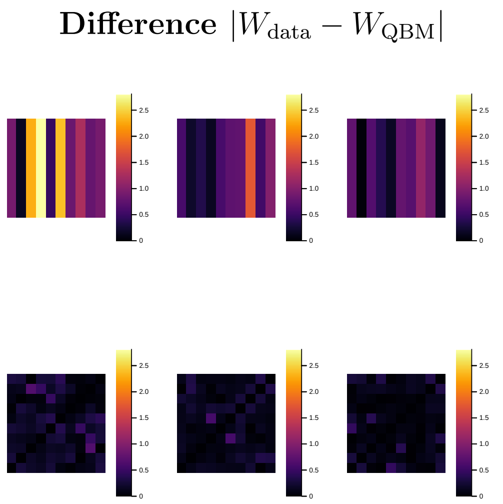


Figure 5.19: The QBM weights that generate Hamiltonian from which the ground state statistics are calculated.

Figure 5.20: The learned weights do not correspond to the exact weights of the spin glass Hamiltonian.

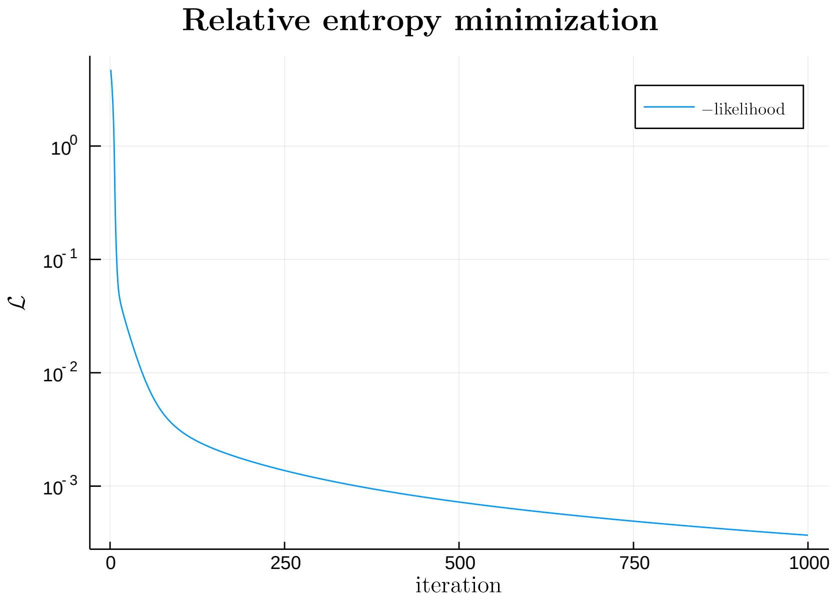


Figure 5.21: The negative likelihood converges smoothly to a low value, indicating that the rank-1 data density matrix η has been retrieved by updating ρ according to the QBM update rule. Learning stopped forcibly after 1000 iterations.

N=10 Classical data

Data statistics

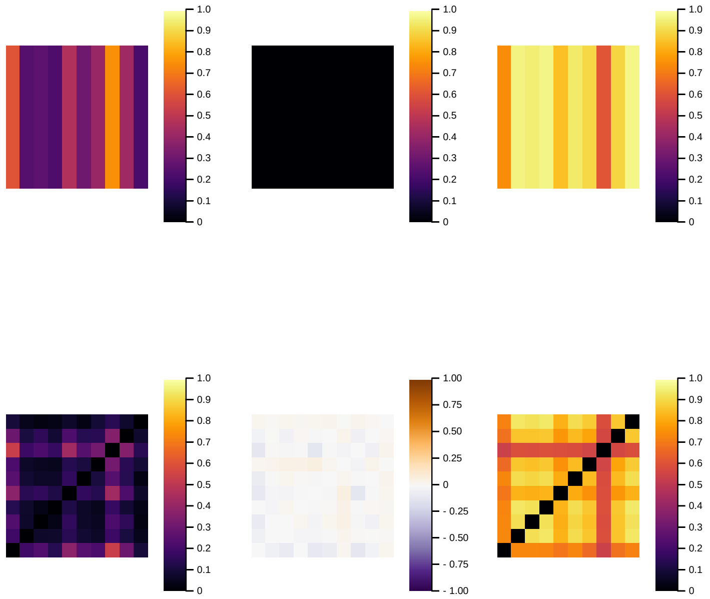


Figure 5.22: The single site (top row) and correlation (bottom row) x, y, z-statistics (respectively in the first to third column) of the rank-1 data density matrix η, constructed from a neuronal activity movie of a salamander. This classical probability vector can therefore only be projected on the manifold parameterized by W_{QBM}.

Model statistics

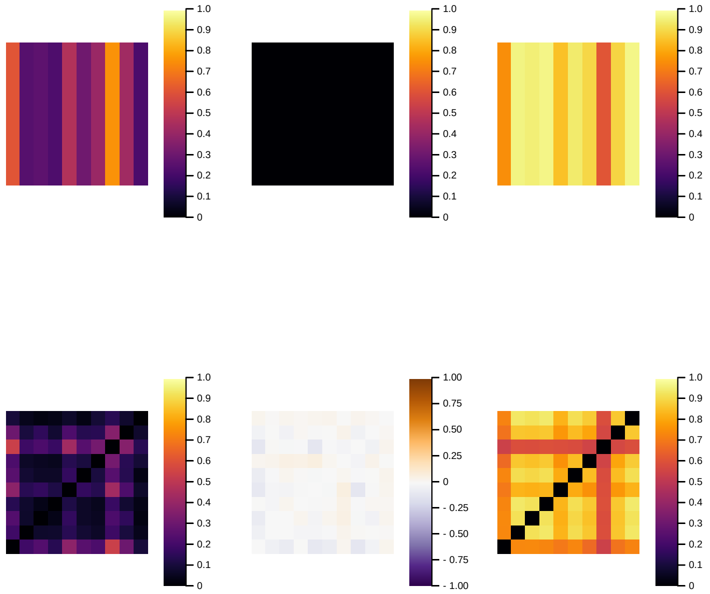


Figure 5.23: The single site (top row) and correlation (bottom row) x, y, z-statistics (respectively in the first to third column) of (rank-1) ρ after optimization. RMS errors shown in figure 5.25.

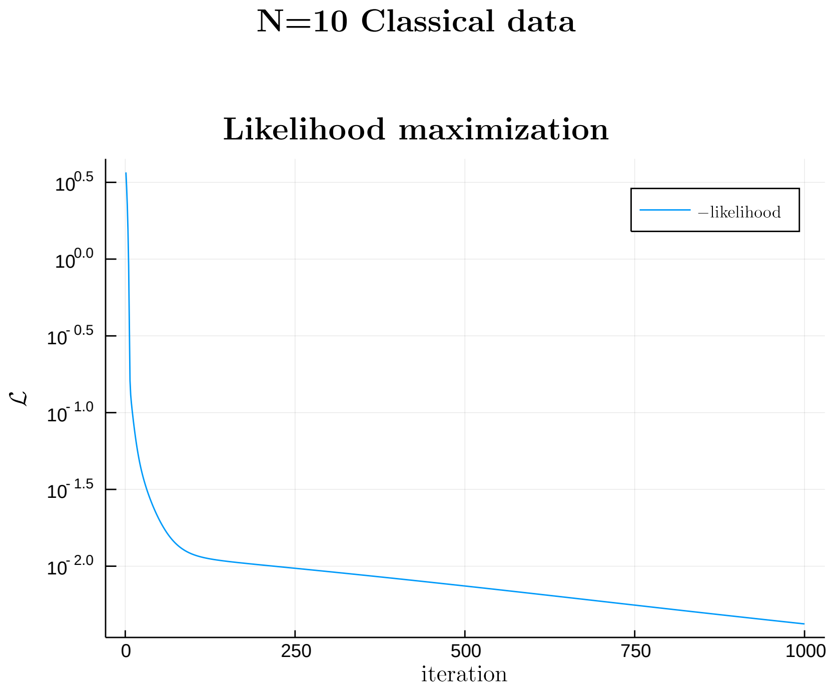


Figure 5.24: The negative likelihood converges smoothly, even though the salamander data is not taken from the QBM parameterization class.

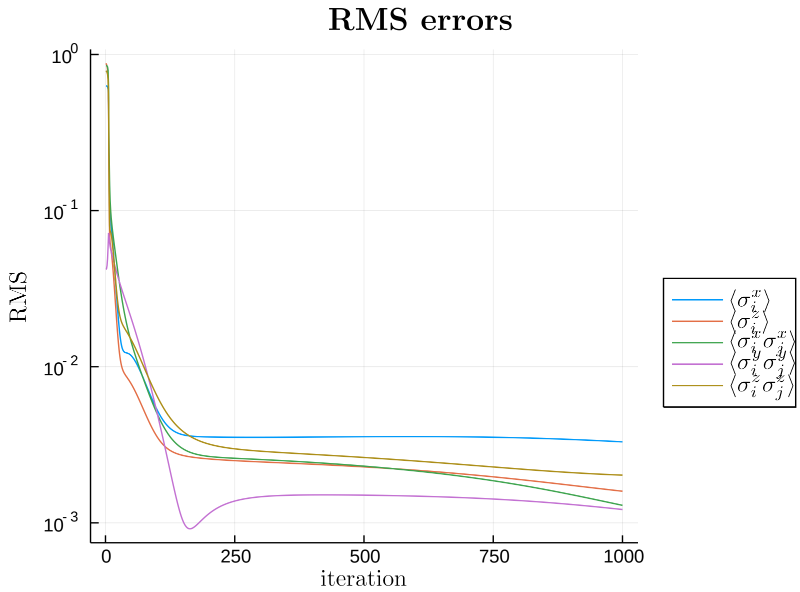


Figure 5.25: RMS error convergence of the relevant statistics.

5.6 Conclusion

Using the full spectrum of the model Hamiltonian, the QBM provides a rich model to learn Hamiltonians with the full-rank algorithm. Alternatively, with the rank-1 approximation the QBM can be used to train a generative model that reproduces the data statistics of Hamiltonians in the QBM parameterization class with high accuracy. For the full rank and rank-1 experiments, two different instances of normally distributed weights were used to see how the QBM would respond to more challenging Hamiltonians. Aside from slower convergence, this was shown not to be an issue. Moreover, for classical data the QBM was proven to provide a generative model that is able to reproduce data statistics accurately. These results demonstrate the potential of the QBM. If these experiments could be carried out on quantum hardware, they would provide a way to learn a classical probability distribution by repeated measurements on a quantum system.

In [54], it has been shown that these trained QBMs provide probability models with increased predictive power compered to the classical BM. This power was used to perform better on train-test-validation tasks, in particular for the classical neuronal data set. Moreover, the QBM is shown to be able to learn a parity problem with only visibile units.

In order to extend these promising results for larger data sets (spin models), numerical methods are needed to estimate the unclamped statistics. This is done in chapter 7, where training the QBM with a VMC method (see chapter 6) is explored.

Chapter 6

Neural Quantum States

Hilbert space scales exponentially with the number of particles. In order to investigate the physical properties of large systems, conventional methods like spectral analysis of the Hamiltonian become intractable. A general spin-1/2 quantum Hamiltonian with 30 particles represented by Pauli matrices needs $2^{60} \cdot 64 = 7 \cdot 10^{19}$ bits, or $70,000$ petabyte of classical storage. This number can be reduced somewhat by using hermiticity and sparseness, but none of these reductions is exponential. As long as these states cannot be represented faithfully by a quantum computer, the best option is to rely on numerical methods to approximate the physics at interest. The ground state and its energy are of particular interest, the understanding of which could help illuminate the phase of a material and its quantum phase transitions accordingly. Numerical methods aim to investigate these features with polynomial representations of relevant Hilbert space in polynomial time.

A recent method for finding ground state wavefunctions of spin-1/2 quantum Hamiltonians was designed by Carleo and Troyer [14] in 2017. In this paper, a variational Artificial Neural-net (ANN) wavefunction ansatz is presented. Furthermore, a reinforcement learning scheme is introduced to train the black-box model to mimic, as closely as possible, the exact ground state. The quality of their algorithm is tested by computing the ground state energy and spin correlations of large systems with specific Hamiltonians suited for comparison with other exact numerical methods (as was briefly discussed in section 3.1.2). Besides learning the ground state wavefunction, the ANN can further be used to describe dynamics as well, however this will not be subject of further discussion in this thesis.

While the original paper only discusses stoquastic (sign-positive) groundstate wavefunctions of 1D and 2D lattices, additional results are gathered for instances of the QBM Hamiltonians (equation 5.13), which are fully connected and non-uniform. These results indicate that the Neural Quantum State can be used to generate groundstate statistics for stoquastic QBM Hamiltonians. Training the (stoquastic) QBM with the approach discussed in this chapter is done in chapter 7.

6.1 Restricted Boltzmann Machines for quantum state embeddings

In chapter 5, a probabilistic model is attached to Restricted Boltzmann Machines. In particular, the occupation probability is inverse-exponential in the energy of the state. This idea is extended in the proposed RBM representation of a quantum spin lattice. In the quantum case, the probabilistic model is given by the Born description

$$P(\mathbf{s}) = |\Psi_{\mathcal{W}}(\mathbf{s})|^2, \tag{6.1}$$

where $\Psi_{\mathcal{W}}(\mathbf{s})$ denotes the local wavefunction of the visible state $\mathbf{s}$ denoted by its binary eigenvalues in the σ^z-basis[1] of the RBM parameterized by a set of complex-valued weights $\mathcal{W} = \{\mathbf{a}, \mathbf{b}, W\}$. The single-unit factors $\mathbf{a}$ and $\mathbf{b}$ are also referred to as biases. Specifically,

$$\Psi_{\mathcal{W}}(\mathbf{s}) = \sum_{\{h_i\}} \left[\exp\Big(\sum_j a_j s_j + \sum_i b_i h_i + \sum_{ij} W_{ij} b_i a_j \Big) \right], \tag{6.2}$$

where $\mathbf{h}$ denotes the state of the hidden units which take on values $h_i = \{-1, +1\}$. The hidden units are basis-independent and can therefore be seen as classical units. Adding hidden units increases the expressive power of the RBM. The ratio of the number of hidden N_h and visible units N_v is given by $\alpha \equiv N_h/N_v$.

The notation of the wavefunction can be simplified since the hidden units take on binary values and there are no intra-layer connections, so that[2]

$$\Psi_{\mathcal{W}}(\mathbf{s}) = \exp\Big(\sum_j a_j s_j \Big) \prod_i 2\cosh\Big(b_i + \sum_j W_{ij} s_j \Big). \tag{6.3}$$

For completeness sake, we write the total wavefunction as

$$|\Psi_{\mathcal{W}}\rangle = \sum_{\mathbf{s}} \Psi_{\mathcal{W}}(\mathbf{s}) \, |\mathbf{s}\rangle \, , \tag{6.4}$$

which is intractable for large systems. This intractability is circumvented since the learning algorithm relies only on "local values" (individual elements) of the wavefunction for gradients and optimization, with an evaluation complexity scaling only quadratically with the number of qubits.

The original publication [14] shows for the AFH model only the low errors of the classical nearest neighbour correlations $\left\langle \sigma_i^z \sigma_{i+1}^z \right\rangle$. For the Transverse Field Ising model, the nearest neighbour quantum correlations $\left\langle \sigma_i^x \sigma_{i+1}^x \right\rangle$ are also shown to be learned with relatively low error. With regards to the QBM, low errors are required for all statistics: classical and quantum, single site expectation values and correlations, neighbouring $(i, i+1)$ and all other $(i, i+j)$ correlations.

6.1.1 Using symmetries

Some Hamiltonians are endowed with a ground-state symmetry. The RBM is readily adapted to incorporate these symmetries. This reduces the computational complexity of optimization (see section 6.4) by getting rid of unnecessary expressibility. For

[1] The computational base the most natural choice, however, other basis representations may be used. In fact, it is necessary to use multiple basis representations to do quantum state tomography with the neural-net quantum state [60].

[2] The factor of 2 in the product can generally be omitted, since normalization is omitted in general.

example, translation-invariant Hamiltonians share this property with their groundstates, so that $a_j = a$, $b_i = b$ and W consists of a weights matrix where its columns are necessarily cyclic permutations of other columns. For $\alpha = 1$ this implies

$$W = \begin{pmatrix} w_1 & w_2 & w_3 \\ w_2 & w_3 & w_1 \\ w_3 & w_1 & w_2 \end{pmatrix}. \tag{6.5}$$

This implication can be easily extended for $\alpha > 1$. In a similar fashion, symmetries w.r.t. spin flips and rotations may be imposed on the ansatz.

Besides the usage of symmetries to reduce the complexity of operations performed on the ansatz, the quality of the network parameters can also be increased experimentally by imposing a spin-sector restriction on the sampling procedure. The spin-sector refers to a subset of configurations in Hilbert space that meets a certain condition. In section 3.1 it is discussed that the ground state of the AFH model is a singlet state for bipartite lattices. This can be used to the advantage of the sampler by only addressing states with net-zero spin. This is done by using a sampling method with total spin $S = 0$ conservation. This effectively reduces the size of the Hilbert space of an N-particle system from 2^N to a system of size $\binom{N}{N/2}$, thereby reducing the sampling error considerably for small systems (see 6.5.2). Not only does this reduce sampling noise, it also relieves the variational ansatz of the task to learn $\Psi_{\mathcal{W}}(\mathbf{s}) = 0$ for configurations $\sum_i s_i \neq 0$ by allowing it to take on arbitrary values at these configurations instead. This focuses the expressive power of the Ansatz to a smaller state space, and relieves the MCMC sampler from local minima.

The drawback of enforcing this symmetry is that quantum spin correlations on the x and y axis are not learned correctly. For example, $\sigma_i^x \approx \frac{1}{N_{\text{samples}}} \sum_{s^k \in \text{MCMC}} \frac{\psi(F_i s^k)}{\psi(s^k)}$ and σ_{ij}^x, and $\psi(F_i s^k) = \psi(s^{k'})$ with $\sum_i s_i^{k'} \neq 0$ is not learned correctly.

6.1.2 Representational power

Neural networks are often regarded as universal function approximators. If the function at hand satisfies some natural and obvious smoothness conditions, the network is expected to approximate the function reasonably well, given enough free parameters to vary. In quantum machine learning, there is a special reason to tread carefully with the universal function interpretation.

The role of entanglement

The neural-net representation of wavefunctions relies on an RBM to represent a wavefunction along with its entanglement properties. However, the entanglement scaling law of a Hamiltonian is a limiting factor for the ability to express its groundstate wavefunction by an RBM [61]. In particular, an RBM where the hidden units are only connected to a subset of visible units[3] is shown to follow the area law. However, RBM's without this restriction are capable of representing volume law entanglement.

[3] In this case, the neural-net representation can be rewritten as a Matrix Product State (MPS). These MPS's, or tensor networks, form the basis of DMRG, which is known to describe area-law entanglement only.

While the ability to represent volume-law states with RBMs does not imply the necessary existence of a functioning learning scheme that actually finds the representation and its energy, the capability of the neural-net to capture volume-law entanglement represents a big advantage over tensor network approaches such as DMRG. The architecture is able to represent even massively entangled (large entanglement entropy) ground states, demonstrated in [61] for the Haldane-Shastry Hamiltonian

$$H = \sum_{j<k}^{N} \frac{1}{d_{jk}^2}(-\hat{\sigma}_i^x\hat{\sigma}_j^x - \hat{\sigma}_i^y\hat{\sigma}_j^y + \hat{\sigma}_i^z\hat{\sigma}_j^z) \tag{6.6}$$

of N spin-1/2 particles with $d_{ij} = (N/\pi)|\sin(\pi(j-k)/N)|$.

Shallowness

Another inevitable limitation of the RBM representation is its shallowness. An example of the ground state of a two-dimensional Hamiltonian with no RBM representation is given in Ref. [61]. Furthermore it is shown that adding hidden layers increases the representational power such that this pathological state does have a DBM (Deep Boltzmann Machine) representation. Although representability theorems give some indication (see e.g. [62]), the exact class of physical states with an RBM representation is still unknown [63]. The flexibility of the RBM is confirmed in section 6.5.

6.2 Variational Monte Carlo in an Artificial Neural-net context

This section will describe the method to optimize the weights of the RBM in order to learn the ground-state of the Hamiltonian at hand. The optimization relies fundamentally on calculating gradients of the energy w.r.t. the variational parameters using a Markov Chain Monte Carlo method, known as *Variational Monte Carlo.* The normalized expectation value of the Hamiltonian H w.r.t. some variational wavefunction $|\Psi_{\mathcal{W}}\rangle$ is given by the Rayleigh quotient

$$R[\Psi_{\mathcal{W}}] = \frac{\langle\Psi_{\mathcal{W}}|H|\Psi_{\mathcal{W}}\rangle}{\langle\Psi_{\mathcal{W}}|\Psi_{\mathcal{W}}\rangle}. \tag{6.7}$$

Minimization of the Rayleigh quotient is based on the famous Rayleigh-Ritz variational principle w.r.t. the exact ground state energy E_0 of H:

$$R[\Psi_{\mathcal{W}}] \geq E_0. \tag{6.8}$$

6.2.1 General VMC Error Analysis

The variational error is denoted $\epsilon = R[\Psi_{\mathcal{W}}] - E_0$. Now we expand the variational wavefunction in the complete basis set of eigenvectors $\{\psi_0, \psi_1, \ldots\}$ of H,

$$|\Psi_{\mathcal{W}}\rangle = \sum_i c_i |\psi_i\rangle . \tag{6.9}$$

It follows that the error ϵ in the variational energy can be written as

$$\epsilon = \sum_{i \neq 0} |c_i|^2 (E_i - E_0) \geq \Delta \sum_{i \neq 0} |c_i|^2, \tag{6.10}$$

where $\Delta = E_1 - E_0 > 0$ denotes the spectral gap. The normalization condition implies

$$\epsilon \geq \Delta (1 - |c_0|^2). \tag{6.11}$$

Note that the error in the variational energy scales quadratically with the error in the variational wavefunction. Interestingly, the lower bound of ϵ scales linearly with the spectral gap. In practice, VMC is aided by a large spectral gap.

6.2.2 Sign problem

As discussed in chapter 4, MCMC lends itself well to explore probability distributions from statistical physics, particularly the Boltzmann distribution. Classical distributions are associated with purely positive weights $\gamma(\mathbf{x}) \geq 0$. Sampling issues in Monte Carlo experiments arise when quantum mechanical distributions are given classical interpretations. In Diffusion and Green Function Monte Carlo the problem problem arises due to the Green function not being stochastic [64][65], which on its turn is caused by positive off-diagonal elements in the Hamiltonian H. A dirty fix is to carry the minus-sign over to the sampled observable $\hat{O}$ at interest. Consequently, this requires the sampling of ($\langle\langle \text{sign} \rangle\rangle_+$ and $\langle\langle \hat{O} \cdot \text{sign} \rangle\rangle_+$) which both go to zero exponentially with the system size. This scaling nullifies the intended benefit of Monte Carlo: the replacement of an exponential problem by a polynomial approximation.

