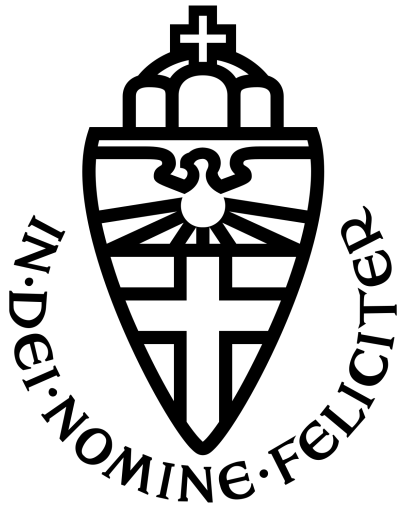
NQS of frustrated magnets: generalization and sign structure

N. Astrakhantsev

in collaboration with

A. Bagrov, K. Tikhonov, T. Westerhout



Topological materials group meeting

September 20, 2019, Zurich

The problem

- Consider the spin$-1/2$ AFM (or essentially any spin model) on a graph G:

$$\hat{H} = \sum_{\langle i,j \rangle} J_{ij} \vec{\sigma}_i \vec{\sigma}_j .$$

- The sum runs over the pairs of spins (nearest neighbors, possibly further) $\langle i, j \rangle$, $\vec{\sigma}_i$ is the Pauli matrices vector.
- The Hilbert space of the model consists of all possible «strings» of the form $|\uparrow\downarrow\downarrow \ldots\rangle$ is 2^N–dimensional. The ground state (GS) wave-function is a set of 2^N complex numbers.
- If the system is frustrated, the ground state is highly degenerate and is probably the so-called quantum spin liquid.
- One of the possible approximate ground states is the RVB, though the problem is *exponentially hard* and the solution is yet unknown.

Symmetries?

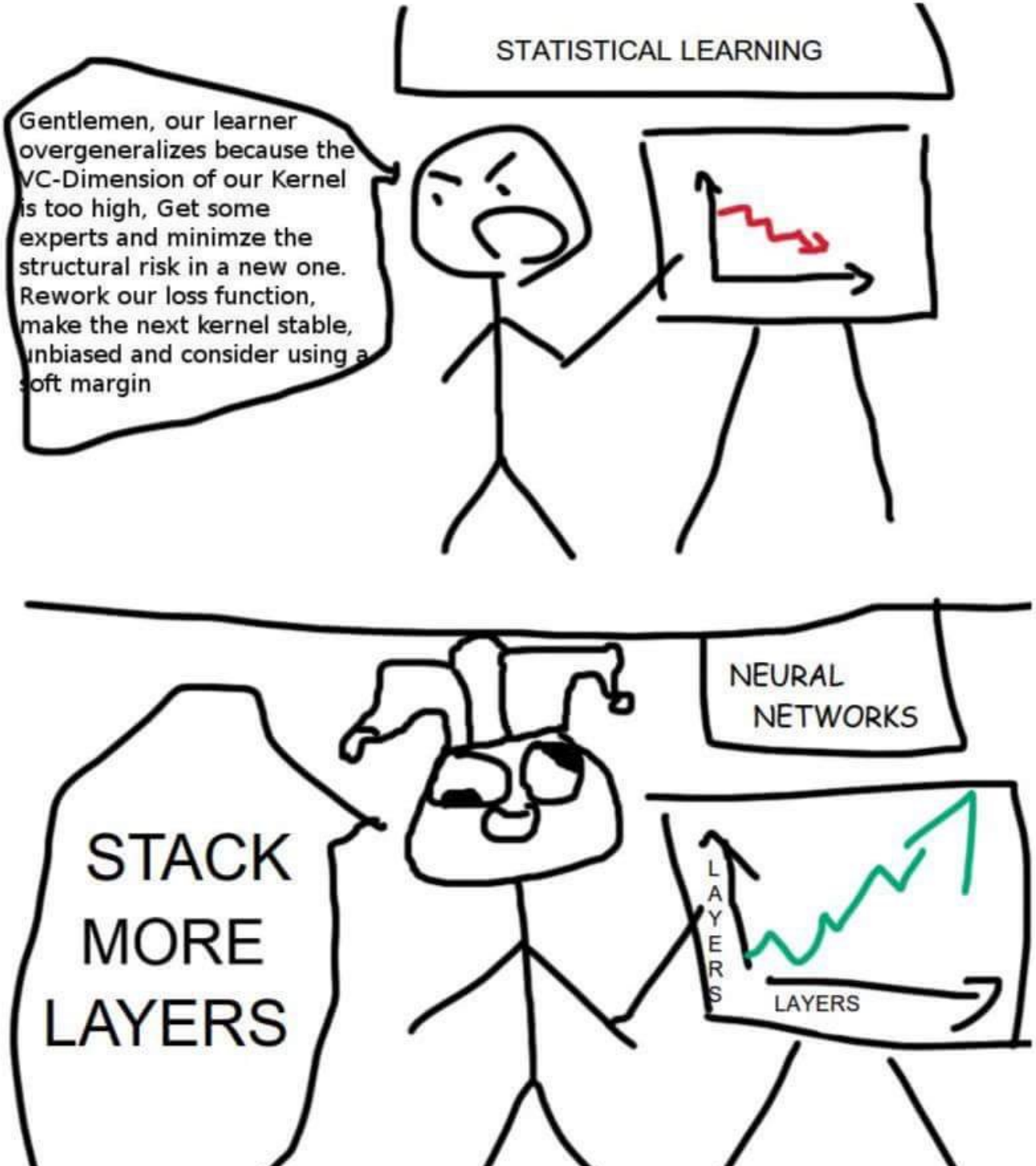
- The ground state preserves the total magnetization, so all the configurations entering the GS should have the same magnetization. In fact, the smallest possible (0 or 1). This reduces the Hilbert space dimensionality by the factor of $\sim \mathrm{Binom}(N/2, N)$.
- If the periodic BC are applied on, say, square lattice, than the total momentum is also conserved. All the configurations matching via some translation, have the same phase and amplitude. This further reduces the dimensionality by the factor of $\sim N$.
- Anyway, $2^N/(N\mathrm{Binom}(N/2, N))$ is still exponentially hard.

Carleo & Troyer's idea

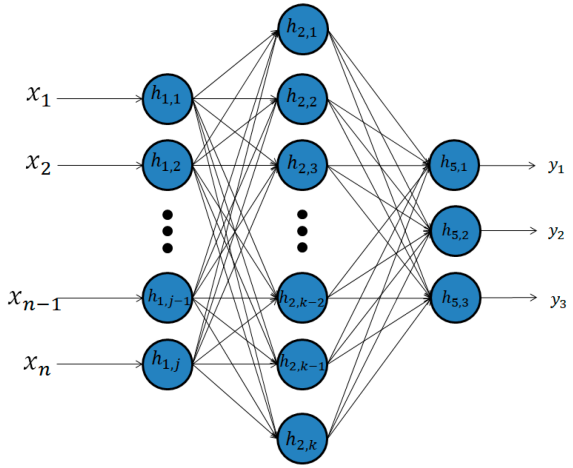
- Let us apply the good-old *variational approach*. Let $\psi(w) : s \to \mathbb{C}$ be a complex-valued function of some parameters set w and the spin configuration $s \sim |\uparrow\downarrow\downarrow \ldots\rangle$.
- It is then obvious that

$$E_{gs} \leqslant \min_w \frac{\langle\psi(w)|\hat{H}|\psi(w)\rangle}{\langle\psi(w)|\psi(w)\rangle}$$

- If the choice of the w–parametrization is «good», the variational WF gives a close GS approximation.
- How do we choose the «good» approximation? One of the ways is the physics insight (which I would prefer).
- Why not pick the neural network?