The sign problem is basis-dependent. However, finding the transformation to relieve H of its positive off-diagonal elements is in general as hard as exact diagonalization. However, for some Hamiltonians efficient techniques have been developed. In section 3.1, it is noted that the the transformation for the XXX anti-ferromagnetic Heisenberg model is given by Marshall-Peierls sign rule: the sign structure of the ground state is known. In other words, a transformation to circumvent the sign problem is known. Sign problem free Hamiltonians (no positive off-diagonal elements) are known as *stoquastic* (portmanteau of *stochastic* and *quantum*).

Finding a stoquastic basis for a given Hamiltonian is an active field of research. For the XYZ model an efficient algorithm was recently developed to find the unitary transformation to make an n-qubit XYZ Hamiltonian (without single qubit terms) stoquastic in $\mathcal{O}(n^3)$ time [66].

For Variational Monte Carlo (*VMC*), however, sampling is done w.r.t. the absolute square of the wavefunction. Therefore, one would not expect the quality of VMC estimates to depend on the sign structure of the Hamiltonian or its ground state. Nevertheless, as will be shown in section 6.5.2, stoquasticity does change the quality of the results. It is argued in [67] that the generalization of sign-structure based on a small sample from Hilbert space is a problem on its own, independent of the sampling sign problem. In this paper, hardness of learning sign structure is quantified in terms of frustration for several 2D lattices. Moreover, there is no known theory that guarantees non-randomness of the sign structure of ground states of general spin Hamiltonians. Truly random sign structure cannot be learned.

6.2.3 Zero-variance property

For increasingly better approximations $|\Psi_{\mathcal{W}}\rangle$ of the true ground state Φ_0, the variation between Markov chain samples of the local energy is reduced due to the *zero-variance property.* A perfectly optimized variational wavefunction (excluding expressibility errors) is an exact eigenfunction of H, so that

$$\frac{\langle s|H|\Psi_{\mathcal{W}}\rangle}{\langle s|\Psi_{\mathcal{W}}\rangle} = E\frac{\langle s|\Psi\rangle}{\langle s|\Psi\rangle} = E. \tag{6.12}$$

Although an idealised approximation, the variance of the local energy over a Markov chain is in fact reduced in when the variational parameter set represents the ground state (or any other eigenstate) with sufficient accuracy. Inconveniently, this property does not apply to observables that do not commute with H. Reduction of the variance of the local energy can be used as an alternative to Rayleigh quotient minimization in order to optimize a variational wavefunction. This is also known as the *sigma variational principle.*

6.2.4 Pathological ground states

For ill-behaved Markov chains the sampling results are capricious (see section 4.3.2). Chains might get stuck on configurations since wavefunctions are not necessarily smooth distributions. This becomes a tangible problem when only local moves are allowed in the chain. A simple demonstration of this is the 1-local (Markov moves with Hamming distance 1) sampling of the ground state of the anti-ferromagnetic Heisenberg system consisting of 2 spin-$\frac{1}{2}$ particles. The ground state of this system is the singlet state, denoted in the computational basis by $\Phi_0 = \frac{1}{\sqrt{2}}(0, 1, -1, 0)^T$. Reducible 1-local samplers do not exist for this system, since neither the $(\downarrow\uparrow)$ state, nor $(\uparrow\downarrow)$ can be escaped by flipping one spin as it has to move through one of the ferromagnetic “bottleneck” configurations with transition probability 0. The two configurations are said to be on two different *islands* in spin-configuration space.

This problem can be alleviated by using more advanced sampling methods. The most straightforward solution to this problem is allowing steps with larger Hamming distance. This approach can be further improved by parallelization. Furthermore, one would expect the number of islands separated by bottlenecks of small Hamming distance to scale inversely with dimensionality (having an island of size N in a system of size N becomes unlikely for large N). The validity of this assumption is supported by the results in section 6.5.2.

6.2.5 Sampling of RBM Gradients

A multitude of different routines exist in order to optimize equation 6.7, each using the gradients in a slightly different manner. The energy gradients w.r.t a variational wavefunction are denoted as

$$f_k = -\frac{\partial E(\mathcal{W})}{\partial \mathcal{W}_k} = -\frac{\partial}{\partial \mathcal{W}_k}\frac{\langle\Psi_{\mathcal{W}}|H|\Psi_{\mathcal{W}}\rangle}{\langle\Psi_{\mathcal{W}}|\Psi_{\mathcal{W}}\rangle}. \tag{6.13}$$

A Taylor expansion of the variational wavefunction in its variational parameters yields (for the k'th component of $\mathcal{W}$):

$$\Psi_{\mathcal{W}+\delta\mathcal{W}_k}(\mathbf{s}) = \Psi_{\mathcal{W}}(\mathbf{s}) + \delta\mathcal{W}_k\frac{\partial\Psi_{\mathcal{W}}(\mathbf{s})}{\partial\mathcal{W}_k} + \mathcal{O}(\partial\mathcal{W}_k^2). \tag{6.14}$$

Assuming $\Psi_{\mathcal{W}}(\mathbf{s}) \neq 0$, equation 6.14 can be rewritten in terms of local operators O_k[4] , corresponding to the variational parameter $\mathcal{W}_k$ defined by diagonal matrix elements[5]

$$\langle \mathbf{s}|O_k|\mathbf{s}'\rangle = O_k(\mathbf{s})\delta_{\mathbf{s},\mathbf{s}'} \tag{6.15}$$

$$O_k(\mathbf{s}) = \frac{\partial \ln \Psi_{\mathcal{W}}(\mathbf{s})}{\partial \mathcal{W}_k} = \frac{1}{\Psi_{\mathcal{W}}(\mathbf{s})}\frac{\partial \Psi_{\mathcal{W}}(\mathbf{s})}{\partial \mathcal{W}_k}, \tag{6.16}$$

so that

$$\Psi_{\mathcal{W}+\delta\mathcal{W}_k} = (1+\delta\mathcal{W}_k O_k)\Psi_{\mathcal{W}}. \tag{6.17}$$

Considering an explicitly normalized wavefunction,

$$|v_{0,\mathcal{W}}\rangle \equiv \frac{\Psi_{\mathcal{W}}}{||\Psi_{\mathcal{W}}||}, \tag{6.18}$$

we define for every variational parameter the perturbed state

$$|v_{k,\mathcal{W}}\rangle \equiv (O_k - \overline{O}_k)\,|v_{0,\mathcal{W}}\rangle\,, \tag{6.19}$$

where

$$\overline{O}_k = \langle v_{0,\mathcal{W}}|O_k|v_{0,\mathcal{W}}\rangle = \frac{\langle \Psi_{\mathcal{W}}|O_k|\Psi_{\mathcal{W}}\rangle}{\langle \Psi_{\mathcal{W}}|\Psi_{\mathcal{W}}\rangle}. \tag{6.20}$$

The normalized version of equation 6.17 can therefore be written as

$$|v_{0,\mathcal{W}+\delta\mathcal{W}_k}\rangle \propto |v_{0,\mathcal{W}}\rangle + \delta\mathcal{W}_k\,|v_{k,\mathcal{W}}\rangle + \mathcal{O}(\delta_k^2). \tag{6.21}$$

The derivative of the energy can thus be worked out using equation 6.21 and by definition of the infinitesimal difference:

$$\begin{aligned}
\frac{\partial E(\mathcal{W})}{\partial \mathcal{W}_k} &= \lim_{\delta\mathcal{W}_k\to 0}\frac{\langle v_{0,\mathcal{W}+\delta\mathcal{W}_k}|H|v_{0,\mathcal{W}+\delta\mathcal{W}_k}\rangle - \langle v_{0,\mathcal{W}}|H|v_{0,\mathcal{W}}\rangle}{\delta\mathcal{W}_k} && (6.22)\\
&= \langle v_{k,\mathcal{W}}|H|v_{0,\mathcal{W}}\rangle + \langle v_{0,\mathcal{W}}|H|v_{k,\mathcal{W}}\rangle && (6.23)\\
&= 2\,\mathrm{Re}\left[\frac{\langle \Psi_{\mathcal{W}}|H(O_k-\overline{O}_k)|\Psi_{\mathcal{W}}\rangle}{\langle \Psi_{\mathcal{W}}|\Psi_{\mathcal{W}}\rangle}\right] && (6.24)\\
&= 2\,\mathrm{Re}\left[\frac{\sum_{\mathbf{s}}\langle \Psi_{\mathcal{W}}|H|\mathbf{s}\rangle\,\langle \mathbf{s}|\,(O_k-\overline{O}_k)\,|\Psi_{\mathcal{W}}\rangle}{\sum_{\mathbf{s}}\langle \Psi_{\mathcal{W}}|\mathbf{s}\rangle\,\langle \mathbf{s}|\Psi_{\mathcal{W}}\rangle}\right]. && (6.25)
\end{aligned}$$

Components $\langle \mathbf{s}|H|\Psi_{\mathcal{W}}\rangle \,/\, \langle \mathbf{s}|\Psi_{\mathcal{W}}\rangle$ are interpreted as the *local energy* E_{loc} of the system at state $\mathbf{s}$ or, equivalently, the $\mathbf{s}$-component of the eigenvalue-equation $H\,|v_{0,\mathcal{W}}\rangle = E\,|v_{0,\mathcal{W}}\rangle$. This quantity is evaluated during sampling to estimate the energy. Equation 6.25 can now be written as:

$$\begin{aligned}
\frac{\partial E(\mathcal{W})}{\partial \mathcal{W}_k} &= 2\,\mathrm{Re}\left[\frac{\sum_{\mathbf{s}} E^*_{\mathrm{loc}}(\mathbf{s})\,\langle \Psi_{\mathcal{W}}|\mathbf{s}\rangle\,\langle \mathbf{s}|\,(O_k-\overline{O}_k)\,|\Psi_{\mathcal{W}}\rangle}{\sum_{\mathbf{s}}|\,\langle \Psi_{\mathcal{W}}|\mathbf{s}\rangle\,|^2}\right] && (6.26)\\
&= 2\,\mathrm{Re}\left[\frac{\sum_{\mathbf{s}} E^*_{\mathrm{loc}}(\mathbf{s})(O_k-\overline{O}_k)|\Psi_{\mathcal{W}}(\mathbf{s})|^2}{\sum_{\mathbf{s}}|\Psi_{\mathcal{W}}(\mathbf{s})|^2}\right]. && (6.27)
\end{aligned}$$

[4] Strictly speaking O_k depends on $\mathcal{W}$, but this dependency is implied to avoid cluttered notation.

[5] Note that the RBM ansatz is a complex exponential e^z. Consequently, it satisfies the Cauchy-Riemann equations and is holomorphic. This allows complex differentiation of $\Psi_{\mathcal{W}}$ w.r.t. a parameter $\mathcal{W}_k \in \mathbb{C}$, even though $E(\mathcal{W})$ is real and not holomorphic.

The energy derivatives are expectation values, evaluated by a Monte Carlo process on the state space under the distribution $|\Psi_{\mathcal{W}}|^2$. With respect to a two-local quantum Hamiltonian (equation 5.29), the local energy for spin-configuration $\mathbf{s}$ can be denoted

$$E_{\text{loc}}(\mathbf{s}) = \sum_{i=1,i<j}^{n} \frac{\Psi_{\mathcal{W}}(F_i F_j \mathbf{s})}{\Psi_{\mathcal{W}}(\mathbf{s})}(w_{ij}^x - w_{ij}^y s_i s_j) + \sum_{i=1}^{n} \frac{\Psi_{\mathcal{W}}(F_i \mathbf{s})}{\Psi_{\mathcal{W}}(\mathbf{s})}(w_i^x + i s_i w_i^y) + \sum_{i=1,i<j}^{n} w_{ij}^z s_i s_j + \sum_{i=1}^{n} s_i w_i^z. \tag{6.28}$$

Equivalently, other spin observables are determined locally as:

$$\sigma_i^x(\mathbf{s}) = \frac{\Psi_{\mathcal{W}}(F_i \mathbf{s})}{\Psi_{\mathcal{W}}(\mathbf{s})} \tag{6.29}$$

$$\sigma_i^y(\mathbf{s}) = i s_i \sigma_i^x(\mathbf{s}) \tag{6.30}$$

$$\sigma_i^z(\mathbf{s}) = s_i. \tag{6.31}$$

The explicit local gradients of the RBM are

$$O_{a_i}(\mathbf{s}) = s_i \tag{6.32}$$

$$O_{b_j}(\mathbf{s}) = \tanh \theta_j(\mathbf{s}) \tag{6.33}$$

$$O_{W_{ij}}(\mathbf{s}) = s_i \tanh \theta_j(\mathbf{s}) = O_{a_i} O_{b_j}, \tag{6.34}$$

where $\theta_j(\mathbf{s}) = b_j + \sum_i W_{ij} s_i$.

Subsequently, stochastic estimates of the gradients are employed in the usual descent scheme with some small learning parameter η:

$$\mathcal{W}_k' = \mathcal{W}_k + \delta \mathcal{W}_k \tag{6.35}$$

$$\delta \mathcal{W}_k = -\eta \frac{\partial E(\mathcal{W})}{\partial \mathcal{W}_k}. \tag{6.36}$$

6.3 Stochastic Reconfiguration

Although many gradient descent algorithms are available, Stochastic Reconfiguration (*SR*) [65] is the most popular choice for updating the weights of the neural-net ansatz (e.g. [68][14][60]). In literature, SR is often referred to as similar [60] or equal [69] to the natural gradient method. While natural gradient follows a geometric approach, SR comes to the same conclusion via a more physical route. In order to shed some light on how Natural Gradient is equivalent to Stochastic Reconfiguration, the two are first explained separately. In section 6.3.3 they are unified.

6.3.1 Stochastic Reconfiguration in detail

Approximating the energy function by Taylor expansion in a small region around some set of parameters $\mathbf{w}$ (with $\partial E(\mathbf{w})/\partial w_k \neq 0$), the energy $E(\mathbf{w}')$ can be approximated as:

$$E(\mathbf{w}') = E(\mathbf{w}) + \sum_k \frac{\partial E(\mathbf{w})}{\partial w_k} \delta w_k + \mathcal{O}(\eta^2) \tag{6.37}$$

$$= E(\mathbf{w}) - \eta \sum_k \Big(\frac{\partial E(\mathbf{w})}{\partial w_k}\Big)^2 + \mathcal{O}(\eta^2), \tag{6.38}$$

so that the energy corresponding to the new parameters $\mathbf{w}'$ is always smaller:

$$\Delta E = E(\mathbf{w}') - E(\mathbf{w}) = -\eta \sum_k \left(\frac{\partial E(\mathbf{w})}{\partial w_k} \right)^2 < 0. \tag{6.39}$$

The step magnitude can be controlled by normalization:

$$\delta s^2 \equiv \sum_k \delta w_k^2. \tag{6.40}$$

This constraint can be incorporated by minimization of the Lagrangean with Lagrange multiplier μ:

$$\Delta E + \mu \delta s^2 = \sum_k \frac{\partial E(\mathbf{w})}{\partial w_k} \delta w_k + \mu \delta w_k^2, \tag{6.41}$$

yielding the condition

$$\delta w_k = \frac{\partial E(\mathbf{w})}{\partial w_k} (2\mu)^{-1}. \tag{6.42}$$

The learning rate is thus constrained by the normalization condition as $\eta = (2\mu)^{-1}$. However, equation 6.40 implicitly assumes a Euclidean metric on the parameter space. Overshooting is only prevented as long as this metric describes the energy landscape accurately. A different metric is chosen to reflect a restriction on the magnitude of change in the wavefunction, instead of its parameters. A measure that takes into account the proximity between the initial and updated wavefunctions is:

$$\delta s^2 = \min_\theta || \exp(-i\theta) v_{0,\mathbf{w}+w_k} - v_{0,\mathbf{w}} ||^2. \tag{6.43}$$

Minimization over θ is done to ignore a global phase difference. Substitution of the normalized wavefunction that has been updated in all parameters w_k yields

$$\delta s^2 = \sum_{k,k'} \langle v_{k,w} | v_{k',w} \rangle \, \delta w_k \delta w_{k'} + \mathcal{O}(|\delta w|^2). \tag{6.44}$$

The minimization of the Lagrangean with this new metric yields (compare with equation 6.42)

$$\sum_{k'} S_{kk'} \delta w_k = \frac{\partial E(\mathbf{w})}{\partial w_k} (2\mu)^{-1}. \tag{6.45}$$

This linear system of equations is solved by

$$\delta \mathbf{w} = S^{-1} \nabla E(\mathbf{w}) (2\mu)^{-1}, \tag{6.46}$$

where the learning rate $\eta = (2\mu)^{-1}$ should be set small enough in order to satisfy the same convergence property as determined in equation 6.39.

The matrix S, where $S_{kk'} = \langle v_{k,w} | v_{k',w} \rangle$ can be identified as the matrix defining the metric (and is of similar form as eq. 6.53). Writing S explicitly:

$$\begin{aligned}
S_{kk'} = \langle v_{k,w} | v_{k',w} \rangle &= \langle v_{0,\mathbf{w}} | (O_k - \overline{O}_k)^* (O_{k'} - \overline{O}_{k'}) | v_{0,\mathbf{w}} \rangle && (6.47) \\
&= \langle v_{0,\mathbf{w}} | O_k^* O_{k'} - O_k^* \overline{O}_{k'} - \overline{O}_k^* O_{k'} + \overline{O}_k^* \overline{O}_{k'} | v_{0,\mathbf{w}} \rangle && (6.48) \\
&= \langle O_k^* O_{k'} \rangle + \langle O_k^* \rangle \langle O_{k'} \rangle - \langle O_{k'} \rangle \langle O_k^* \rangle - \langle O_k^* \rangle \langle O_{k'} \rangle && (6.49) \\
&= \langle O_k^* O_{k'} \rangle - \langle O_k^* \rangle \langle O_{k'} \rangle && (6.50)
\end{aligned}$$

Finally, the parameters are updated through

$$\mathbf{w}' = \mathbf{w} - \eta S^{-1} \nabla E. \tag{6.51}$$

Note that by assuming a diagonal metric, conventional gradient descent is retrieved immediately. Numerical problems due to an ill-conditioned S are alleviated by shifting the diagonal of S, effectively making the gradient more Euclidean.

6.3.2 Natural Gradient

The natural gradient provides the steepest direction of descent by geometry arguments [70]. Let $S = \{\mathbf{w} \in \mathbb{R}^n\}$ be the parameter space on which $L(\mathbf{w})$ is defined. When S is Euclidean with an orthonormal coordinate system $\mathbf{w}$, the squared length of a small incremental vector $d\mathbf{w}$ connecting $\mathbf{w}$ and $\mathbf{w} + d\mathbf{w}$ is given by

$$|d\mathbf{w}|^2 = \sum_{i=1}^{n} (dw_i)^2, \tag{6.52}$$

where dw_i are components of $d\mathbf{w}$. When the coordinate system is nonorthonormal, the squared length is given b the quadratic form

$$|d\mathbf{w}|^2 = \sum_{i,j} g_{ij} dw_i dw_j. \tag{6.53}$$

When S is a curved manifold, there is no orthonormal linear coordinates, and the length of $d\mathbf{w}$ is always written as in the last equation. Such space is Riemannian. $G = (g_{ij})$ is the Riemannian metric tensor. In the Euclidean case it reduces to

$$g_{ij}(\mathbf{w}) = \delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j \end{cases} \tag{6.54}$$

The steepest descent direction of a function $L(\mathbf{w})$ is defined as the vector $d\mathbf{w}$ that minimizes $L(\mathbf{w} + d\mathbf{w})$ where $|d\mathbf{w}| = \epsilon^2$ is fixed.

Steepest descent in Riemannian space

The steepest descent direction of $L(\mathbf{w})$ in Riemannian space is given by[6]:

$$-\tilde{\nabla} L(\mathbf{w}) = -G^{-1}(\mathbf{w}) \nabla L(\mathbf{w}), \tag{6.55}$$

where $G^{-1} = (g^{ij})$ is the inverse of the metric $G = (g_{ij})$, and ∇L is the conventional gradient

$$\nabla L(\mathbf{w}) = \Big(\frac{\partial}{\partial w_1} L(\mathbf{w}), \dots, \frac{\partial}{\partial w_n} L(\mathbf{w}) \Big)^T. \tag{6.56}$$

In the case of a Euclidean parameter space S with an orthonormal coordinate system, the metric is simply the identity $G^{-1} = G = I$.

The learning rule becomes

$$\mathbf{w}' = \mathbf{w} - \eta G^{-1}(\mathbf{w}) \nabla l(z, \mathbf{w}). \tag{6.57}$$

[6] A proof is included in Appendix A.

The Quantum Geometric Tensor

The proper metric to accurately determine the proximity of wavefunctions was first formalized in [71]. The metric is formed by consideration of quantum *rays*, as opposed to quantum *states*. Rays have the property that two states are described by the same points (on the manifold of rays) if they differ only by a phase, meaning two normalized vectors Ψ' and Ψ of some Hilbert space defined as

$$\Psi' = \exp(i\theta)\Psi \tag{6.58}$$

describe the same ray. Consider a family of normalized vectors $\{\Psi_{\mathbf{w}}\}$ in this space which depend smoothly on an n-dimensional parameter $\mathbf{w} = (w_1, \cdots, w_n) \in \mathbb{R}$. The distance measure is developed as usual by the inner product on this Hilbert space

$$||\Psi_{\mathbf{w}+d\mathbf{w}} - \Psi_{\mathbf{w}}||^2 = \langle \Psi_{\mathbf{w}+d\mathbf{w}} - \Psi_{\mathbf{w}} | \Psi_{\mathbf{w}+d\mathbf{w}} - \Psi_{\mathbf{w}})\rangle \tag{6.59}$$

$$= \sum_{i,j} \left\langle \frac{\partial \Psi_{\mathbf{w}}}{\partial w_i} \middle| \frac{\partial \Psi_{\mathbf{w}}}{\partial w_j} \right\rangle dw_i dw_j, \tag{6.60}$$

where the last equality is a second order approximation. By separation of the real and imaginary parts of the Hermitian inner product,

$$\left\langle \frac{\partial \Psi_{\mathbf{w}}}{\partial w_i} \middle| \frac{\partial \Psi_{\mathbf{w}}}{\partial w_j} \right\rangle = \gamma_{ij} + i\sigma_{ij}, \tag{6.61}$$

the distance measure reduces, due to the antisymmetry of the imaginary part, to the simpler form

$$||\Psi_{\mathbf{w}+d\mathbf{w}} - \Psi_{\mathbf{w}}||^2 = \sum_{i,j} \gamma_{ij} dw_i dw_j. \tag{6.62}$$

However, identifying γ as the metric tensor on the manifold of rays is wrong due to its dependence on phase. Evaluating γ under a phase shift yields:

$$\gamma'_{ij} = \mathrm{Re}\left[\left\langle \frac{\partial \Psi'_{\mathbf{w}}}{\partial w_i} \middle| \frac{\partial \Psi'_{\mathbf{w}}}{\partial w_j} \right\rangle \right] \tag{6.63}$$

$$= \gamma_{ij} + \beta_i \frac{\partial \theta}{\partial w_j} + \beta_j \frac{\partial \theta}{\partial w_j} + \frac{\partial \theta}{\partial w_i}\frac{\partial \theta}{\partial w_j}, \tag{6.64}$$

with $\beta_i = i\left\langle \Psi_{\mathbf{w}} \middle| \frac{\partial \Psi_{\mathbf{w}}}{\partial w_i} \right\rangle$[7]. This metric differs from γ_{ij}, ergo the metric is gauge-variant. This can be fixed by noting how the the β_i terms transform:

$$\beta_i \to \beta'_i = \beta_i + \frac{\partial \theta}{\partial w_j}, \tag{6.65}$$

leading to the improved metric

$$G_{ij} = \gamma_{ij} - \beta_i \beta_j, \tag{6.66}$$

for which $G'_{ij} = G_{ij}$ by design.