STATISTICAL LEARNING
Gentlemen, our learner overgeneralizes because the VC-Dimension of our Kernel is too high, Get some experts and minimze the structural risk in a new one. Rework our loss function, make the next kernel stable, unbiased and consider using a soft margin
NEURAL NETWORKS
STACK MORE LAYERS
LAYERS
LAYERS

A lyrical digression: neural networks in a nutshell



- Neural networks take *features*: the initial vector of N numbers x_i and outputs the result of M numbers y_j.
- The transform is done layer-by-layer. Take all the features at i–th layer into one single vector $\vec{h}_i$ and say

$$\boxed{\vec{h}_{i+1} = \text{nonlinearity}(W^i h_i) + b_i}.$$

- The nonlinearity is essentially any function, b_i is the *bias*.
- The parameters are: matrices W^i and bias vectors b^i. They are somehow chosen to give the right predictions.
- NN is nothing but *a way to parametrically encode a very complicated function.*

Carleo & Troyer's idea

- The $\psi(w) : s \to \mathbb{C}$ is the neural network. It inputs the vector of N values $\{+1, +1, -1, -1, \ldots, +1\}$ and produces $\mathrm{Re}\psi(s)$ and $\mathrm{Im}\psi(s)$.
- The parameters w are *optimized* to produce the state with the smallest possible energy.
- The optimization in Deep Learning is usually done via the gradient descent technique:

$$w := w + h\frac{\partial \mathcal{L}}{\partial w},$$

where $\mathcal{L}$ is a functional to optimize (in our case, energy evaluated at the $\psi(w)$ state.

- Important assumption that we make here: *exponentially large WF can be encoded «good enough» with the polynomial number of parameters.*

Carleo & Troyer's idea

- Problem: in order to compute $\partial\mathcal{L}/\partial w$, we need to evaluate the sum containing *exponentially large number of terms*: the curse of dimensionality remanifests as another entity.
- To address this, C&T proposed a clever sum evaluation algorithm, which unfortunately makes one more assumption (which is only sometimes justified).

C&T's Monte-Carlo

We are evaluating the sum

$$E^{\text{var}} = \frac{1}{\mathcal{Z}} \sum_{s \in \mathbb{H}} |\psi(s)|^2 \langle s|\hat{H}|s\rangle.$$

C&T proposed to replace the whole sum with the sum over some restricted set of configurations $S = \{s_1, s_2, \ldots, s_n\}$ distributed with the canonical weight $s_i \sim |\psi(s)|^2/\mathcal{Z}$:

$$E^{\text{var}} \approx \sum_{s \in S} \langle s|\hat{H}|s\rangle.$$

C&T's assumption: *by looking at some small set of spin configurations S we can make a conclusion about the whole E^{var} and correctly evaluate the derivative $\partial\mathcal{L}/\partial w$.*

How the S set is generated?

The idea is close to the DQMC.

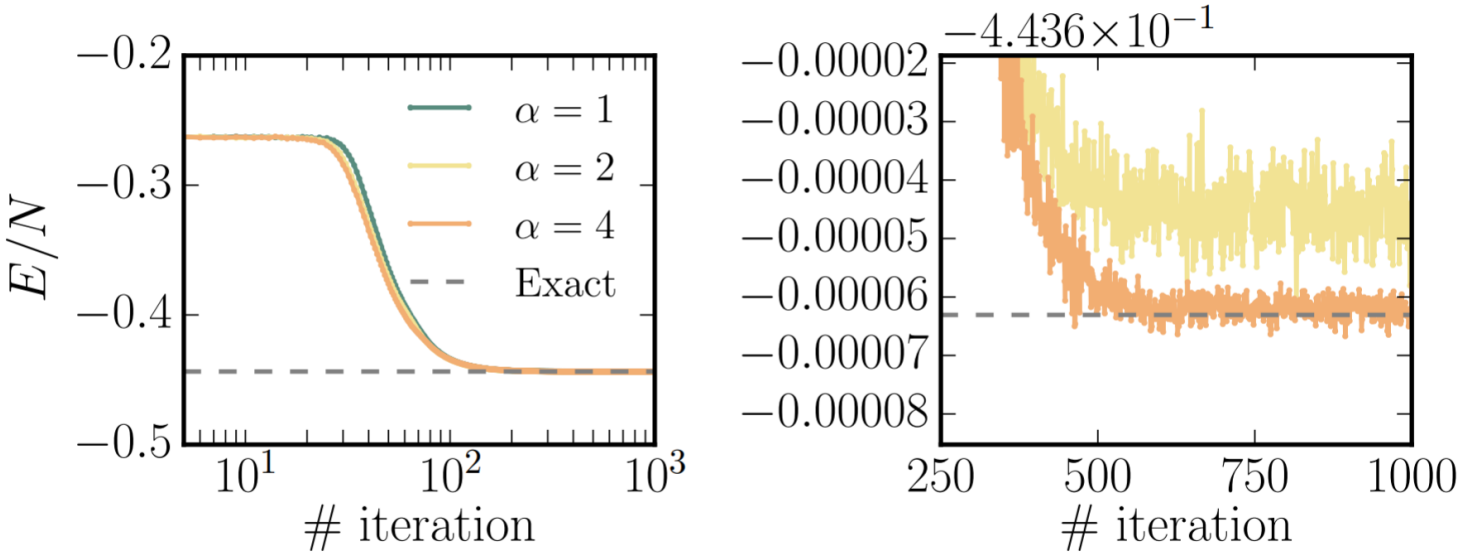
- Consider the current spin configuration s_τ.
- Perform several random spin flips. We accept or reject the final spin configuration $s_{\tau'}$ with the probability

$$P(s_\tau \to s_{\tau'}) = \min\left(1, \frac{|\psi(s_\tau)|^2}{|\psi(s_{\tau'})|^2}\right).$$

- It can be proven that in such a process the spin configurations will be distributed canonically $s_i \sim |\psi(s)|^2/\mathcal{Z}$.

The C&T procedure success

The C&T demonstrated that this approach can find the GS of the 10×10 square lattice AFM with the nearest-neighbor hoppings. The community went very excited.



Where have we cheated?[1]

[1] If we have not cheated it is mathematics, not physics.