[7] Note that β_i is real since $2\,\mathrm{Re}\left[\left\langle \Psi_{\mathbf{w}} \middle| \frac{\partial \Psi_{\mathbf{w}}}{\partial w_i} \right\rangle\right] = \left\langle \Psi_{\mathbf{w}} \middle| \frac{\partial \Psi_{\mathbf{w}}}{\partial w_i} \right\rangle + \left\langle \frac{\partial \Psi_{\mathbf{w}}}{\partial w_i} \middle| \Psi_{\mathbf{w}} \right\rangle = \frac{\partial}{\partial w_i}\langle \Psi_{\mathbf{w}} | \Psi_{\mathbf{w}} \rangle = 0$.

6.3.3 Equivalence with Quantum Geometric Natural Gradient

The choice of metric by Sorella [72] (equation 6.43) for the SR method leads to an update equation for the variational parameters which is reminiscent of the update equation that emerges in the natural gradient formalism. As introduced above, the Quantum Geometric Tensor is the (gauge invariant) metric for measuring "physical" distances between rays in projective Hilbert space, also referred to as the natural Fubini-Study metric on quantum space [73]. Rewriting this natural metric (equation 6.66)

$$
\begin{aligned}
G_{kk'} &\equiv \left\langle \partial_{w_k} \Psi_{\mathbf{w}} \middle| \partial_{w_{k'}} \Psi_{\mathbf{w}} \right\rangle - \left\langle \partial_{w_k} \Psi_{\mathbf{w}} | \Psi_{\mathbf{w}} \right\rangle \left\langle \Psi_{\mathbf{w}} \middle| \partial_{w_{k'}} \Psi_{\mathbf{w}} \right\rangle & (6.67)\\
&= \left\langle \Psi_{\mathbf{w}} | O_k^* O_{k'} | \Psi_{\mathbf{w}} \right\rangle - \left\langle \Psi_{\mathbf{w}} | O_k^* | \Psi_{\mathbf{w}} \right\rangle \left\langle \Psi_{\mathbf{w}} | O_{k'} | \Psi_{\mathbf{w}} \right\rangle & (6.68)\\
&= \langle O_k^* O_{k'} \rangle - \langle O_k^* \rangle \langle O_{k'} \rangle & (6.69)\\
&= S_{kk'} & (6.70)
\end{aligned}
$$

yields the S-matrix used by SR (equation 6.50). By substitution of the S-metric established above, the natural gradient in Riemannian space (equation 6.57) becomes

$$
\begin{aligned}
\mathbf{w}' &= \mathbf{w} - \eta \tilde{\nabla} E(\mathbf{w}) & (6.71)\\
&= \mathbf{w} - \eta G^{-1} \nabla E(\mathbf{w}) & (6.72)\\
&= \mathbf{w} - \eta S^{-1} \nabla E(\mathbf{w}), & (6.73)
\end{aligned}
$$

which is equivalent to the update rule as proposed with stochastic reconfiguration (equation 6.51). Therefore, SR can be interpreted as a Natural Gradient method on some Hilbert space endowed with a Fubini-Study metric. Originally, SR uses a different approach to ensure proper (non-Euclidean) normalization of the gradient by usage of a phase-independent geometry. This consideration is also incorporated by the Quantum Geometric Tensor. In SR it leads to a linear system of equations described by the S-matrix, which is equal to the Quantum Geometric Tensor which follows from geometric arguments more succinctly.

6.4 Complexity

In order to analyze the computational bottleneck of the ANN learning algorithm, the complexity of the sampling and optimization process should be analyzed. In this discussion $N_p \equiv |\mathcal{W}|$ denotes the number of parameters.

6.4.1 Sampling

The Metropolis-Hastings sampler uses fractions of wavefunctions $|\Psi(\mathbf{s}', \mathcal{W})/\Psi(\mathbf{s}, \mathcal{W})|^2$ in order to sample gradients of the Rayleigh quotient and consequently update parameters $\mathcal{W} \to \mathcal{W}'$. A lookup-table $\theta = b + W^T s$ can be used to save time. In order to gather one local energy sample in an N_v-spin system, one needs $\mathcal{O}(N_v^2 N_p) = \mathcal{O}(\alpha N_v^3)$ operations, due to the necessary evaluations of equation 6.28. For nearest-neighbors models with uniform weights (e.g. the XXX AFH model), the complexity is reduced to $\mathcal{O}(\alpha N_v^2)$.

6.4.2 Optimization

Storage of the S-matrix can be prevented by making use of its product structure in combination with an iterative solver. This leads to two modi operandi:

- **Naive inversion**. Build the S-matrix explicitly using $\mathcal{O}(N_p^2 N_{\text{MC}})$ time and $\mathcal{O}(N_p^2)$ space. Subsequently, invert it exactly using $\mathcal{O}(N_p^3)$ operations to get the updated wavefunction parameters.

- **Matrix-free inversion**. Use a conjugate gradient (CG) method, which requires at most $\mathcal{O}(N_p)$ steps to converge, each step requiring a matrix-vector product $S \cdot \mathbf{z}$ on some test vectors $\mathbf{z}$. The computational cost of the product can be reduced (and the storage of S prevented) by not forming S explicitly, but rather using $S\mathbf{z} = (O^* - \overline{O}^*) \cdot ((O - \overline{O}) \cdot \mathbf{z}) = (O^* - \overline{O}^*) \cdot \mathbf{z}' = \mathbf{z}''$. This matrix-vector product requires $\mathcal{O}(N_p N_{\text{MC}})$ operations. Since CG requires at most $\mathcal{O}(N_p)$ of such matrix-vector products, the final complexity is of dominant order $\mathcal{O}(N_p^2 N_{\text{MC}})$. In practice, the number of CG steps needed to solve the linear system with reasonable precision is much smaller than N_p. In terms of computation time, this makes CG a viable alternative for naive optimization already for $N_{\text{MC}} \approx N_p$. In terms of storage, usage of this method becomes inevitable for very large numbers of variational parameters where $N_{\text{MC}} < N_p$ and $S \sim \mathcal{O}(N_p^2)$ too large to store. In terms of implementation, the copying of large arrays to and from CPUs can be prevented by performing the matrix-vector products in parallel, and collecting the resulting vectors in shared memory.

6.5 Experiments

The analysis of the performance of the Neural-net Quantum State is a complicated matter with many knobs to turn. Tuning these parameters and determining their correlations is an art. However, the chosen experiments all aim to answer the question: "Does the polynomially scaling (in time and space w.r.t. the number of spins) ANN model find the correct ground state energy and statistics?". Sometimes, inevitably arbitrary decisions have to be made in order to fairly compare results. For example, comparison of differently sized spin systems is done on the basis of a linear scaling of samples. All the relevant parameters are:

- Number of spins N.

- Weights of the model at interest.

- The number of workers/CPUs, each one corresponding to an independent Markov chain.

- The number of proposed flips by the sampler N_{samples}. *Single*: only flips with Hamming distance 1. *Single/double*: 50% probability of proposing a single flip or a double flip of Hamming distance 2.

- The (initial) learning rate η. This is a real scalar that determines the step size in the first optimization step. Default value $\eta = 0.05$.

- Decay factor γ. At iteration t, the learning rate is $\gamma^t \eta$. Default value $\gamma = 1$.

- Regularization λ. Improve the condition number of the S-matrix. Default value $\lambda = 0.05$.
- Hidden unit density $\alpha = N_h/N_v$.
- Iterations/repetitions/optimization steps.
- Samples per iteration N_{MC}.
- MCMC steps per sample (samples per *sweep*) to increase sample independence. Default value equal to number of spins N.
- Thermalization steps. Default value set to $0.1 \cdot N_{\mathrm{MC}}$.
- Conjugate gradient error tolerance ϵ. Default value 10^{-3}.

6.5.1 The misleading accuracy of sampling

Although rarely used in practice, it is still a good sanity check to determine the quality of the complete variational wavefunction, denoted $\Psi_{\mathcal{W}}$ in the following analysis. Some surprising results are collected for the XXX anti-ferromagnetic Heisenberg model with naive (not $S = 0$ sector) sampling.

In figure 6.1, a discrepancy can be noticed between the mean local energy $\overline{E_{\mathrm{loc}}}$ and the Rayleigh Quotient $R[\Psi_{\mathcal{W}}] = \frac{\langle \Psi_{\mathcal{W}} | H | \Psi(\mathcal{W}) \rangle}{\langle \Psi_{\mathcal{W}} | \Psi_{\mathcal{W}} \rangle}$. The mean local energy $\overline{E_{\mathrm{loc}}}$ is defined as the sampling average over the Markov chain of the last optimization step. The mean local energies gathered by the double flip sampler are clustered in above- and below-diagonal measurements that shift to and from either side upon reinitialization and rethermalization (see figure 6.3).

Why it is surprising

While the measurements above the diagonal are expected due to sampling noise, the cluster of measurements of the double flip sampler under the diagonal is surprising. The sampling noise could, while very unlikely, decrease the error in the ground state energy, however the large cluster far beneath the diagonal is not explained away by this. These measurements demonstrate that the sampling average w.r.t. the $|\Psi(\mathcal{W})|^2$ distribution could give a better estimate of the ground state energy than the exact average over the same distribution. The shifts upon reinitalisation indicate that the quality of the measurements changes due to differences in the Markov chain rather than the quality of the variational wavefunction (figure 6.3). This is furthermore confirmed by figure 6.2.

Why it happens

The *raison d'être* for the below-diagonal measurements becomes clear when "flooring" the optimized wavefunction by setting elements $|\Psi_{\mathcal{W}}(s^k)|^2 < 10^{-4}$ to 0. The variational ansatz approximates the wavefunction with great precision, but zero-elements of the true wavefunction settle to absolute values of around 10^{-4} in the variational wavefunction. The effect of all these small (but non-zero) absolute value elements adds up in the error of the Rayleigh quotient. However, these illegitimate states (corresponding to wavefunction elements that should have absolute value zero) are not sampled in the cluster underneath the diagonal. When a Markov chain does not encounter such low probability states, it effectively samples from the flattened wavefunction. Taking into account this fact figure 6.1 is redrawn in figure 6.4 (see also figure 6.8), showing that all measurements end up above the diagonal as expected. This explains the measurements under the diagonal.

However, when a finite Markov chain does encounter a state s^k with relatively low absolute value (e.g. $|\Psi_{\mathcal{W}}(s^k)|^2 \approx 10^{-2}$), it could move to such neighbouring illegitimate states with non-negligible probability (about 0.01) by the acceptance criterion $\min\left(1, \left|\frac{\Psi_{\mathcal{W}}(F_{ij}s^k)}{\Psi_{\mathcal{W}}(s^k)}\right|^2\right)$, where F_{ij} again denotes the flipping operator acting on site i and j (see also section 3.1). When the sampler arrives at these illegitimate states, unstable samples of the local energy are collected[8], the computation of which involves fractions over connected states $\frac{\Psi_{\mathcal{W}}(F_{ij}s^k)}{\Psi_{\mathcal{W}}(s^k)}$. Investigation of the Markov chains belonging to the measurements of the double flips sampler reveals that the above-diagonal cluster measurements all encountered an illegitimate state, while all below-diagonal cluster measurements did not. Omitting the illegitimate states and their local energy samples from their respective Markov chains results in their measurements moving below the diagonal as well (figure 6.6). This explains the measurements of the double flip sampler above the diagonal in figure 6.1.

Initially, there is no reason to expect that the single flip sampler would show different behaviour, however figure 6.1 shows no clear above- and below-diagonal clustering. The reason for this is the poor ergodicity of the Markov chains with single flip transitions and consequently also the poorer network parameter solutions found. Flooring the variational wavefunctions found by single flip sampling results in slightly lower errors w.r.t. the true wavefunction, but nowhere near the low errors of the double flip sampler (see figure 6.5 and 6.7). Correcting for illegitimate states encountered by the single flip sampler has negligible effect, since the errors caused by this are of lesser order than the error due to the poorly optimised network parameters.

The errors of both the "fixed" (floored) Rayleigh quotients and mean local energies (illegitimate samples removed) are shown in figure 6.9, showing almost all errors are moved above-diagonal.

Conclusion

The node structure of the ground state wavefunction for this particular model is known, and the encountered problems could easily be solved by using a $S = 0$-sector sampler. However, this is not the case for any 2-local qubit Hamiltonians. These results therefore show that learning -and sampling from- wavefunctions with nodes can lead to large errors, effectively caused by the failure of the probability distribution $|\Psi_{\mathcal{W}}|^2$ to meet the MCMC smoothness requirement.

[8] Note that the absolute values of local wavefunctions at illegitimate states $\Psi_{\mathcal{W}}(s^k_{\text{illeg}})$ and $\Psi_{\mathcal{W}}(F_{ij}s^k_{\text{illeg}})$ are of the same order of magnitude, thus introducing large errors in the MCMC mean over local observable samples.

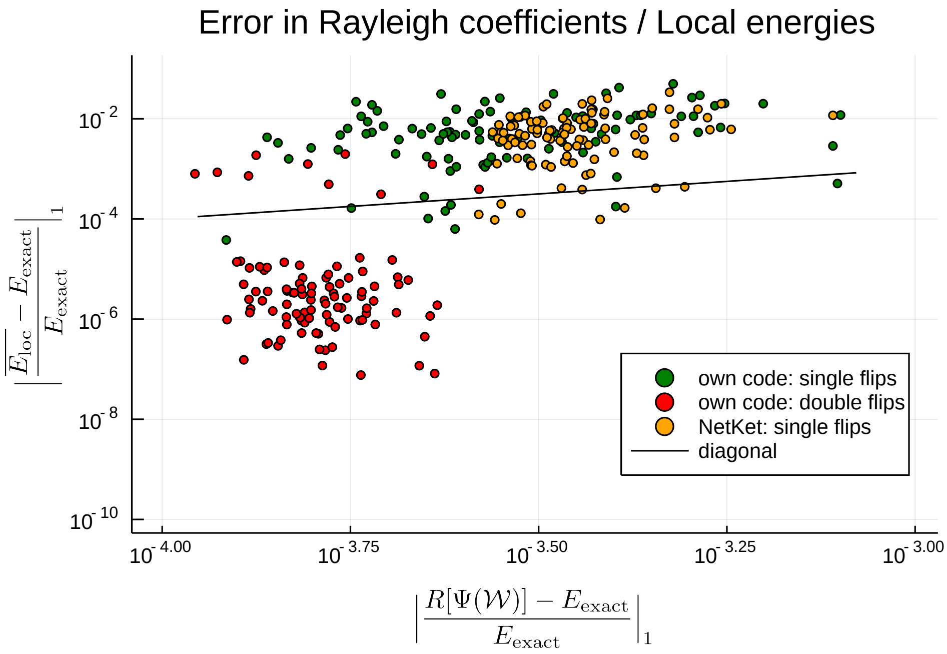


Figure 6.1: Comparison between own software written in Julia and open-source many-body quantum system software package NetKet [74]. 100 runs are performed in each modus on the stoquastic AFH model with of a chain of qubits of length $L = 6$ with periodic boundary conditions and $\alpha = 3$. The figure also confirms the proper workings of own software. The discrepancy between the mean local energy and the Rayleigh quotient is visible when double spin flips are used to propose new states in the local Metropolis sampler in own software. The double flip sampler picks (with probability 0.5) 1 or 2 random sites of the chain to flip. The flip is accepted with the Metropolis-Hastings probability $\min(1, |\frac{\Psi_{\mathcal{W}}(F_{ij}s^k)}{\Psi_{\mathcal{W}}(s^k)}|^2)$. The total optimization algorithm consists of 800 steps, with 400 samples per step. The learning rate at optimization step t is defined as $\eta = \gamma^t \eta_{\text{init}}$ with $\eta_{\text{init}} = 0.05$ and the decay factor $\gamma = 0.995$. The single flip sampler encounters on average 2.93 unique states in the last Markov chain before termination, while the double flip sampler encounters 13.75 unique states. The exact wavefunction has $\binom{n}{n/2} = 20$ non-zero absolute value elements corresponding to the anti-ferromagnetic states $\{s^{\text{AF}} | \sum_i s_i^{\text{AF}} \equiv 0\}$.

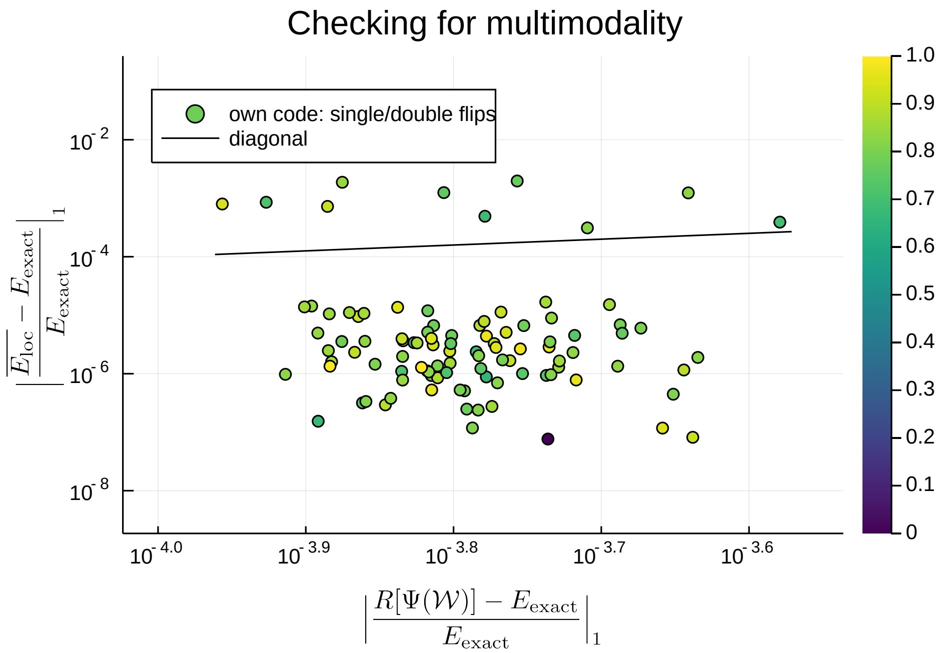


Figure 6.2: Zoom of figure 6.1, including only the single/double flip runs of own code. The colormap indicates the difference between the network parameter set corresponding to the energy with the highest error and others by $|\mathcal{W}_{\max} - \mathcal{W}_j|_2$. There is no indication of clustering due to a multi-modal network parameter set.

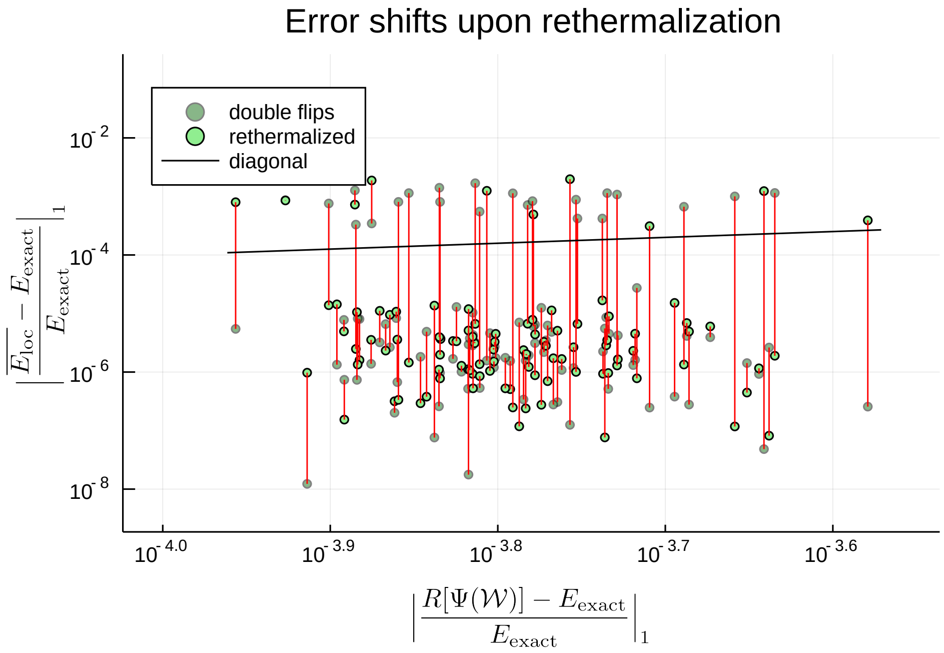


Figure 6.3: Sampling the local energy from a reinitialized and rethermalized Markov Chain indicates that the dichotomy is due to issues in the sampling process.

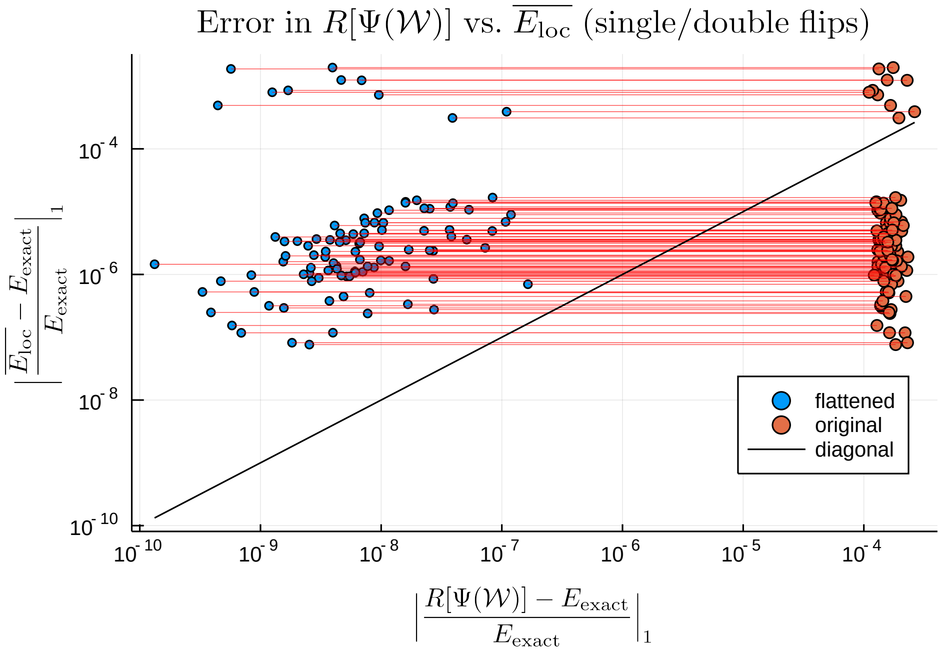


Figure 6.4: Flooring the variational wavefunctions optimised through double flip sampling so that elements $|\Psi_{\mathcal{W}}(s^k)|^2 < 10^{-4}$ are put to 0, results in above-diagonal errors in the Rayleigh quotient.

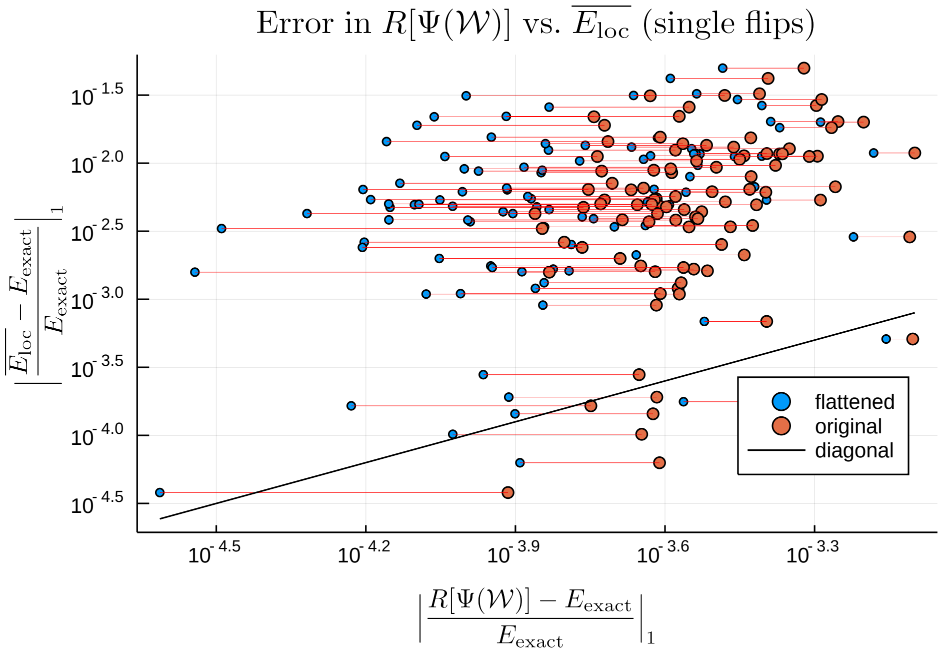


Figure 6.5: Flooring the variational wavefunctions optimised through single flip sampling so that elements $|\Psi_{\mathcal{W}}(s^k)|^2 < 10^{-4}$ are put to 0, results in above-diagonal errors in the Rayleigh quotient.

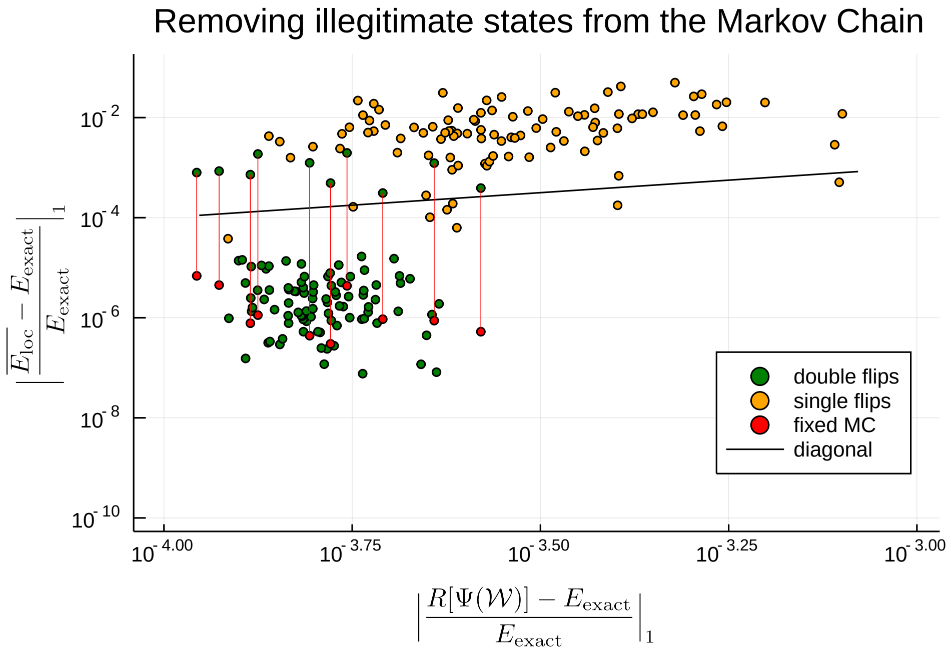


Figure 6.6: Excluding states s^k with $|\Psi_{\mathcal{W}}(s^k)|^2 < 10^{-4}$ (and the collected local energies at these states) from the Markov chains of above-diagonal measurements gathered with the double flip sampler results in a below-diagonal shift of the error in $\overline{E_{\text{loc}}}$

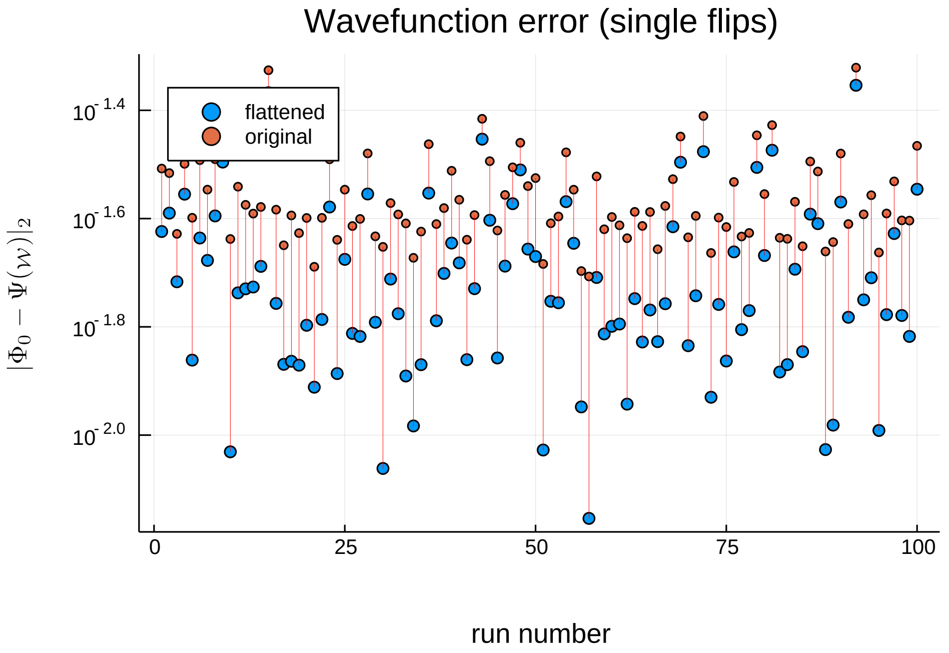


Figure 6.7: The quality of the variational wavefunctions optimized by single flip sampling is increased by flooring elements $|\Psi_{\mathcal{W}}(s^k)|^2 < 10^{-4}$.

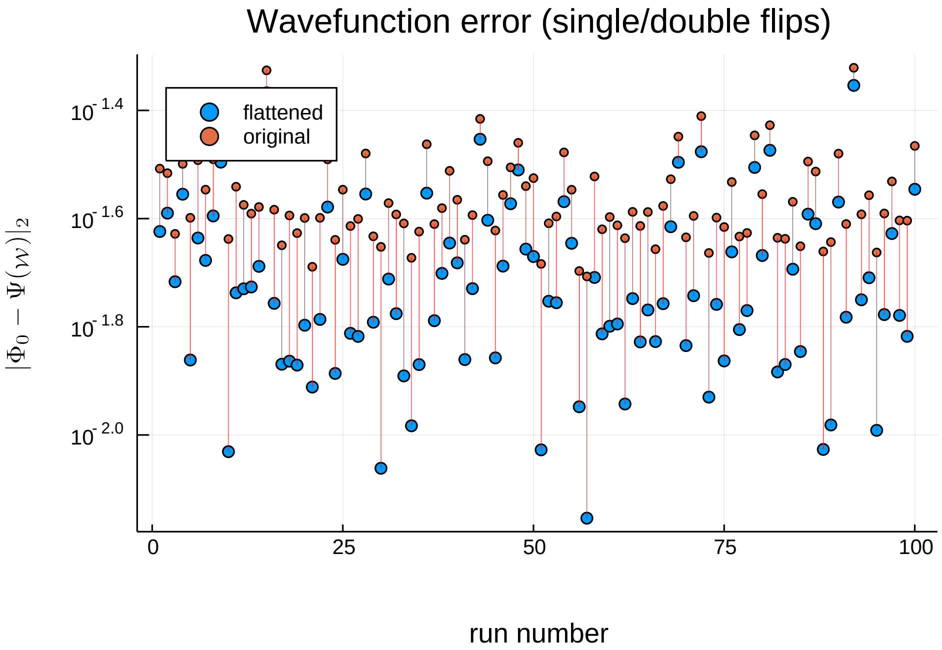


Figure 6.8: The quality of the variational wavefunctions optimized by double flip sampling is increased by flooring elements $|\Psi_{\mathcal{W}}(s^k)|^2 < 10^{-4}$. The extent to which the error is lowered is greater in comparison to the single flip variational wavefunctions. This is explained by the fact that the error in the non-zero elements is also higher for the single flip wavefunctions.

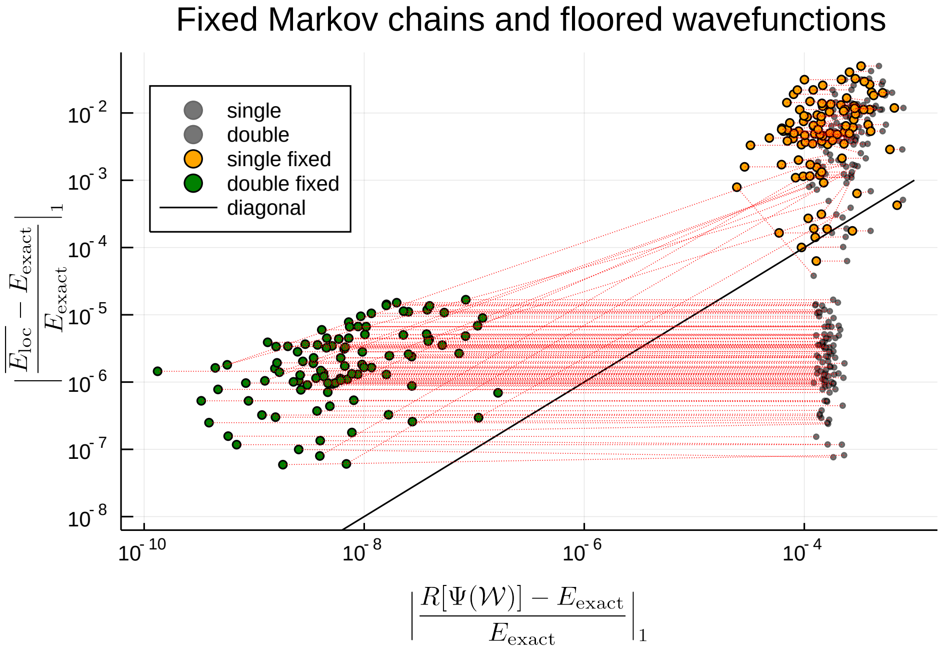


Figure 6.9: The measurements corresponding to floored wavefunction Rayleigh quotients (lateral shift) and fixed Markov chains (vertical shifts). Only a few single flip errors remain below-diagonal, which can be attributed to “fortunate noise”.

6.5.2 Stoquastic XXX AFH model

The XXX AFH model (section 3.1) is a straightforward choice to test the capabilities of the Neural-net Quantum State for a system of non-diagonalizable size. Several statistical tests are performed on this model.

In section 6.5.1 the error in $\overline{E_{\mathrm{loc}}}$ w.r.t. the exact energy depends on the number of flips the sampler performs each step (single or single/double flips). To see if this result generalizes to systems of larger size, the same experiment is repeated for N=40 in figure 6.10. The error clusters corresponding to different samplers in figure 6.1 for N=6 are absent for N=40: using a sampler which proposes single and double flips does not lead to more accurate variational wavefunctions for larger system sizes.

Using a sampler that only visits $\sum_i s_i = 0$ states does improve convergence of $\overline{E_{\mathrm{loc}}}$ to the exact GS energy: see figure 6.11.

One reaches a point of diminishing returns when increasing the hidden unit density α above some value. Figure 6.12 shows that the increased expressibility of the model only reduces the error in the ground state energy for larger ($N > 20$) systems. Even for N=40, using $\alpha = 6$ does not give significantly better results than $\alpha = 12$. This result was also found in [68].

As a proof of concept, a NQS was also trained for an N=100 system, returning relative energy errors similar to N=40 for $\alpha = 2, 6$ (see appendix B).

With regards to the Quantum Boltzmann Machine, it is important that spin statistics are retrieved with low error. Looking specifically at estimations of $\langle \sigma^x \rangle$, the error scales with the system size (table 6.1). Moreover, the error increases with large α. Other statistics and spin correlation functions are retrieved with one to two orders of magnitude higher accuracy (table 6.2).

Average NQS estimates $\widehat{\langle \sigma^x \rangle}$

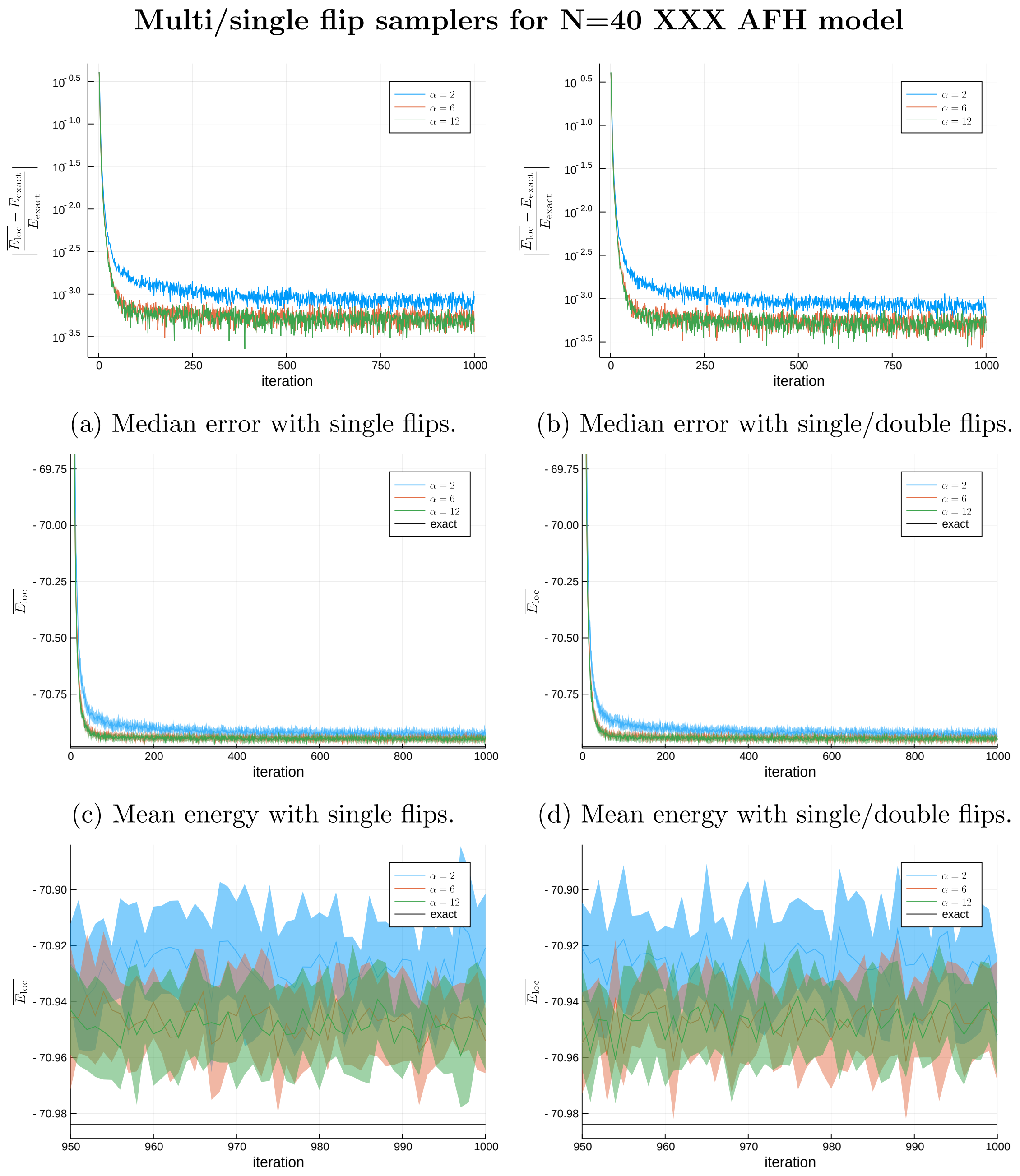
		N			
		6	10	20	40
α	2	0.011	0.013	0.026	0.073
	6	0.018	0.025	0.044	0.093
	12	0.019	0.028	0.048	0.096

Table 6.1: Mean NQS estimates of $\langle \sigma^x \rangle$ statistics after 10 runs. Exact $\langle \sigma^x_i \rangle = 0$.

NQS spin statistics RMS errors

	N			
	6	10	14	20
$\langle \sigma^y_i \rangle$	0.002	0.001	0.001	0.001
$\langle \sigma^z_i \rangle$	0.005	0.005	0.005	0.006
$\left\langle \sigma^x_{ij} \right\rangle$	0.002	0.002	0.003	0.002
$\left\langle \sigma^y_{ij} \right\rangle$	0.002	0.002	0.003	0.002
$\left\langle \sigma^z_{ij} \right\rangle$	0.003	0.003	0.004	0.005

Table 6.2: Mean RMS errors of the spin statistics of the NQS for $\alpha = 2$ after 10 independent runs per value. Exact statistics retrieved by diagonalization. In correspondence with the results of table 6.1, errors are larger for larger values α (results not shown).



(a) Median error with single flips.

(b) Median error with single/double flips.

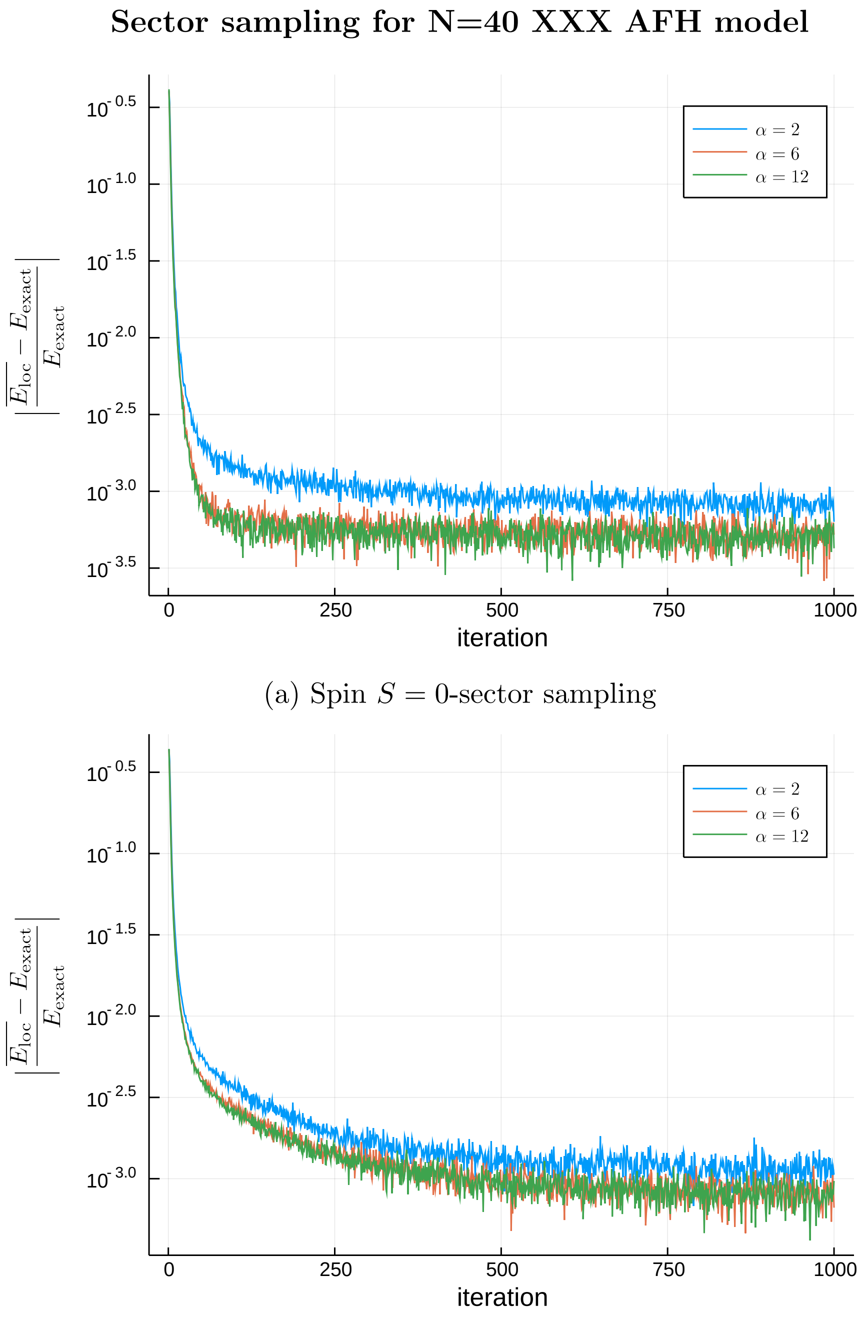
(c) Mean energy with single flips.

(d) Mean energy with single/double flips.

(e) Inset of figure 6.10c.

(f) Inset of figure 6.10d.

Figure 6.10: Convergence of the local energy sampling in the $S = 0$-sector. Learning rate $\eta = 0.05$ and decay factor $\gamma = 0.997$. Sampling on 25 CPUs each collecting 100 samples per iteration for hidden unit densities $\alpha = 2, 6, 12$. Results averaged over 10 trials are nearly identical for the single and single/double flip samplers, for all hidden unit densities.



(a) Spin $S = 0$-sector sampling

(b) Naive sampling

Figure 6.11: Errors in the local energy medialized over 10 runs. Learning rate $\eta = 0.05$ and decay factor $\gamma = 0.997$. Sampling on 25 CPUs each collecting 100 samples per iteration for hidden unit densities $\alpha = 2, 6, 12$. Results averaged over 10 trials are different depending on the sector that is sampled. The $S = 0$-sector sampler (figure 6.11a) converges in less iterations to a lower minimum with a lower variance.