The C&T main secret

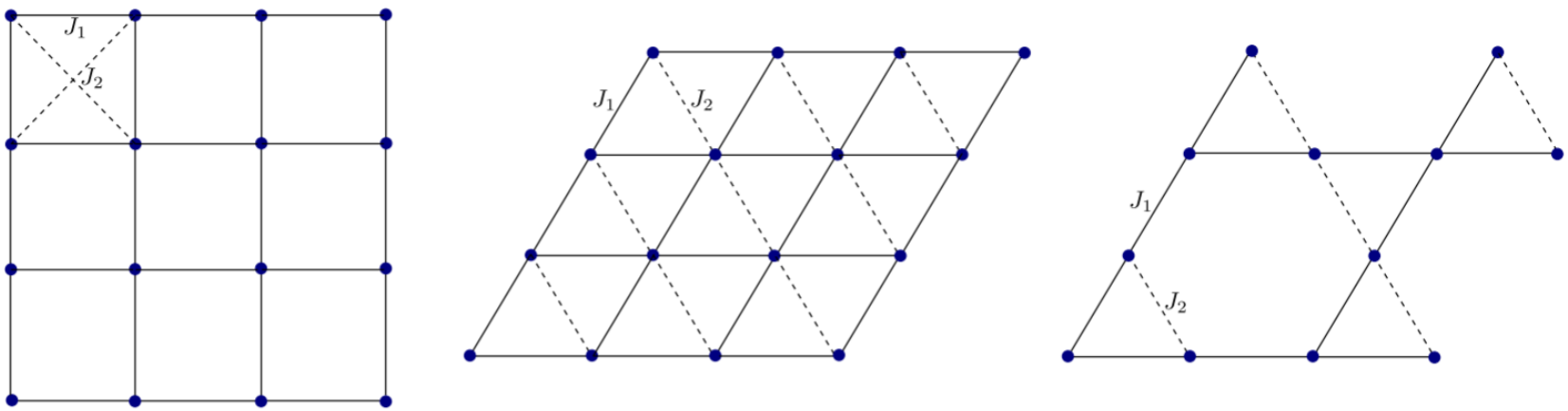
- The Heisenberg AFM Hamiltonian has positive off-diagonal terms
$$\langle \uparrow\downarrow | \hat{H} | \downarrow\uparrow \rangle > 0.$$
- The resulting GS wavefunction should be *very oscillating* so that for every neighboring spins, $\psi(\uparrow\downarrow)\psi^*(\downarrow\uparrow) < 0$. Only then the off-diagonal contribution will be negative.
- However, such oscillations are very hard for the NN to fit. It is more-or-less *smooth function*.
- Moreover, the oscillating sum in E^{var} is very hard to estimate (sign problem as is!).
- The solution: perform the unitary basis transform and flip the basis on the A-sublattice. The sign problem is basis-dependent! (this is only possible on the bipartite lattices). However, C&T have never mentioned this fact in their paper, but this transformation can be found deep within their Github repository.

But what if the lattice is not bipartite?

- The unitary transform can not be performed: the sign problem is present.
- On non-frustrated lattices only a small subset of the Hilbert space gives significant contribution. However, in the frustrated systems *this might not be the case*!
- What is even worse, all the significant states can be *different*: by looking at the small subset one is unable to draw conclusion about the rest set.

Our work: study the *generalization* of the neural networks

Consider three lattices with the main hopping J_1 (solid lines) and the frustration controlled by the J_2 hopping (dashed lines).



- Square lattice with next-to-nearest neighbors;
- Triangular lattice as is ($J_2 = 1$ corresponds to the famous triangular AFM lattice);
- the Kagome lattice;

The problem setup

- since all Hamiltonian matrix elements are real, the Hamiltonian can be diagonalized on $\mathbb{R}$, the GS has the form

$$|gs\rangle = A_1\sigma_1|s_1\rangle + \ldots + A_n\sigma_n|s_n\rangle,$$

  where A_i are the positive amplitudes, $\sigma_n = \pm 1$ are the signs and $n = \mathrm{Binom}(n, n/2)$, since all states have the same magnetization 0.
- empirically, the NN can be successfully transformed to have two outputs: $\log A$ and the probability to have the + sign.
- Learning of $\log A$ is relatively easy, the main problem is predicting the sign σ_i of the state $|\sigma_i\rangle$.

The problem setup

- we take the lattice of ~ 24 spins, with the $M = 0$ sector size of 2704156 vectors;
- we perform the exact diagonalization and obtain $\tilde{A}_i$, $\tilde{\sigma}_i$ exact amplitudes and signs;
- we consider amplitudes $A_i = \tilde{A}_i$ known (we do not learn them), but only try to learn phases σ_i to fit $\tilde{\sigma}_i$;
- to reproduce the C&T algorithm condicions, we sample $n_t = r_t \times 2704156$ vectors with the canonical probability $|\psi(s)|^2/\mathcal{Z}$, where $r_t \ll 1$; we train our network to correctly predict the signs on n_t training vectors and see how this prediction *generalizes* on all the 2704156 spins;
- essentially, this is the most typical machine learning problem: given some small subset of the existing data, one has to correctly make predictions for the whole set;