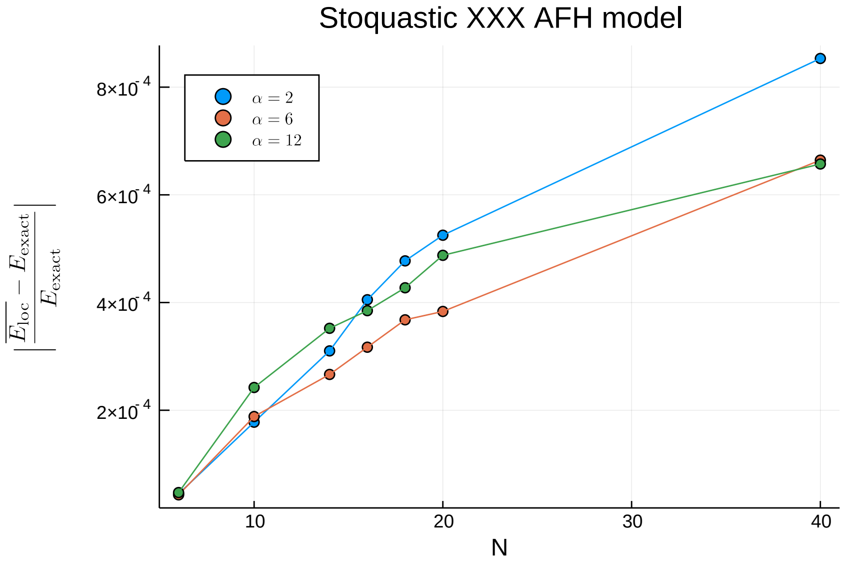


Figure 6.12: Median errors after 10 runs. Naive (not $S = 0$-sector) sampling method with single/double flips. Learning rate $\eta = 0.05$ with decay rate $\gamma = 0.997$. Samples per iteration equal to $10 \cdot N$ per CPU on 25 CPUs. 750 iterations total, local energy averaged over last 50 optimizations (iter $701 - 750$). Added expressibility only lowers error for large system sizes.

6.5.3 Non-stoquastic XXX AFH model

Training the NQS for the non-stoquastic AFH model (not transformed according to the Marshall sign rule) leads to poor results. See figure 6.13 for N=40. In order to investigate the quality of the learned sign structures, a converged (10^{-3} error) and non-converged run (10^{-1} error) are compared. The signs of the wavefunctions are checked according to Marshalls rule (equation 3.9). The global phase is estimated by the phase of the Neel-ordered configuration $\left|s^{\text{Neel}}\right\rangle = |1, -1, 1, \ldots, -1\rangle$, since the ground state wavefunction has large norm at this configuration and thus a low relative error in the phase of $\| \left\langle s^{\text{Neel}} \middle| \Psi_{\mathcal{W}} \right\rangle \|$. For 10^8 uniformly sampled spin configurations, the norm mass with correct sign is divided by the total sampled norm mass (weighted average), resulting in $M^{\text{low}}_{\text{corr.sign}} = 0.9988$ and $M^{\text{high}}_{\text{corr.sign}} = 0.9964$ for the low and high error wavefunctions respectively.

Seeing as the non-stoquastic model fails to converge for large N, the breakdown behaviour of the NQS is tested for varying amount of exact samples and α in figure 6.14. In this figure it is shown that the probability of retrieval decreases with system size and increases with the number of samples. Also, for higher α, the retrieval rate is lower: training the NQS to learn signs is hard and becomes harder for more complex Ansatze.

In order to deepen the insight in the effects of non-stoquasticity, the comparison is also made for the stoquastic and non-stoquastic N=6 XXX AFH model. Ten separate NQS optimizations were performed for both modes using exact sampling to rule out any Markov Chain issues. The results for the stoquastic model were significantly better, sampling one order of magnitude lower errors in the ground state energy after convergence (figure 6.15). The explanation for this can be found by assessment of the quality of the wavefunctions: the overlaps of the (non-) stoquastic NQS wavefunctions are 0.999 ± 0.003 and $0.99977 \pm 6 \cdot 10^{-5}$ respectively.

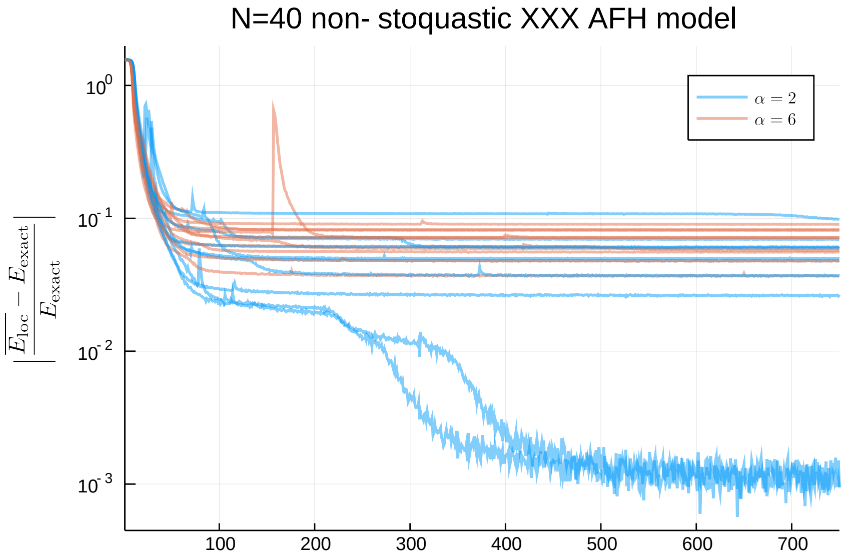


Figure 6.13: Naive (not $\sum_i s_i = 0$ sector) MCMC sampling method with single/double flips of the non-stoquastic N=40 AFH model. Learning rate $\eta = 0.05$ with decay rate $\gamma = 0.997$. Samples per iteration equal to $10 \cdot N$ per CPU on 25 CPUs. Retrieval of the ground state is hard: out of twenty runs, only two converged for $\alpha = 2$.

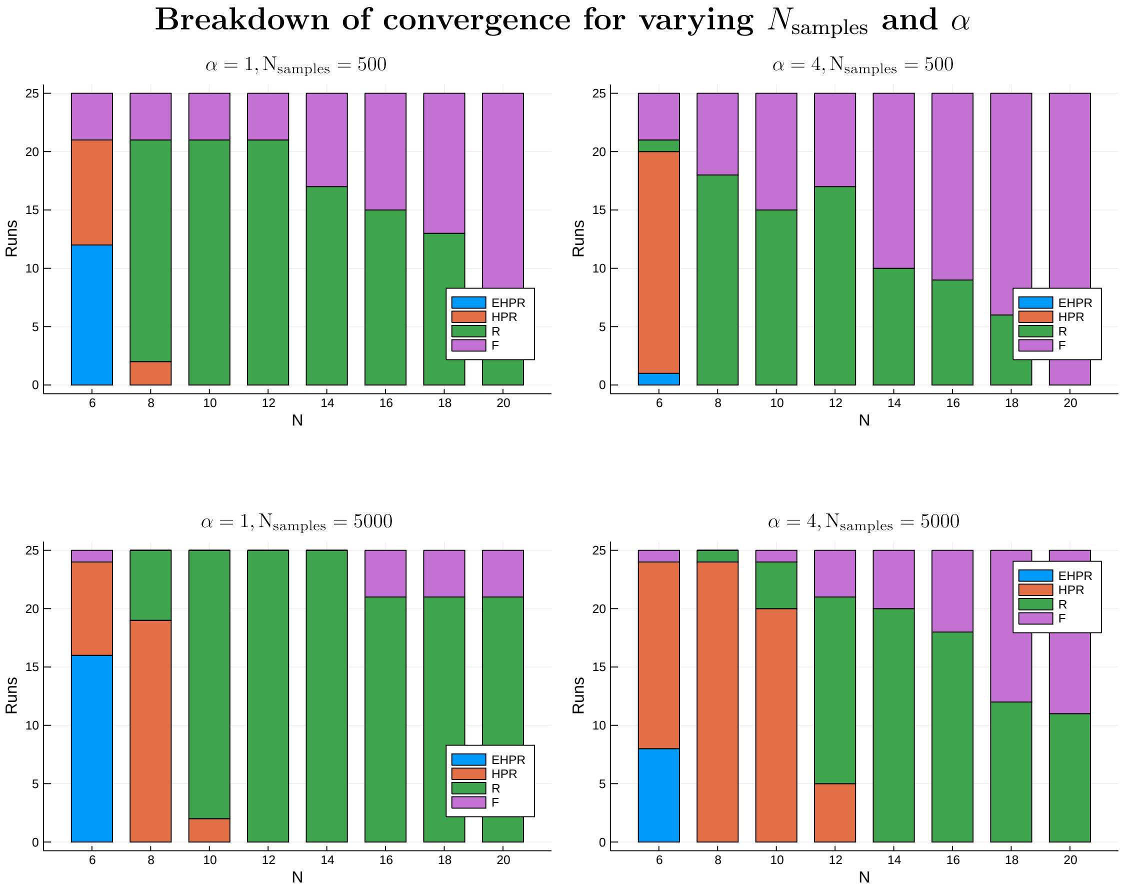
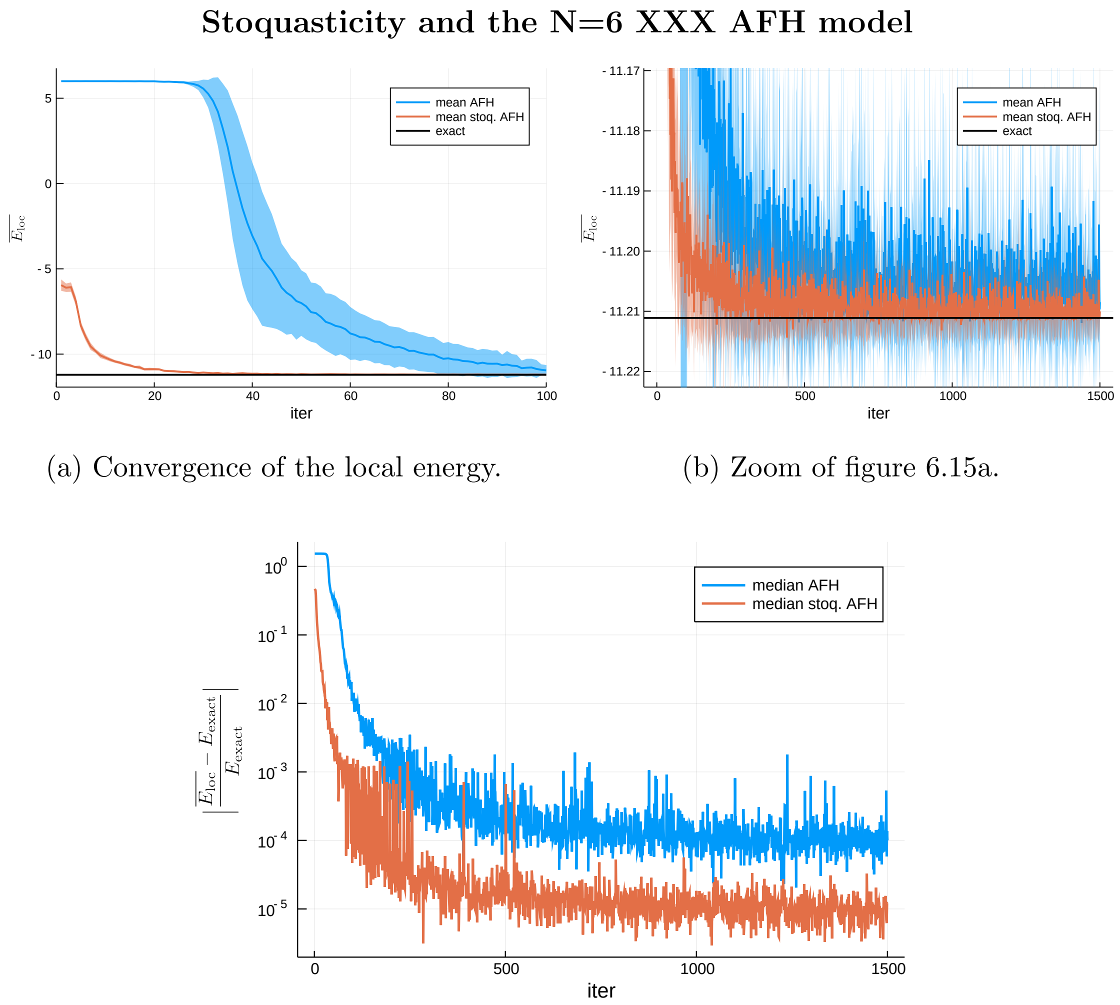


Figure 6.14: Retrieval rate of the ground state energy after 25 runs of 1000 epochs for system sizes N=6 to N=20. Learning rate $\eta = 0.05$ and $\gamma = 0.997$. NQS optimized with exact samples. Extremely High Precision Retrieval (*EHPR*): relative error $\epsilon_{\text{rel}} = \|\overline{E_{\text{loc}}}/E_{\text{exact}} - 1\| < 10^{-6}$; High Precision retrieval (*HPR*): $10^{-6} \leq \epsilon_{\text{rel}} < 10^{-4}$; Retrieval ($R$): $10^{-4} \leq \epsilon_{\text{rel}} < 10^{-2}$; Failure ($F$): $\epsilon_{\text{rel}} > 10^{-2}$.



(a) Convergence of the local energy.

(b) Zoom of figure 6.15a.

(c) Median errors as a function of iteration. Note that visual comparison of the variance of both models is misleading due to the logarithmic scale.

Figure 6.15: Energy convergence of 10 runs on both the stoquastic and non-stoquastic AFH model for 6 spins (transparent colors denote standard deviation). Optimization done with $\alpha = 3$ and 200 exact samples per optimization. Learning rate $\eta = 0.05$ kept constant during run. S-matrix inverted exactly.

6.5.4 Stoquastic connected models

The bad convergence results of the ground state energy in the non-stoquastic XXX AFH model impose a limit to the scope of the Neural-net Quantum State algorithm, in particular with regards to its application to a Quantum Boltzmann Machine. Before delving further into results w.r.t. the QBM, more (stoquastic) models should be tested in order to answer the question: "Suppose the QBM encounters only stoquastic Hamiltonians in its learning process, can we trust the ground state spin statistics of the NQS for these Hamiltonians?" The XXX AFH model gives only a partial answer to this question: for nearest-neighbor, uniformly coupled, non-frustrated Hamiltonians.

The performance of the NQS is further tested for systems of diagonalizable size, with arbitrary stoquastic couplings $w_i^{x,y,z}$, $w_{ij}^{x,y,z}$. The weights of one model are shown in figure 6.16, dubbed the *RSC* (Random Stoquastic Connected) model[9]. The sampled statistics of one run are shown in figure 6.18. Highly accurate convergence to the true ground state energy is found for systems up to 20 spins in figure 6.19. Of course, the RSC model is not representative of all stoquastic connected models. However, the RSC model results indicate that fully connected stoquastic models of non-diagonalizable size can in fact be learned by the NQS.

[9] The weights are not strictly random, but generated by an arbitrary but stoquastic and deterministic recipe for different system sizes.

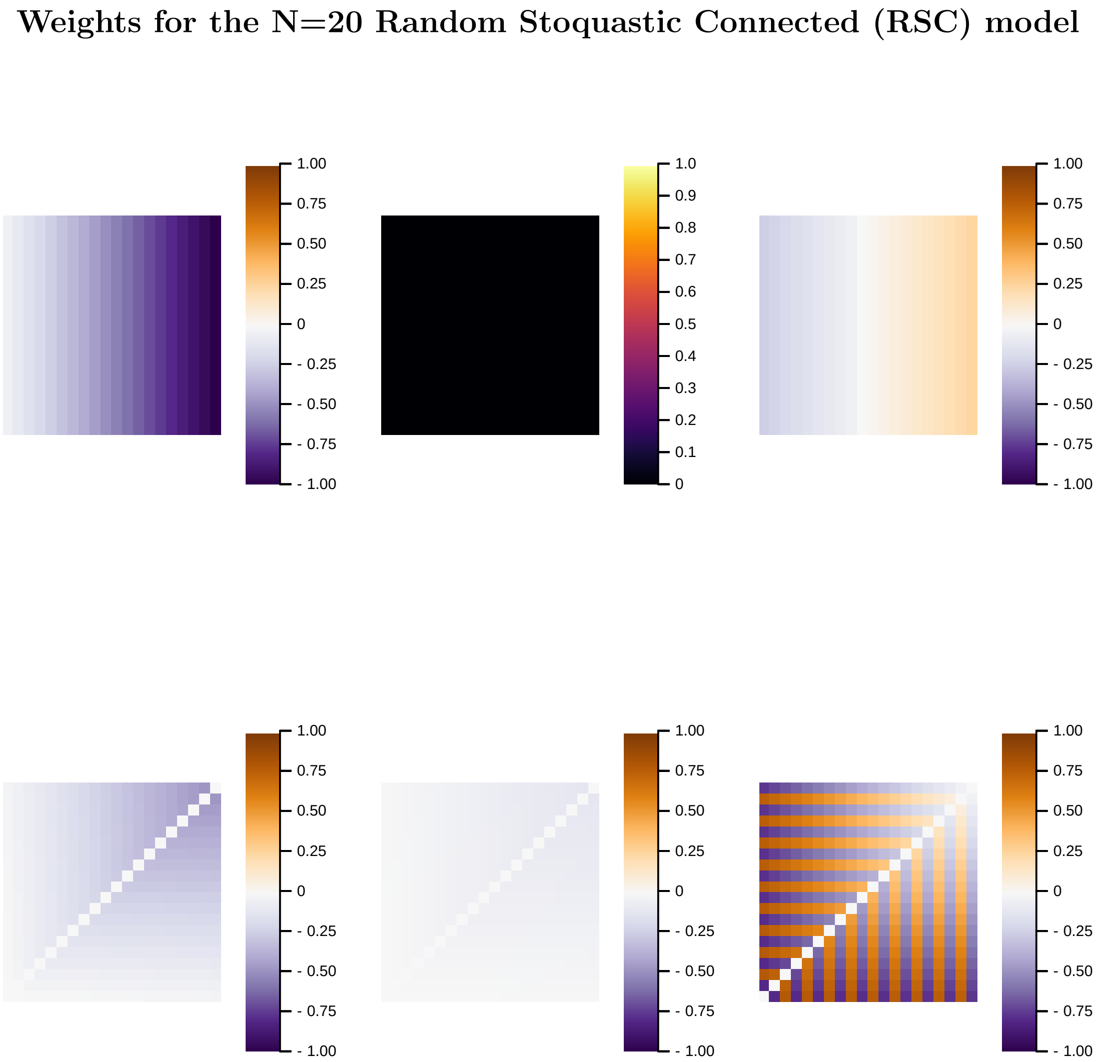


Figure 6.16: In natural reading order, the: w_i^x, w_i^y, w_i^z, w_{ij}^x, w_{ij}^y and w_{ij}^z weights of the N=20 RSC model.

RSC Hamiltonian statistics comparison

Exact statistics

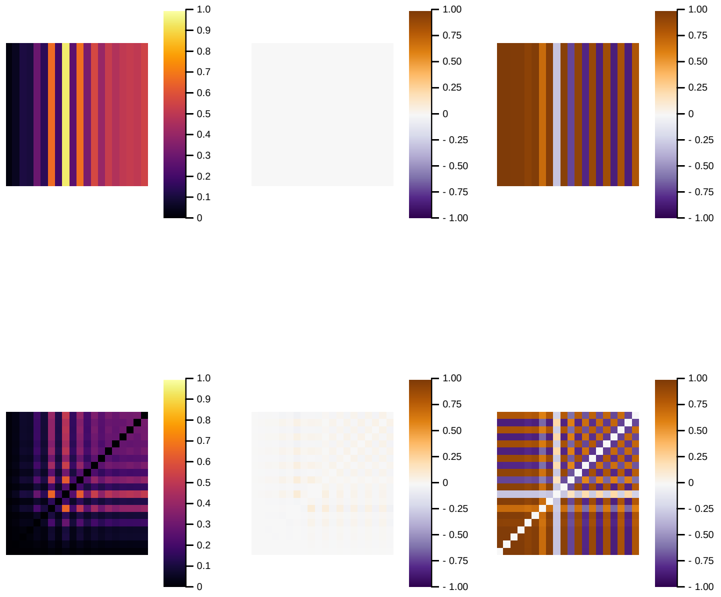


Figure 6.17: The single site (top row) and correlation (bottom row) x, y, z-statistics (respectively in the first to third column) of the RSC Hamiltonian.

NQS statistics

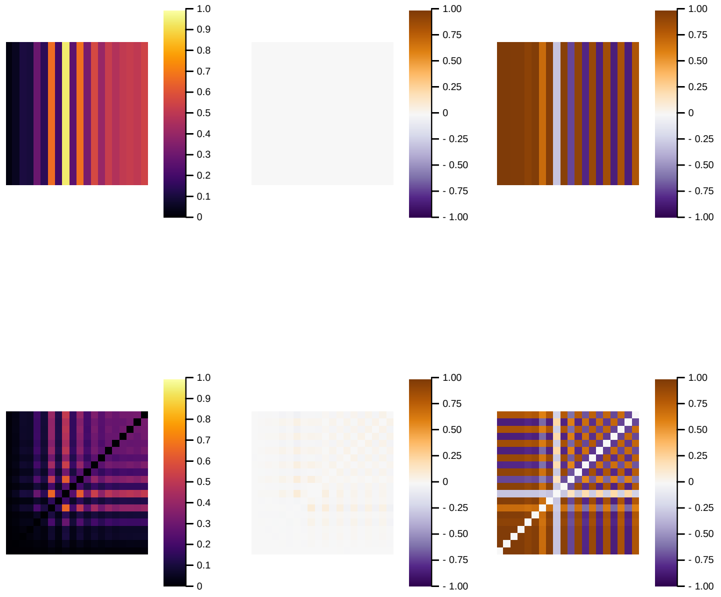


Figure 6.18: The single site (top row) and correlation (bottom row) x, y, z-statistics (respectively in the first to third column) after NQS optimization. NQS optimized in 750 iterations by sampling the optimized wavefunction on 25 CPUs with $N_{\text{samples}} = 200$ samples per CPU. RMS errors are of order $\mathcal{O}(10^{-3})$, except the error in σ^y, which is of order $\mathcal{O}(10^{-4})$.

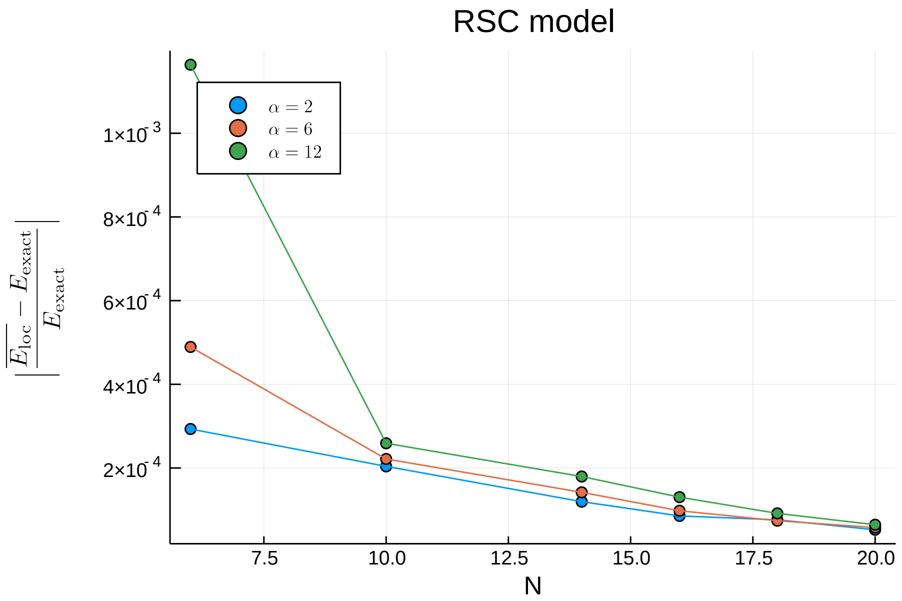


Figure 6.19: Median errors of the RSC model after 10 runs. Local energy averaged over the last 50 iterations. Learning rate $\eta = 0.05$ with decay parameter $\gamma = 0.997$. Double flip sampling used for all measurements. Surprisingly, increasing the hidden unit density increases the error of the ground state energy for the RSC model: a simpler model works better for the fixed number of MC samples ($10 \cdot N$ samples per CPU on 25 CPUs). Moreover, the energy error is inversely related to the system size. Comparing this figure to the stoquastic AFH model in figure 6.12, it is undecided but probable that increasing the hidden unit density improves performance of the NQS for $N > 20$ (see also table 6.4)

.

6.5.5 Varying the number of MC samples

The results of figure 6.12 and 6.19 illustrate that simply increasing the hidden unit density is not a guarantee for a more accurate NQS wavefunction. Increasing the number of samples is another obvious way in which the accuracy of the sampled gradients -and therefore the NQS- could be increased. In order to illustrate the interplay between the amount of samples N_{samples} and α, various experiments are done for the stoquastic AFH and RSC model. The ranges of tested parameters are $\alpha = 2, 6, 12$ and $N_{\text{samples}} = 100, 200, 500$ per CPU on ten CPUs for N=20 systems with $\eta = 0.05$ and decay rate $\gamma = 0.997$. All experiments were repeated ten times to produce averages and variances.

Increasing the number of samples does not decrease the number of iterations needed to converge to a stable local energy (figure 6.20). However, as expected, using more samples reduces the variance of the local energy (figure 6.21-6.22). As was shown before in figure 6.12 (AFH) and figure 6.19 (RSC), increasing the hidden unit density does not decrease errors beyond $\alpha = 6$. This is again confirmed by comparison of the $\alpha = 6, 12$ results in this subsection.

The overlaps of the wavefunction w.r.t. results from exact diagonalization are shown in tables 6.3 and 6.4.

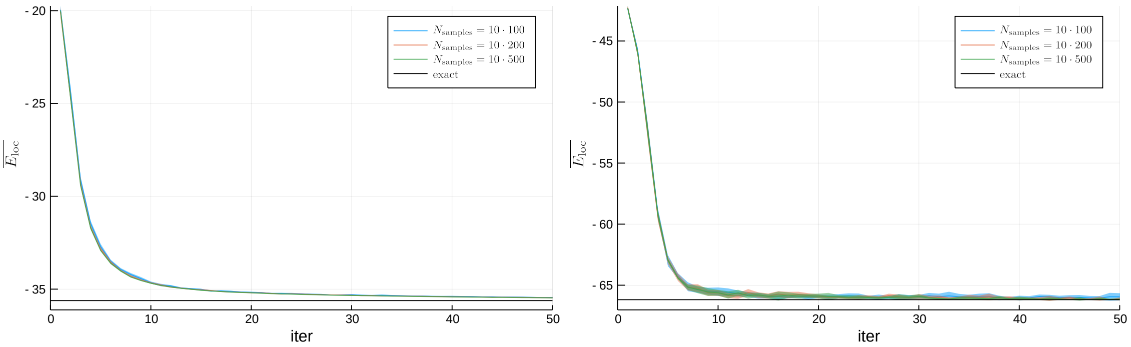
		N_{samples}		
		1000	2000	5000
α	2	0.9973 ± 2e−4	0.9981 ± 1e−4	0.9987 ± 1e−4
	6	0.9963 ± 2e−4	0.9974 ± 2e−4	0.9980 ± 1e−4
	12	0.9959 ± 6e−5	0.9970 ± 6e−5	0.9976 ± 3e−5

Table 6.3: Overlaps with the true wavefunction for the AFH model. For N=20, small hidden unit density still suffices to capture the ground state. Better overlaps only by decreasing MC noise.

		N_{samples}		
		1000	2000	5000
α	2	0.9989 ± 2e−4	0.9993 ± 4e−4	0.9996 ± 2e−4
	6	0.9996 ± 4e−4	0.9997 ± 2e−4	0.9999 ± 9e−5
	12	0.9996 ± 3e−4	0.9998 ± 9e−5	0.9999 ± 3e−5

Table 6.4: Overlaps with the true wavefunction for the N=20 RSC model. At variance with the AFH model (table 6.3), added expressibility results in marginally increased overlaps with slightly lower variances. The RSC ground state is retrieved with overall increased accuracy w.r.t. the AFH model.

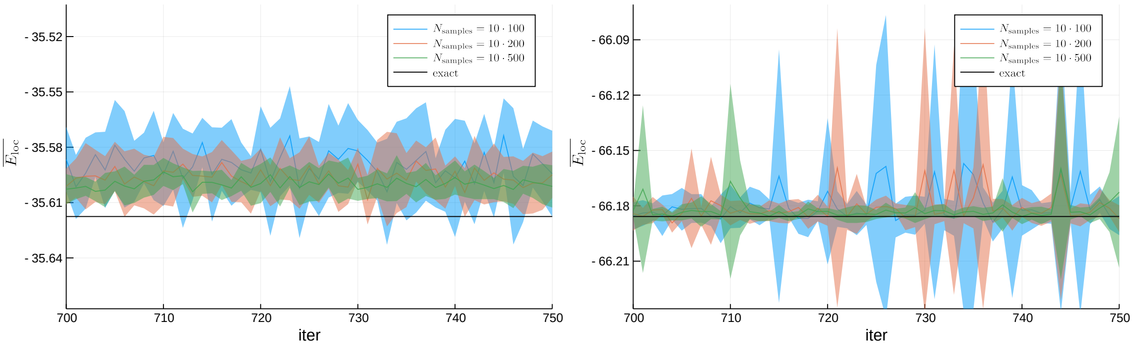
Convergence of the N=20 AFH (left) and RSC (right) models averaged over 10 separate runs for varying N_{samples}



(a) Stoq. XXX AFH

(b) RSC

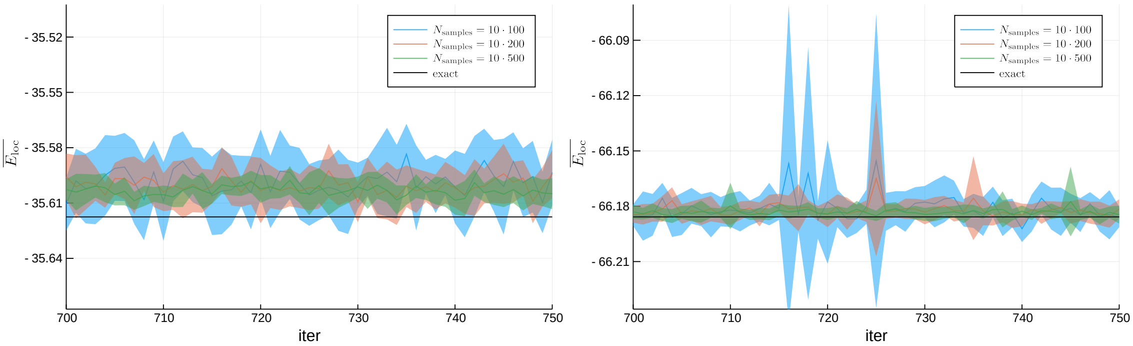
Figure 6.20: First 50 iterations for $\alpha = 12$. Increasing the number of samples does not reduce convergence time using Markov chains of length approximately $1\%, 2\%$ and 5% the size of Hilbert space. Learning done on 10 CPUs.



(a) Stoq. XXX AFH

(b) RSC

Figure 6.21: Last 50 iterations for $\alpha = 2$.



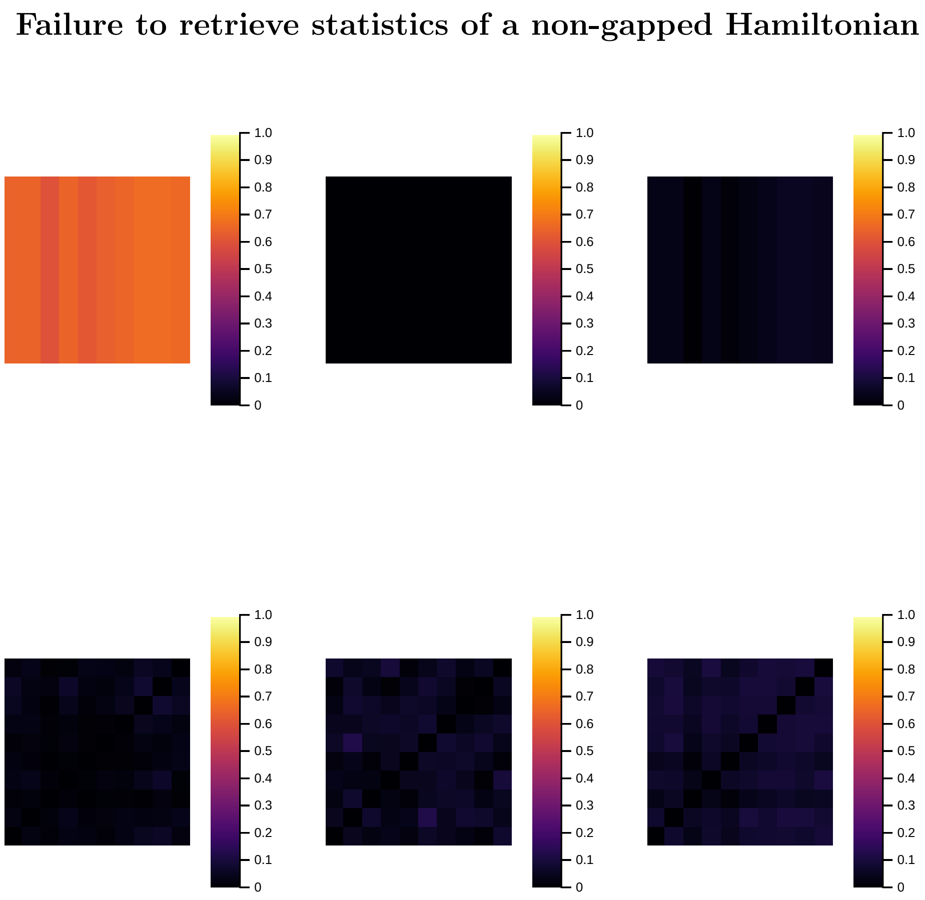
(a) Stoq. XXX AFH

(b) RSC

Figure 6.22: Last 50 iterations for $\alpha = 6$. Variances between separate runs and iterations are reduced for both models w.r.t. the $\alpha = 2$ experiments in figure 6.21.

6.5.6 Failure to converge without a spectral gap

The Rayleigh quotient is a problematic objective function for finding the ground state wavefunction when the ground state energy is degenerate. In such cases, the found ground state energy is found with small relative error, while other statistics are not guaranteed to be correct. A demonstration is given in figure 6.23. This Hamiltonian is generated by a converged Quantum Boltzmann Machine trained w.r.t. a classical data density matrix generated by the asymmetric parity probability distribution. This demonstrates that the QBM may fail to converge when a NQS is used as a substitute for exact diagonalization if the target Hamiltonian has a small spectral gap.



Failure to retrieve statistics of a non-gapped Hamiltonian

Figure 6.23: The absolute values of differences in spin expectation values of the exact ground state wavefunction and the NQS. In natural reading order the expectation values in the x, y, z directions on the first row, and x, y, z correlations on the second row. The spectral gap is of order 10^{-6}, with a ground state energy of -10.879249. Upon examination, the NQS was found to correspond to $\Psi_{\mathcal{W}} \approx \frac{1}{\sqrt{2}}(\Psi_1 - \Psi_2)$ with an overlap of 0.9996, where Ψ_1 and Ψ_2 denote the ground state and first excited state. The relative energy error is 10^{-5}, but the correct $\langle \sigma^x \rangle$ statistics are not retrieved.

6.6 Conclusion

The Neural-net Quantum State is able to capture the statistics and ground state energy of stoquastic models with great precision for systems far beyond the scope of exact diagonalization. While the accuracy of the statistics scales proportionally to the hidden unit density α for the (stoquastic) XXX AFH model, the fully connected RSC model shows different results. There is no one-size-fits all solution to produce the statistics of randomly connected models with a given accuracy. Nevertheless, for moderately sized ($N \approx 30$) stoquastic models there is hope that the NQS can be used to train a Quantum Boltzmann Machine even for small hidden unit densities with errors on spin expectation values of order $O(10^{-2})$.

For the non-stoquastic AFH model, the NQS fails to converge reliably. As discussed, this might relate to the sign-structure generalization problems explored [67]. Moreover, in this paper it has been shown that the performance to learn sign structure heavily relies on neural-net design. Particularly Convolutional Neural Networks seem to outperform Feed-forward Neural Networks and RBMs.

Moreover, the rate of convergence of the NQS does not depend on the amount of samples. In order to maximize efficiency, the amount of samples as well as the hidden unit density should therefore be increased adaptively.

A requirement left out of previous discussions is the smoothness of the probability distribution $|\Psi_{\mathcal{W}}|^2$: for Hamiltonians with ground states that have heavily concentrated probability mass ("spiked" ground state wavefunctions), MCMC is likely to fail. While this did not prove to be a problem for the tested Hamiltonians, adherance to this condition can, strictly speaking, not be ensured for QBM Hamiltonians with random couplings. The smoothness of the NQS proved to be particularly important for the estimation of quantum statistics. The estimator for the σ^x statistics is $\widehat{\langle \sigma_i^x \rangle} = \sum_{s \in |\Psi_{\mathcal{W}}|^2} \frac{\Psi_{\mathcal{W}}(F_i s)}{\Psi_{\mathcal{W}}(s)}$. The ground state of the Heisenberg model has only non-zero norm for zero-sum spin configuration basis states. Therefore, a perfect NQS would be parameterized such that $\Psi_{\mathcal{W}}(F_i s) = 0$ for all $s \in |\Psi_{\mathcal{W}}|^2$. This requires a stepping behaviour from the neural net: a large -strictly speaking infinite- difference in output for inputs separated by a Hamming distance of 1. For systems of size N=6 this can be done with $\mathcal{O}(10^{-2})$ errors, while for N=40 systems, the error approaches $\mathcal{O}(10^{-1})$. While possibly an expressibility issue, the problem was shown not to be solved by increasing the hidden unit density. In order to increase the effective Hamming distance between states with large norm difference, different architectures or data preprocessing strategies (such as basis transformations) should be researched.

Chapter 7

Results

As shown, the Quantum Boltzmann Machine can be trained using only ground state statistics, in what is called the “rank-1 approximation”. The final step is to generate these statistics numerically, using the Neural Quantum State method, so that the QBM can be trained for Hamiltonians beyond the scope of exact diagonalization. The total pipeline is visualized in figure 7.1.

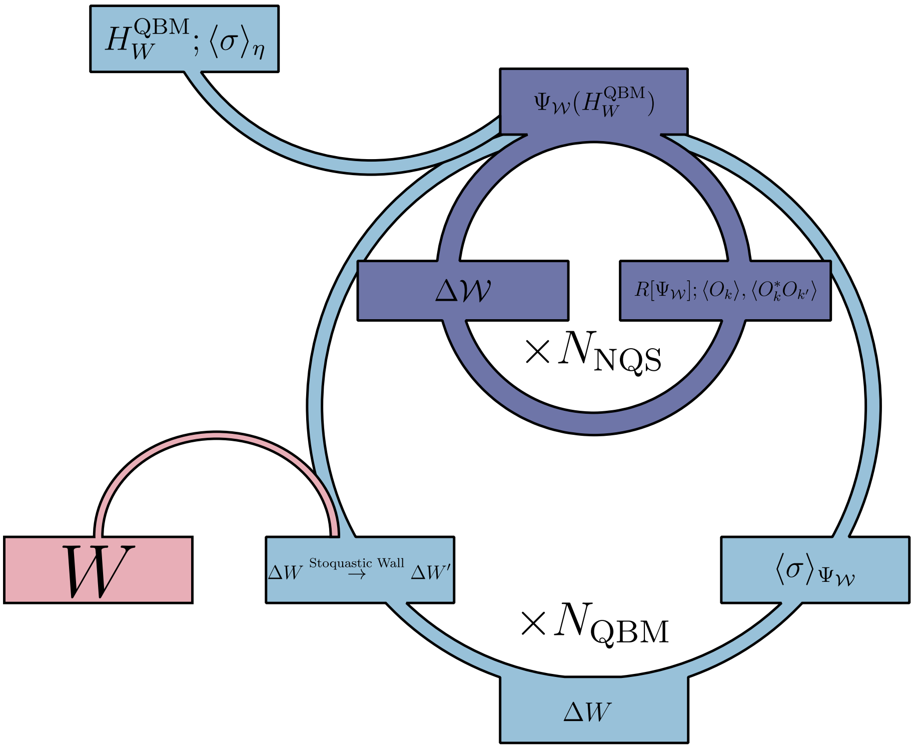

$H_W^{\text{QBM}}; \langle\sigma\rangle_\eta$
$\Psi_{\mathcal{W}}(H_W^{\text{QBM}})$
$\Delta\mathcal{W}$
$R[\Psi_{\mathcal{W}}]; \langle O_k\rangle, \langle O_k^* O_{k'}\rangle$
$\times N_{\text{NQS}}$
W
$\Delta W \overset{\text{Stoquastic Wall}}{\rightarrow} \Delta W'$
$\langle\sigma\rangle_{\Psi_{\mathcal{W}}}$
$\times N_{\text{QBM}}$
ΔW

Figure 7.1: The pipeline for training a QBM with usage of ground state statistics, calculated from the Neural Quantum State. W parameterizes the QBM Hamiltonian, calligraphic $\mathcal{W}$ the NQS, $R[\Psi_{\mathcal{W}}]$ denotes the Rayleigh quotient and $\langle O_k\rangle$ the expectation value of the variational derivative w.r.t. network parameter k. The inner loop is repeated N_{NQS} times for every outer loop update.

In the rank-1 approximation, the statistics are degenerate: the ground states of many Hamiltonians may correspond to the same statistics. Performing quantum state tomography with the QBM is therefore impossible in the rank-1 approximation. However, the QBM could still be used as a generative model. The hypothesis is that the QBM, with its inherent quantum features, outperforms classical generative models. The QBM has been shown to outperform the BM for classical data (coming from retinal cells of a salamander) [54].

The source of the data statistics can be either quantum or classical, coming from any source. However, smoothness of the variational ground state wavefunction should be taken into account. Although there is no one-to-one correspondence, data sets that are very biased to a small number of statistics are probable to be generated by non-uniform ground state wavefunctions of the QBM parameterized Hamiltonian. The non-uniformity can be a problem in terms of expressibility of the Ansatz, but more importantly for the ergodicity of MCMC in the training process.

In particular for small absolute value spin statistics on the x and y-axis, which are sampled by evaluating products of local wavefunctions (see equation 5.41), the Restricted Boltzmann Machine Ansatz might suffer from the problem explored in section 6.5.1. In any case, the quantum statistics are the most difficult to estimate for the NQS: the error on these estimates will determine the quality of the model statistics and quantum likelihood $\mathcal{L}$.

Moreover, the NQS often fails when it has to learn a sign structure 6.5.3). Therefore, a mechanism should be put in place to keep the QBM weights W from forming a non-stoquastic Hamiltonian.

7.1 Stoquastic Wall

Checking for stoquasticity is computationally cheap, since it only requires that $w_i^x < 0$ and $w_{ij}^x \leq -|w_{ij}^y|$ for all i, j. The Stoquastic Wall ensures that the QBM Hamiltonian is stoquastic, so that the NQS is not tasked with learning a sign structure. If the QBM proposes a change in weights w_{ij}^x and w_{ij}^y so that the Wall is breached, the learning rate for the individual weights are lowered so that the weights lie exactly on the barrier. This implies

$$w_{ij}^x + \epsilon \Delta w_{ij}^x = -|w_{ij}^y + \epsilon \Delta w_{ij}^y| \tag{7.1}$$

$$\epsilon = \frac{\pm w_{ij}^y - w_{ij}^x}{\Delta w_{ij}^x \mp \Delta w_{ij}^y}, \tag{7.2}$$

where the $\epsilon \geq 0$ solution corresponds to the correct barrier. This is also done in a trivial manner for the w^x weights. The Stoquastic Wall requires negligible computing time relative to the computation of model statistics.

7.2 Experimental parameters

Combining the Quantum Boltzmann Boltzmann Machine and the Neural Quantum State algorithms results in a large number of hyperparameters. These parameters are tuned using experience from results in previous chapters. This includes a learning rate for the NQS of $\eta = 0.05$ (not to be confused with the data density matrix η) with decay parameter $\gamma = 0.996$, $N_{\text{NQS}} = 750$ and hidden unit density $\alpha < 6$. Due to the stochastic nature of the NQS spin statistics, the QBM learning rate is kept at a conservative $\epsilon = 0.02$. The amount of required QBM iterations N_{QBM} is monitored manually by inspection of the likelihood curve $\mathcal{L}$.

7.2.1 Stoquastic Wall

The target Hamiltonian is stoquastic for positive data density matrices $\eta \geq 0$. Therefore, the Stoquastic Wall is expected not to be a limiting factor for classical data although the concern of identifiability (section 5.3.2) still plays a role. The QBM weights W_{QBM} are initialized in the stoquastic region. This includes the Marshall-Peierls sign rule Heisenberg model as possibility, so that the initial Hamiltonian has large spectral gap and its gradients are approximated accurately.

7.3 Results

The RMS errors in expectation values of elements H_r of the QBM Hamiltonian are considered separately, e.g. for $k = xx$:

$$\zeta^{xx} \equiv \sqrt{\frac{1}{2N(N-1)} \sum_{j>i} \left[\left\langle \sigma_i^x \sigma_j^x \right\rangle_\eta - \left\langle \sigma_i^x \sigma_j^x \right\rangle_\rho \right]^2}. \tag{7.3}$$

7.3.1 Proof of concept: N=10 neuronal data

In order to demonstrate the QBM can be trained with the NQS for small (diagonalizable) classical problems, we turn to a neuronal data set. The data statistics are extracted from retinal neurons of a salamander. Six experiments are done in order to illustrate the effects of the approximations done by QBM+NQS learning:

- Exact rank-10 learning with/without barrier,
- Exact rank-1 learning with/without barrier,
- NQS rank-1 learning with/without barrier.

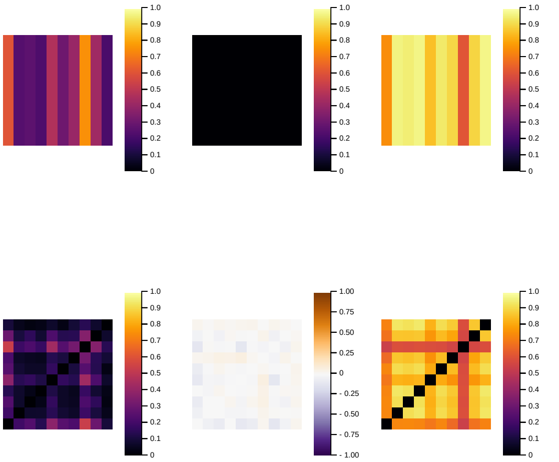
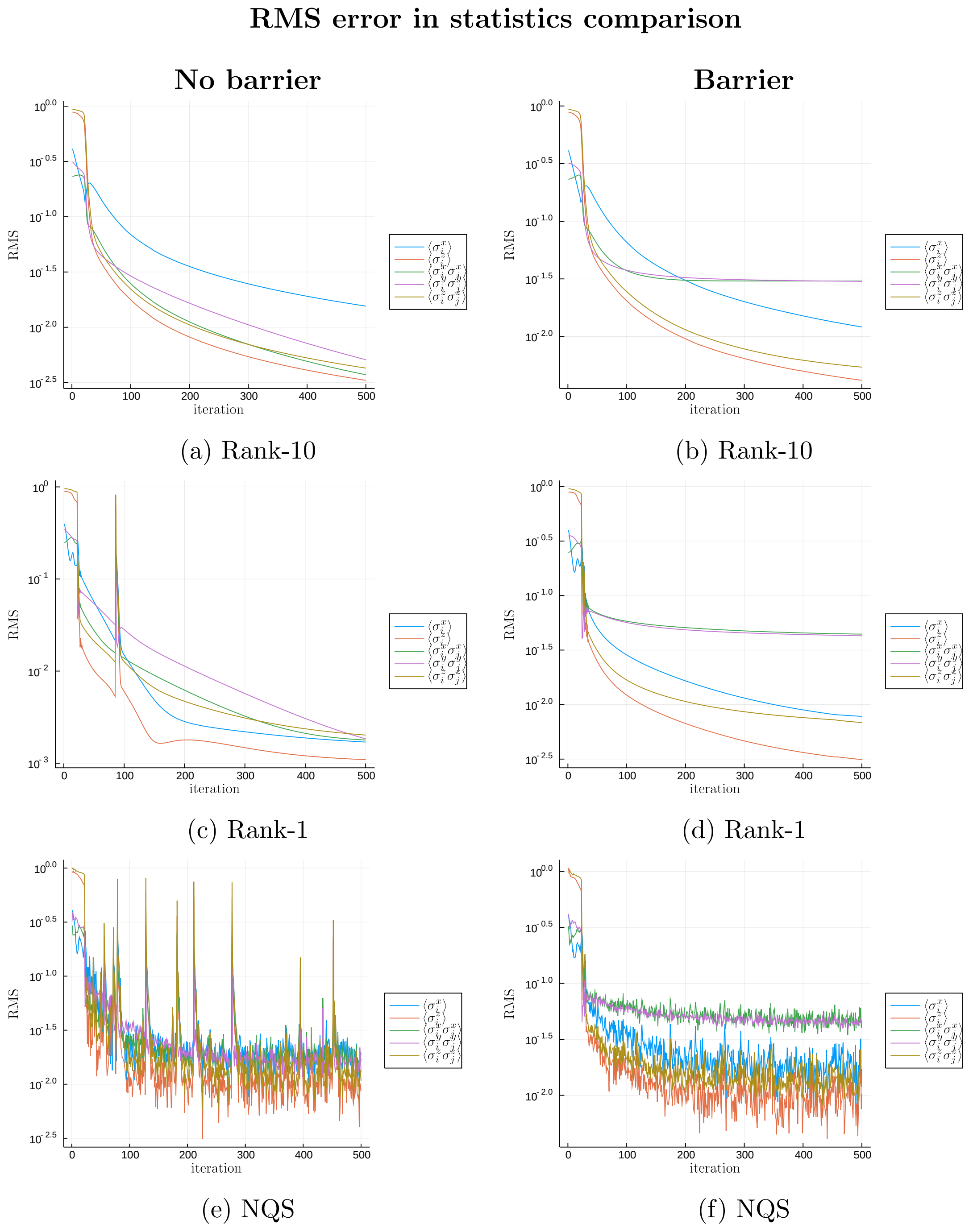


Figure 7.2: Data statistics $\langle \sigma_i^x \rangle$, $\langle \sigma_i^y \rangle$, $\langle \sigma_i^z \rangle$ (top row); $\langle \sigma_i^x \sigma_j^x \rangle$, $\langle \sigma_i^y \sigma_j^y \rangle$, $\langle \sigma_i^z \sigma_j^z \rangle$ (bottom row).



(a) Rank-10

(b) Rank-10

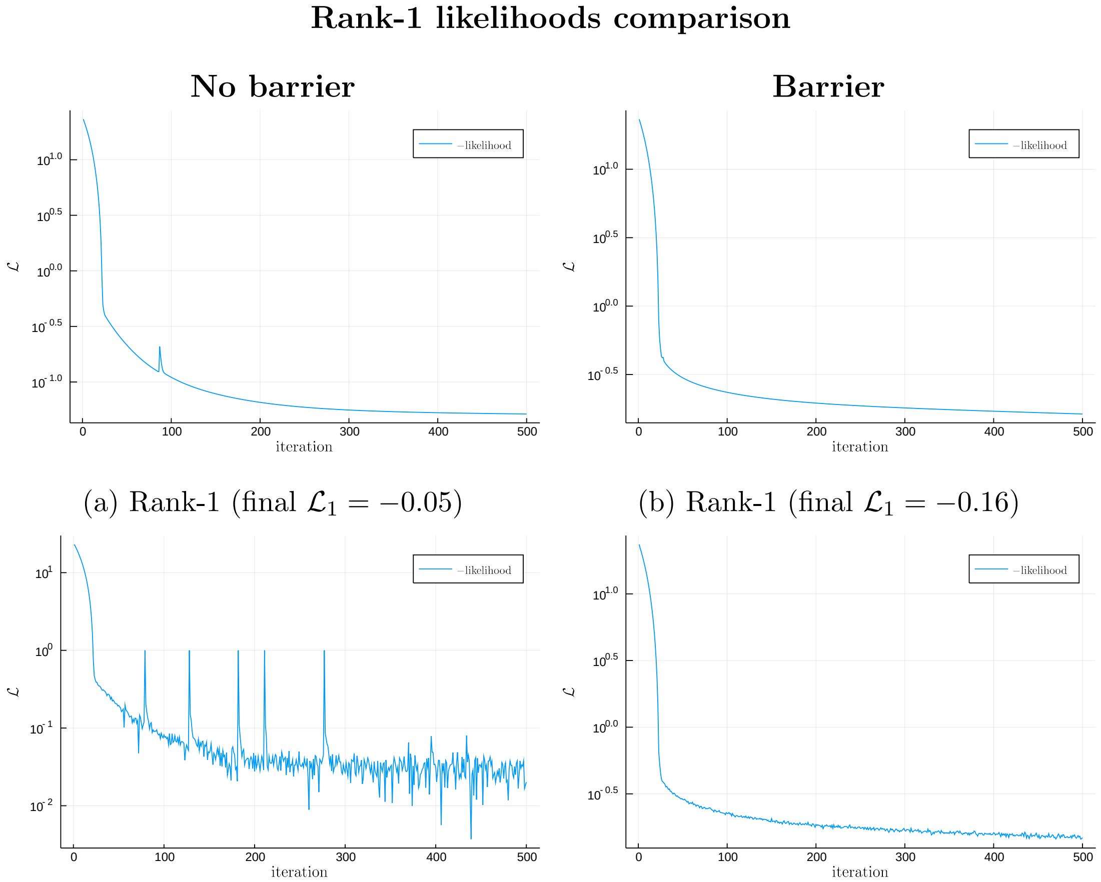
(c) Rank-1

(d) Rank-1

(e) NQS

(f) NQS

Figure 7.3: One QBM iteration corresponds to 750 NQS updates with $\alpha = 2$ and $N_{\text{samples}} = 500$ on a single CPU. QBM and NQS learning rates are $\epsilon = 0.02$ and $\eta = 0.05$ respectively. QBM initialized with XXX Heisenberg model weights for all runs. Note that the rank-10 algorithm does not run into degeneracy spikes. Moreover, the NQS fails to converge and finds incorrect statistics multiple times without the barrier. This is confirmed by spikes in the likelihood (see figure 7.4). The barrier fixes this problem, but leads to an overall worse solution, particularly for $\langle\sigma_i^x\sigma_j^x\rangle$ and $\langle\sigma_i^y\sigma_j^y\rangle$, which can therefore be ascribed to model mismatch.



(a) Rank-1 (final $\mathcal{L}_1 = -0.05$)

(b) Rank-1 (final $\mathcal{L}_1 = -0.16$)

(c) NQS (final $\mathcal{L}_1 \approx -0.04$)

(d) NQS (final $\mathcal{L}_1 \approx -0.15$)

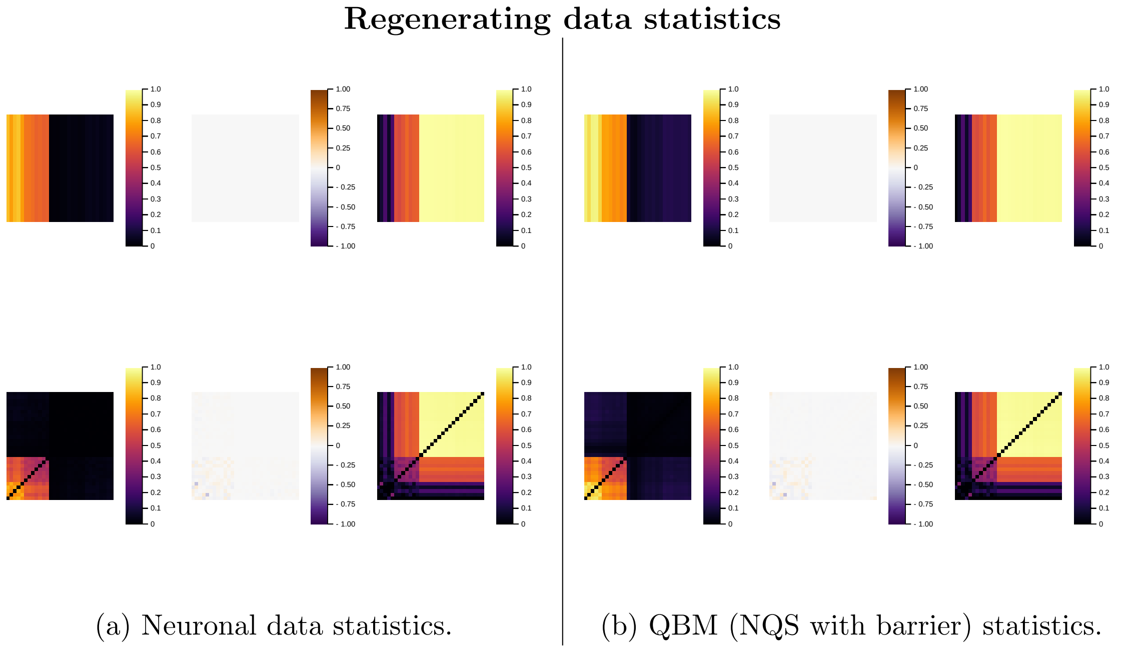
Figure 7.4: The peaks to 10^0 are artificial: when the likelihood was found to be positive, it was set manually to this value. Positive likelihoods indicate failure to converge (the ground state energy is overestimated, see equation 5.53). The barrier leads to worse solutions (lower likelihood) in both the exact and NQS modi. Keep in mind that the plot is logarithmic, so that the noise level may be overestimated for higher values of $\mathcal{L}$ when inspected visually.

7.3.2 N=30 neuronal data

The NQS is only detrimental for system of non-diagonalizable size. A fully connected Hamiltonian for 30 spin-$\frac{1}{2}$ particles is represented by a dense matrix of $\mathcal{O}(10^{18})$ elements, beyond the reach of exact diagonalization on moderately sized computing clusters. The neuronal salamander data set for N=10 is now extended to 30 neurons.

In figure 7.7a, the likelihood can be seen to converge smoothly to a maximum. However, two peaks are shown at iteration 37 and 117. These peaks are preceded by a dip, which indicates that the NQS did not converge properly to the ground state, and subseqently sampled the energy of an excited state. The peaks in this figure correspond to the peaks in the RMS errors in figure 7.6. As was the case for N=10 salamander data with barrier, the classical RMS errors are significantly lower.

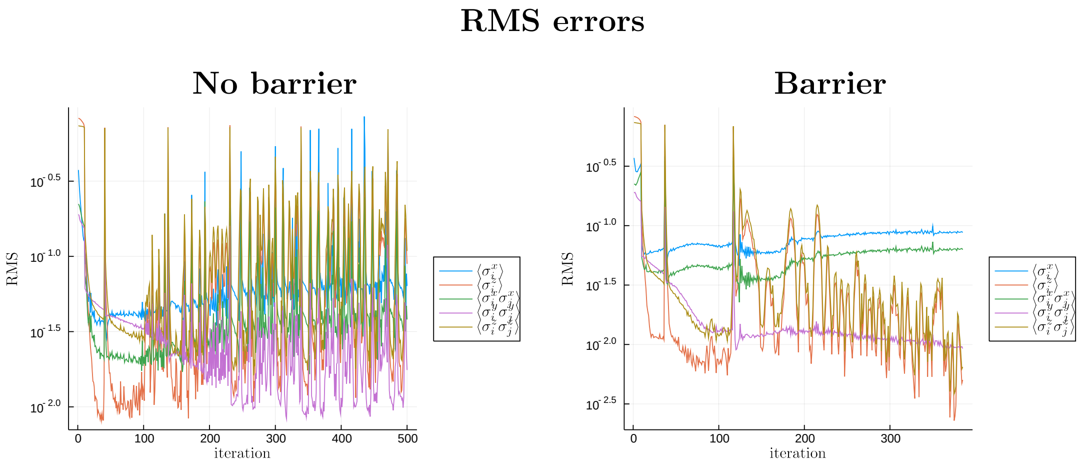
The oscillatory behaviour of RMS values was found to be much more outspoken for the non-stoquastic QBM (figure 7.6a).



(a) Neuronal data statistics.

(b) QBM (NQS with barrier) statistics.

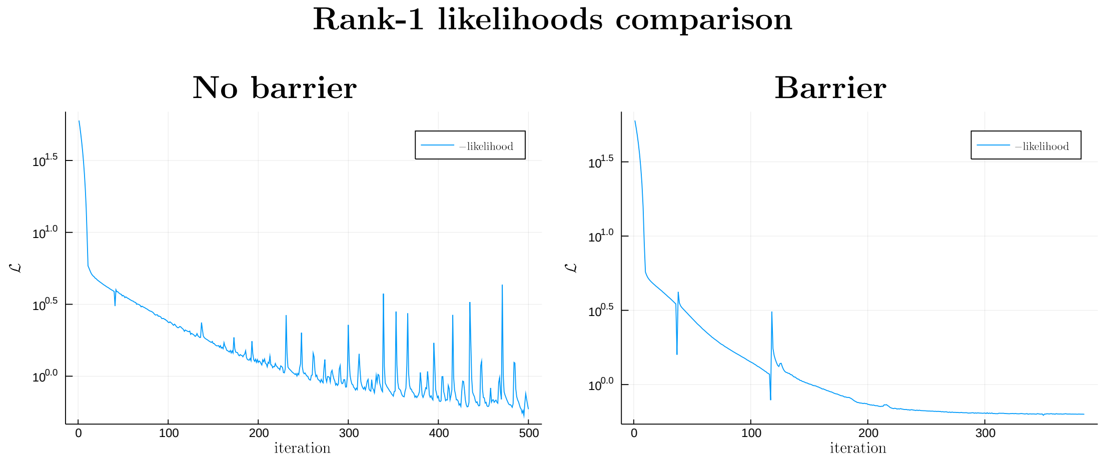
Figure 7.5: The statistics shown in both subfigures (from top left to bottom right) are expectation values $\langle\sigma_i^x\rangle$, $\langle\sigma_i^y\rangle$, $\langle\sigma_i^z\rangle$, $\left\langle\sigma_i^x\sigma_j^x\right\rangle$, $\left\langle\sigma_i^y\sigma_j^y\right\rangle$ and $\left\langle\sigma_i^z\sigma_j^z\right\rangle$. After 350 QBM updates, the data spin statistics are regenerated by the model rank-1 density matrix with total RMS = 0.048. One QBM iteration corresponds to 750 NQS updates with $\alpha = 2$ and $N_{\text{samples}} = 100$ on 30 CPUs each update. QBM and NQS learning rates are $\epsilon = 0.02$ and $\eta = 0.05$ respectively. QBM initialized with XXX Heisenberg model weights.



(a) NQS

(b) NQS

Figure 7.6: The RMS errors ζ^k and as a function of QBM iteration. The results of figure 7.5b correspond to the final iteration in figure 7.6b. Note that the classical statistics have significantly lower RMS errors. In addition, the oscillatory behaviour of RMS values in figure 7.6b -while the likelihood has converged- indicates that the Hamiltonian is degenerate.



(a) NQS (final $\mathcal{L}_1 \approx -0.6$)

(b) NQS (final $\mathcal{L}_1 \approx -0.6$)

Figure 7.7: No positive likelihoods are encountered, which indicates that the ground state energy is found with reasonable accuracy. Keep in mind that the plot is logarithmic, so that the noise level may be overestimated for higher values of $\mathcal{L}$ when inspected visually.

7.4 Conclusion

The QBM was trained with the NQS for N=10 and N=30 classical data sets, originating from recordings of retinal cells of a salamander.

The problems in the rank-1 and NQS modi were all due to a small spectral gap. For the rank-1 modus the problem is simply that the gradient cannot be approximated by only ground state statistics due to a small spectral gap. In addition to this, the NQS modus is sensitive to sampling wrong statistics (converging to a linear combination of ground and excited states) for small spectral gap Hamiltonians:

it samples the wrong gradient when the rank-1 approximation no longer holds.

Although the degeneracy issue seems less apparent for the barrier solutions at N=10, the oscillatory behaviour of RMS errors of the N=30 QBM indicates that the encountered Hamiltonians have small spectral gap.

For the N=30 non-stoquastic (no barrier) QBM, the oscillations were much more prevalent. Two explanations for this are possible. The first is that this QBM run encountered more degenerate Hamiltonians. The second is that the NQS is more sensitive to the spectral gap for non-stoquastic Hamiltonians. The peaks in the likelihood (figure 7.7a) indicate that the NQS often sampled an inaccurate gradient. As discussed, it is unclear if this is failure of the NQS to converge, or due to failure of the rank-1 approximation itself[1]. However, when compared to the catastrophic failure to converge for the non-stoquastic Heisenberg model (section 6.5.3), it is surprising that the NQS is still able to perform quite well for N=30 non-stoquastic full connected Hamiltonians.

Although the N=30 QBM is not expected to outperform a classical BM with current RMS errors, a comparison could be done by e.g. determining the predictive powers of the classical and quantum BM, when trained on a subset of retinal activity data.

This leads us to the conclusion that the answer to the central research question posed in the introduction of this thesis is a provisional "Yes, but with $\mathcal{O}(10^{-1})$ errors in quantum statistics". Further research is needed to determine if different NQS architectures and different data sets might be suitable to bring the RMS error down. Furthermore, the stoquastic wall was shown to impede the N=10 QBM considerably. Alleviating the NQS-QBM of the stoquastic wall requires more investigation w.r.t. learning neural-net representations of signed wavefunctions.

[1] Or even a combination of both.

Chapter 8

Outlook

Training Quantum Boltzmann Machines with classical software requires a lot of resources. For systems with dozens of qubits, the complete spectrum of a density matrix cannot be calculated. Using the rank-1 approximation, the required spectral information is truncated to only the ground state. This ground state was found by the variational optimization of a Neural Quantum state. Since the parameterization of the QBM Hamiltonian is not restricted in any way, the NQS must be able to represent any ground state of 2-local spin Hamiltonians with external fields in x, y and z-directions.

However, the NQS was shown to fail for small spectral gap and the non-stoquastic Heisenberg model. For non-stoquastic fully connected Hamiltonians, however, the NQS performed surprisingly well for N=30. The exact reason for worse convergence of the NQS is still unclear. Although ground states of stoquastic (gapped) Hamiltonians were found reliably, the class of non-stoquastic Hamiltonians for which the NQS fails might be narrowed down further.

When the NQS was applied to the QBM, the problem with non-stoquastic Hamiltonians was circumvented by adding a stoquastic wall. However, gappedness can not be controlled. For this reason, the large QBM could not be trained accurately with the rank-1 approximation. Moreover, the restriction imposed by the barrier seems to reduce the (best possible) quality of the QBM representation significantly.

Determining if a Hamiltonian is gapped is an open problem in physics. In particular, the problem is known to be *undecidable* [75]: “There exists no algorithm to determine whether an arbitrary model is gapped or gapless, and . . . there exist models for which the presence or absence of a spectral gap is independent of the axioms of mathematics.” A general algorithm to steer the QBM away from gapless[1] Hamiltonians requires a significant breakthrough. Additionally, the considered classical data sets showed long-range correlations in their classical/quantum statistics. Long-range correlations are often linked to quantum phase transitions and gaplessness, which are notoriously hard to capture with an NQS. See [76] for a discussion particularly related to quantum criticality and the performance of an NQS with an RBM ansatz. Getting rid of the rank-1 approximation entirely without quantum hardware might be possible in the future, seeing as research is also done to design algorithms/architectures to embed mixed states with neural networks [77].

Irrespective of gappedness and criticality, the accuracy of the quantum statistics

[1] Note that in [75], gappedness is defined as having a continuous spectrum above the ground state energy in the thermodynamic limit in addition to the absence of a spectral gap.

sampled with the NQS was found to deteriorate significantly with increasing system sizes for the 1D Heisenberg model. This problem might be alleviated by experimenting with different architectures viz. FFNNs and Deep Boltzmann Machines. Although the goal in this thesis was to train a QBM instead of performing direct tomography, inspiration for improving the quality of quantum state embeddings and the statistics inferred from them might be found in NQS tomography literature such as [78] and [60].

As was discussed in the Introduction, ideally one would like to perform quantum algorithms on quantum hardware. While systems with dozens of spins are still out of reach, variational wavefunction optimization has been performed on quantum hardware using Variational Quantum Eigensolvers (*VQE*) [12]. In the QBM learning algorithm, the VQE would be a hardware replacement of the Neural Quantum State. Similar to classical neural nets, the representative power of quantum circuits depends on the architecture. Moreover, the efficacy of quantum circuits also depends on their ability to control noise. The design of noise-resilient quantum circuits with great representative power is an ongoing field of research [79]. Moreover, efforts are made to extend quantum hardware implementations of ground state VQEs to learning thermal (mixed) states [80]. Such a device could be the ultimate solution for training a QBM on hardware in the future.

Appendix A

Proof for Riemannian steepest descent

Set $d\mathbf{w} = \epsilon\mathbf{a}$, and search for the a that minimizes

$$L(\mathbf{w} + d\mathbf{w}) = L(\mathbf{w}) + \epsilon\nabla L(\mathbf{w})^T\mathbf{a} \tag{A.1}$$

under the constraint

$$|\mathbf{a}|^2 = \sum g_{ij}a_ia_j = \mathbf{a}^TG\mathbf{a} = 1. \tag{A.2}$$

The Lagrangean equation for this constraint optimization problem becomes

$$\frac{\partial}{\partial a_i}\Big(\nabla L(\mathbf{w})^T\mathbf{a} - \lambda\mathbf{a}^TG\mathbf{a}\Big) = 0. \tag{A.3}$$

Component-wise this gives the equation

$$\begin{aligned}\nabla L(\mathbf{w})_i &= \lambda\frac{\partial}{\partial a_i}\sum_{i,j} a_ig_{ij}a_j && \text{(A.4)}\\ &= \lambda(\sum_j g_{ij}a_j + \sum_j a_jg_{ij}) && \text{(A.5)}\\ &= 2\lambda\sum_j g_{ij}a_j && \text{(A.6)}\\ &= 2\lambda G_{\text{i-th row}}\cdot\mathbf{a}. && \text{(A.7)}\end{aligned}$$

Vector-wise this reduces to

$$\nabla L(\mathbf{w}) = 2\lambda G\mathbf{a}, \tag{A.8}$$

or

$$\mathbf{a} = \frac{1}{2\lambda}G^{-1}\nabla L(\mathbf{w}), \tag{A.9}$$

where λ is determined from the constraint.

The natural gradient of L in Riemannian space is denoted

$$\tilde{\nabla}L(\mathbf{w}) = G^{-1}\nabla L(\mathbf{w}). \tag{A.10}$$

Appendix B

N=100 XXX AFH model

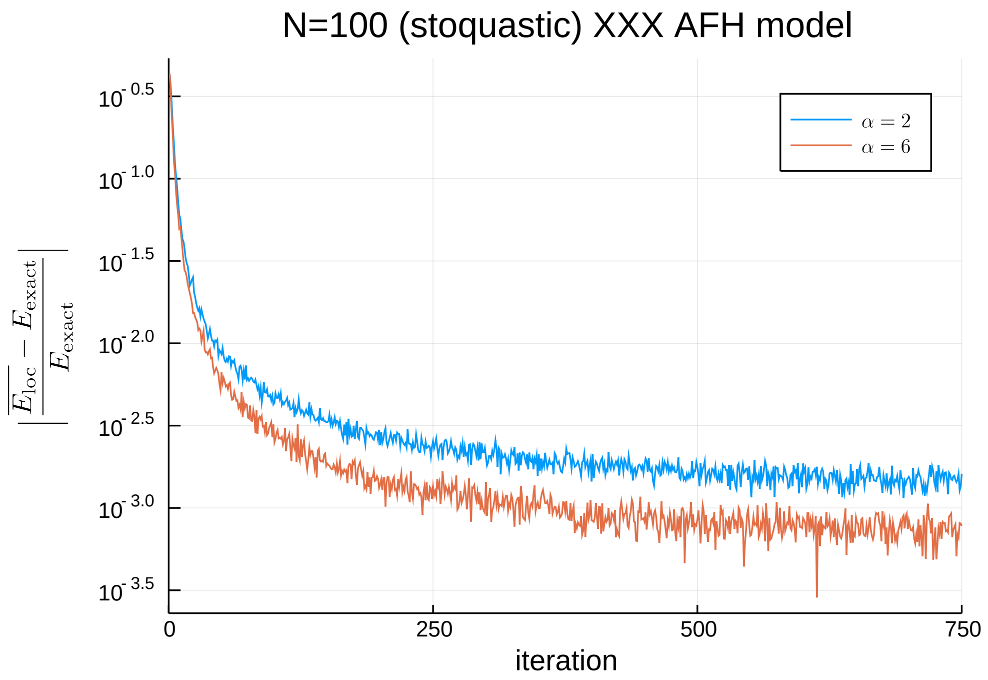


Figure B.1: Energy error decreasing smoothly as a function of iteration for the stoquastic AFH model. Learning rate $\eta = 0.05$ with decay factor $\gamma = 0.997$. Spin $\sum_i s_i = 0$-sector sampling with single flips on 40 CPUs each collecting 100 samples per optimization. $\alpha = 6$ reduces the error in the energy significantly.

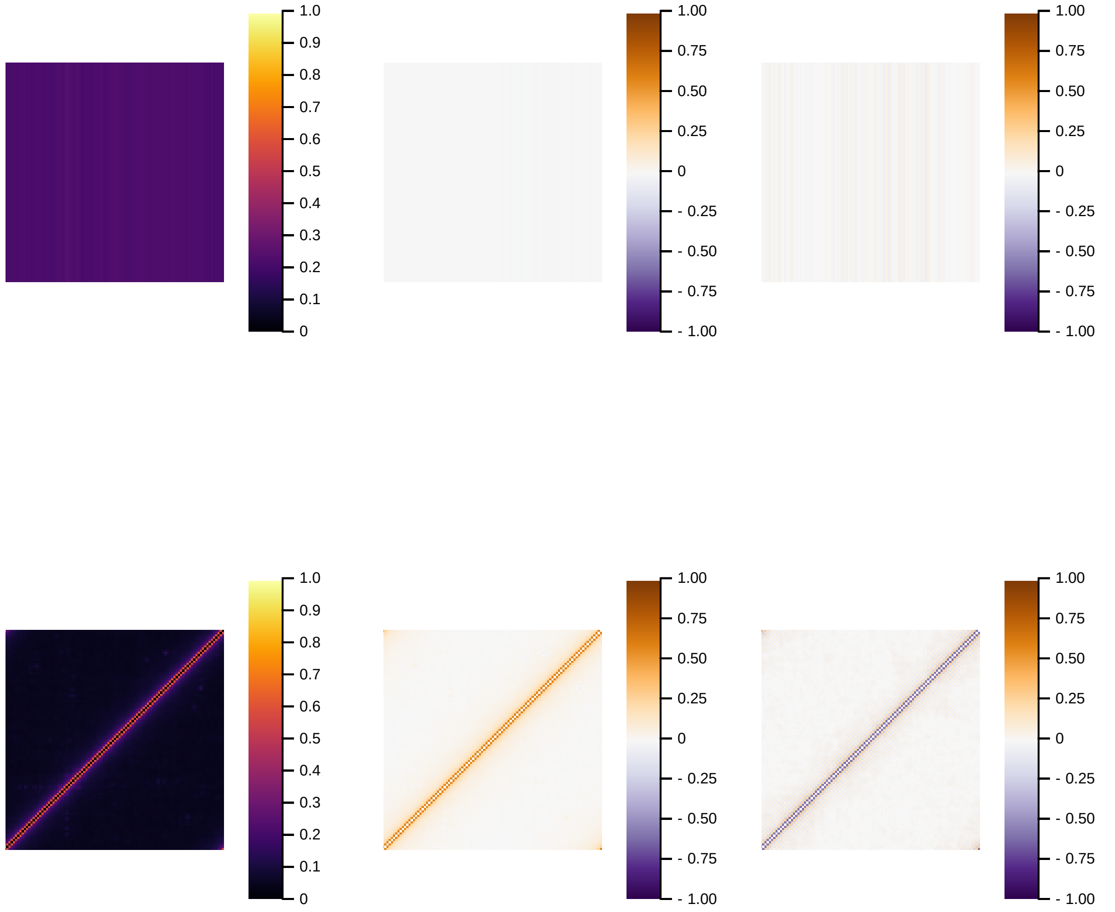


Figure B.2: In natural reading order, the: $\langle \sigma_i^x \rangle$, $\langle \sigma_i^y \rangle$, $\langle \sigma_i^z \rangle$, $\left\langle \sigma_i^x \sigma_j^x \right\rangle$, $\left\langle \sigma_i^y \sigma_j^y \right\rangle$ and $\left\langle \sigma_i^z \sigma_j^z \right\rangle$ estimations of statistics of the NQS for $\alpha = 3$ after optimization. Sample averages of the combination of 25 CPUs collecting 100 samples each. Note that these results are not done with a $S = 0$ sampler, since this would result in bad quantum statistics. Note that $\widehat{\langle \sigma^x \rangle} \approx 0.2$ (while the exact $\langle \sigma^x \rangle = 0$), indicating that this quantum statistic is particularly hard to learn for the NQS. These results are gathered under an hour of user time on ordinary cluster hardware (E5-2698 @ 2.20GHz).

Bibliography

[1] A. Mayor. *Gods and Robots: Myths, Machines, and Ancient Dreams of Technology.* Princeton University Press, 2018.

[2] M. Chen et al. "Disease prediction by machine learning over big data from healthcare communities". In: *Ieee Access* 5 (2017), pp. 8869–8879.

[3] M. I. Jordan and T. M. Mitchell. "Machine learning: Trends, perspectives, and prospects". In: *Science* 349.6245 (2015), pp. 255–260.

[4] D. Silver et al. "Mastering the game of Go with deep neural networks and tree search". In: *nature* 529.7587 (2016), p. 484.

[5] *Introducing a Brain-inspired Computer.* `http://www.research.ibm.com/articles/brain-chip.shtml`. Accessed on 24-11-19.

[6] N. Jones. "How to stop data centres from gobbling up the world's electricity." In: *Nature* 561.7722 (2018), p. 163.

[7] M. Davies et al. "Loihi: A neuromorphic manycore processor with on-chip learning". In: *IEEE Micro* 38.1 (2018), pp. 82–99.

[8] *Safety-first AI for autonomous data centre cooling and industrial control.* `https://deepmind.com/blog/article/safety-first-ai-autonomous-data-centre-cooling-and-industrial-control`. Accessed on 24-10-19.

[9] P. W. Shor. "Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer". In: *SIAM review* 41.2 (1999), pp. 303–332.

[10] M. H. Amin et al. "Quantum Boltzmann Machine". In: *Phys. Rev. X* 8 (2 May 2018), p. 021050.

[11] A. Wietek and A. M. Läuchli. "Sublattice coding algorithm and distributed memory parallelization for large-scale exact diagonalizations of quantum many-body systems". In: *Phys. Rev. E* 98 (3 Sept. 2018), p. 033309. DOI: `10.1103/PhysRevE.98.033309`. URL: `https://link.aps.org/doi/10.1103/PhysRevE.98.033309`.

[12] A. Peruzzo et al. "A variational eigenvalue solver on a photonic quantum processor". In: *Nature communications* 5 (2014), p. 4213.

[13] S. McArdle et al. "Variational ansatz-based quantum simulation of imaginary time evolution". In: *npj Quantum Information* 5.1 (2019), pp. 1–6.

[14] G. Carleo and M. Troyer. "Solving the quantum many-body problem with artificial neural networks". In: *Science* 355.6325 (2017), pp. 602–606.

[15] J. Bezanson et al. "Julia: A fresh approach to numerical computing". In: *SIAM review* 59.1 (2017), pp. 65–98. URL: `https://doi.org/10.1137/141000671`.

[16] M. Compen. *NeuralQuantumState.jl*. 2019. URL: https://github.com/mcompen/NeuralQuantumState.jl.

[17] E. Merzbacher. *Quantum Mechanics*. Wiley, 1998. ISBN: 9780471887027.

[18] K. Husimi. “Some Formal Properties of the Density Matrix”. In: *Proceedings of the Physico-Mathematical Society of Japan. 3rd Series* 22.4 (1940), pp. 264–314. DOI: 10.11429/ppmsj1919.22.4_264.

[19] J. v. Neumann. “Thermodynamik quantenmechanischer Gesamtheiten”. ger. In: *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse* 1927 (1927), pp. 273–291. URL: http://eudml.org/doc/59231.

[20] J. von Neumann. *Mathematical Foundations of Quantum Mechanics: New Edition*. Princeton university press, 2018.

[21] E. Schrödinger. “Discussion of probability relations between separated systems”. In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 31. 4. Cambridge University Press. 1935, pp. 555–563.

[22] J. Bub. “Quantum Entanglement and Information”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Spring 2019. Metaphysics Research Lab, Stanford University, 2019.

[23] J. S. Bell. “On the einstein podolsky rosen paradox”. In: *Physics Physique Fizika* 1.3 (1964), p. 195.

[24] R. F. Werner. “Quantum states with Einstein-Podolsky-Rosen correlations admitting a hidden-variable model”. In: *Phys. Rev. A* 40 (8 Sept. 1989), pp. 4277–4281. DOI: 10.1103/PhysRevA.40.4277. URL: https://link.aps.org/doi/10.1103/PhysRevA.40.4277.

[25] J. Barrett. “Nonsequential positive-operator-valued measurements on entangled mixed states do not always violate a Bell inequality”. In: *Physical Review A* 65.4 (2002), p. 042302.

[26] K. Życzkowski et al. “Dynamics of quantum entanglement”. In: *Physical Review A* 65.1 (2001), p. 012101.

[27] J. Eisert, M. Cramer, and M. B. Plenio. “Colloquium: Area laws for the entanglement entropy”. In: *Rev. Mod. Phys.* 82 (1 Feb. 2010), pp. 277–306. DOI: 10.1103/RevModPhys.82.277. URL: https://link.aps.org/doi/10.1103/RevModPhys.82.277.

[28] R. Horodecki et al. “Quantum entanglement”. In: *Reviews of modern physics* 81.2 (2009), p. 865.

[29] J. P. Keating and F. Mezzadri. “Entanglement in Quantum Spin Chains, Symmetry Classes of Random Matrices, and Conformal Field Theory”. In: *Phys. Rev. Lett.* 94 (5 Feb. 2005), p. 050501. DOI: 10.1103/PhysRevLett.94.050501. URL: https://link.aps.org/doi/10.1103/PhysRevLett.94.050501.

[30] S. Chakravarty, B. I. Halperin, and D. R. Nelson. “Low-temperature behavior of two-dimensional quantum antiferromagnets”. In: *Physical review letters* 60.11 (1988), p. 1057.

[31] A. Auerbach. *Graduate Texts in Contemporary Physics: Interacting Electrons and Quantum Magnetism.* 1994.

[32] L. Capriotti et al. “Quantum Effects and Broken Symmetries in Frustrated Antiferromagnets”. In: (2000).

[33] C. D. Meyer. *Matrix analysis and applied linear algebra.* Vol. 71. Siam, 2000.

[34] S. R. White. “Density-matrix algorithms for quantum renormalization groups”. In: *Physical Review B* 48.14 (1993), p. 10345.

[35] H. A. Bethe. “Statistical theory of superlattices”. In: *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences* 150.871 (1935), pp. 552–575.

[36] M. Karbach et al. “Introduction to the Bethe Ansatz I”. In: *Computers in Physics* 11.1 (1997), pp. 36–43. DOI: 10.1063/1.4822511.

[37] N. Metropolis and S. Ulam. “The monte carlo method”. In: *Journal of the American statistical association* 44.247 (1949), pp. 335–341.

[38] N. Metropolis et al. “The beginning of the Monte Carlo method”. In: *Los Alamos Science* 15.584 (1987), pp. 125–130.

[39] D. J. C. MacKay. *Information Theory, Inference & Learning Algorithms.* New York: Cambridge University Press, 2002. ISBN: 0521642981.

[40] W. R. Gilks, S. Richardson, and D. Spiegelhalter. *Markov chain Monte Carlo in practice.* Chapman and Hall/CRC, 1995.

[41] D. P. Landau and K. Binder. *A Guide to Monte Carlo Simulations in Statistical Physics.* 4th ed. Cambridge University Press, 2014.

[42] D. A. Levin and Y. Peres. *Markov chains and mixing times.* Vol. 107. American Mathematical Soc., 2017.

[43] M. K. Cowles and B. P. Carlin. “Markov chain Monte Carlo convergence diagnostics: a comparative review”. In: *Journal of the American Statistical Association* 91.434 (1996), pp. 883–904.

[44] K. P. Murphy. *Machine learning: a probabilistic perspective.* MIT press, 2012.

[45] J. Schmidhuber. “Deep learning in neural networks: An overview”. In: *Neural networks* 61 (2015), pp. 85–117.

[46] W. S. McCulloch and W. Pitts. “A logical calculus of the ideas immanent in nervous activity”. In: *The bulletin of mathematical biophysics* 5.4 (1943), pp. 115–133.

[47] D. O. Hebb. *The organization of behavior.* New York: Wiley, 1949.

[48] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. “A learning algorithm for Boltzmann machines”. In: *Cognitive science* 9.1 (1985), pp. 147–169.

[49] G. E. Hinton. “Training products of experts by minimizing contrastive divergence”. In: *Neural computation* 14.8 (2002), pp. 1771–1800.

[50] H. C. Nguyen, R. Zecchina, and J. Berg. “Inverse statistical problems: from the inverse Ising problem to data science”. In: *Advances in Physics* 66.3 (2017), pp. 197–261.

[51] T. J. Sejnowski. “Higher-order Boltzmann machines”. In: *AIP Conference Proceedings*. Vol. 151. 1. AIP. 1986, pp. 398–403.

[52] X. Gao and L.-M. Duan. “Efficient representation of quantum many-body states with deep neural networks”. In: *Nature communications* 8.1 (2017), p. 662.

[53] M. Suzuki. “Relationship between d-Dimensional Quantal Spin Systems and (d+1)-Dimensional Ising Systems: Equivalence, Critical Exponents and Systematic Approximants of the Partition Function and Spin Correlations”. In: *Progress of Theoretical Physics* 56.5 (Nov. 1976), pp. 1454–1469.

[54] H. J. Kappen. “Learning quantum models from quantum or classical data”. In: *arXiv preprint arXiv:1803.11278* (2019).

[55] E. Carlen. “Trace inequalities and quantum entropy: an introductory course”. In: *Entropy and the quantum* 529 (2010), pp. 73–140.

[56] I. Bengtsson and K. Życzkowski. *Geometry of quantum states: an introduction to quantum entanglement*. Cambridge university press, 2017.

[57] M. Schuld and F. Petruccione. *Supervised Learning with Quantum Computers*. Vol. 17. Springer, 2018.

[58] K. M. R. Audenaert. “Telescopic Relative Entropy”. In: *Theory of Quantum Computation, Communication, and Cryptography*. Ed. by D. Bacon, M. Martin-Delgado, and M. Roetteler. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 39–52. ISBN: 978-3-642-54429-3.

[59] Z.-X. Gong et al. “Entanglement area laws for long-range interacting systems”. In: *Physical review letters* 119.5 (2017), p. 050501.

[60] G. Torlai et al. “Neural-network quantum state tomography”. In: *Nature Physics* 14.5 (2018), p. 447.

[61] D.-L. Deng, X. Li, and S. D. Sarma. “Quantum entanglement in neural network states”. In: *Physical Review X* 7.2 (2017), p. 021021.

[62] N. Le Roux and Y. Bengio. “Representational power of restricted Boltzmann machines and deep belief networks”. In: *Neural computation* 20.6 (2008), pp. 1631–1649.

[63] G. Carleo et al. “Machine learning and the physical sciences”. In: (2019). arXiv: `1903.10563 [physics.comp-ph]`.

[64] “Chapter Fifteen - Introduction to the Variational and Diffusion Monte Carlo Methods”. In: *Electron Correlation in Molecules – ab initio Beyond Gaussian Quantum Chemistry*. Ed. by P. E. Hoggan and T. Ozdogan. Vol. 73. Advances in Quantum Chemistry. Academic Press, 2016, pp. 285–314. DOI: `https://doi.org/10.1016/bs.aiq.2015.07.003`. URL: `http://www.sciencedirect.com/science/article/pii/S0065327615000386`.

[65] F. Becca and S. Sorella. *Quantum Monte Carlo Approaches for Correlated Systems*. Cambridge University Press, 2017. DOI: `10.1017/9781316417041`.

[66] J. Klassen and B. M. Terhal. “Two-local qubit Hamiltonians: when are they stoquastic?” In: *arXiv preprint arXiv:1806.05405* (2018).

[67] T. Westerhout et al. “Neural Quantum States of frustrated magnets: generalization and sign structure”. In: *arXiv preprint arXiv:1907.08186* (2019).

[68] T. Vieijra. “Machine learning approaches to strongly correlated spin systems”. PhD thesis. Universiteit Gent, 2019.

[69] I. Glasser et al. “Neural-Network Quantum States, String-Bond States, and Chiral Topological States”. In: *Phys. Rev. X* 8 (1 Jan. 2018), p. 011006. DOI: 10.1103/PhysRevX.8.011006. URL: https://link.aps.org/doi/10.1103/PhysRevX.8.011006.

[70] S.-I. Amari. “Natural Gradient Works Efficiently in Learning”. In: *Neural Comput.* 10.2 (Feb. 1998), pp. 251–276. ISSN: 0899-7667. DOI: 10.1162/089976698300017746. URL: http://dx.doi.org/10.1162/089976698300017746.

[71] J. P. Provost and G. Vallee. “Riemannian structure on manifolds of quantum states”. In: *Communications in Mathematical Physics* 76.3 (Sept. 1980), pp. 289–301. DOI: 10.1007/BF02193559. URL: https://doi.org/10.1007/BF02193559.

[72] S. Sorella, M. Casula, and D. Rocca. “Weak binding between two aromatic rings: Feeling the van der Waals attraction by quantum Monte Carlo methods”. In: *The Journal of chemical physics* 127.1 (2007), p. 014105.

[73] O. Bleu et al. “Effective Theory of Nonadiabatic Quantum Evolution Based on the Quantum Geometric Tensor”. In: *Phys. Rev. Lett.* 121 (2 July 2018), p. 020401. DOI: 10.1103/PhysRevLett.121.020401. URL: https://link.aps.org/doi/10.1103/PhysRevLett.121.020401.

[74] G. Carleo et al. “NetKet: A Machine Learning Toolkit for Many-Body Quantum Systems”. In: *SoftwareX* (2019), p. 100311. DOI: 10.1016/j.softx.2019.100311. URL: http://www.sciencedirect.com/science/article/pii/S2352711019300974.

[75] T. S. Cubitt, D. Perez-Garcia, and M. M. Wolf. “Undecidability of the spectral gap”. In: *Nature* 528.7581 (2015), p. 207.

[76] D. Sehayek et al. “Learnability scaling of quantum states: Restricted Boltzmann machines”. In: *Physical Review B* 100.19 (2019), p. 195125.

[77] F. Vicentini et al. “Variational Neural-Network Ansatz for Steady States in Open Quantum Systems”. In: *Phys. Rev. Lett.* 122 (25 June 2019), p. 250503. DOI: 10.1103/PhysRevLett.122.250503. URL: https://link.aps.org/doi/10.1103/PhysRevLett.122.250503.

[78] J. Carrasquilla et al. “Reconstructing quantum states with generative models”. In: *Nature Machine Intelligence* 1.3 (2019), p. 155.

[79] A. G. Rattew et al. “A Domain-agnostic, Noise-resistant Evolutionary Variational Quantum Eigensolver for Hardware-efficient Optimization in the Hilbert Space”. In: *arXiv preprint arXiv:1910.09694* (2019).

[80] G. Verdon et al. “Quantum Hamiltonian-Based Models and the Variational Quantum Thermalizer Algorithm”. In: *arXiv preprint arXiv:1910.02071* (2019).