The problem setup

- we perform the gradient descent to minimize the *binary cross entropy*:

$$\mathcal{L} = -\sum_{i=1}^{n_t} \left(\frac{1+\tilde{\sigma}_i}{2} \log p_i + \frac{1-\tilde{\sigma}_i}{2} \log(1-p_i) \right),$$

where p_i is the probability of the $+$ sign given by our network; if $p_i > 1/2$ we assign the $+$ sign to the state, otherwise the $-$ sign;

- having performed the training on the n_t states, we make prediction on the whole state; the training accuracy is measured in terms of the *overlap* with the ground state;

$$\langle \psi_{nn} | \psi_{gs} \rangle = \frac{\sum_{i=1}^{N} |A_i|^2 \sigma_i \tilde{\sigma}_i}{\sqrt{\langle \psi_{nn} | \psi_{nn} \rangle} \sqrt{\langle \psi_{gs} | \psi_{gs} \rangle}},$$

where σ_i are the signs predicted by the NN, $\tilde{\sigma}_i$ are the exact diagonalization signs;

Square lattice

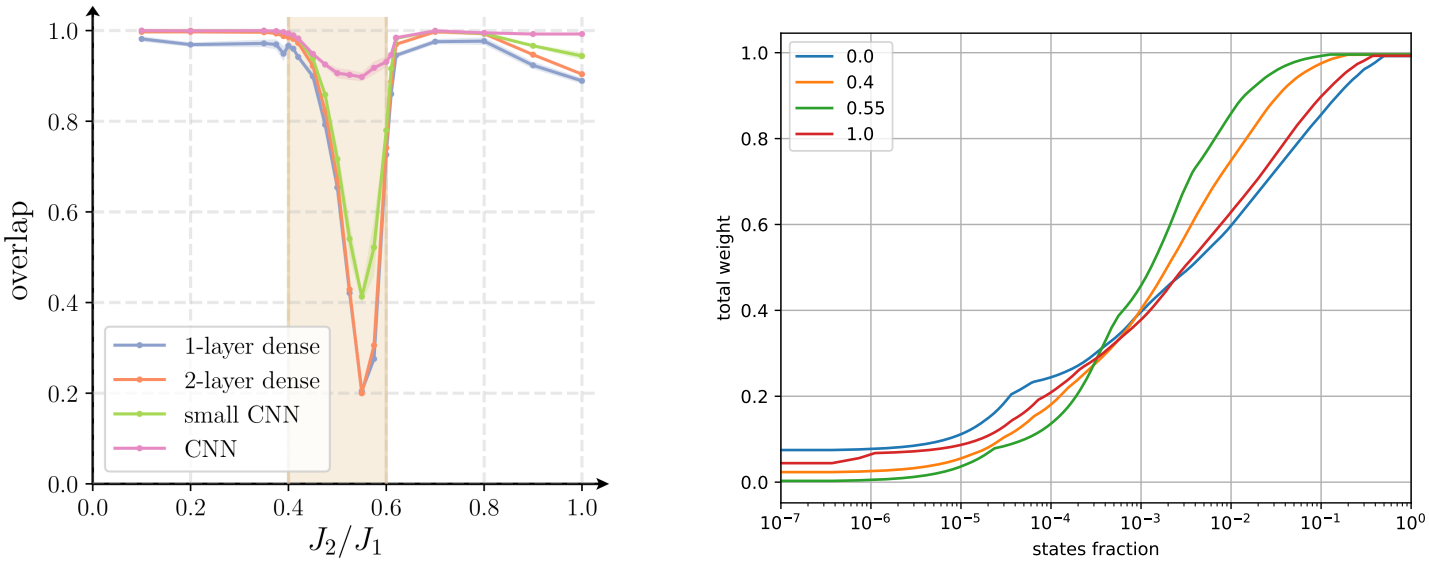


Figure: (Left): The square lattice dependence of overlap on the j_2/j_1 ratio at fixed $r_t = 0.01$. The frustrated region is shown in yellow. (Right): the cumulative amplitude $\sum_i |\psi_i|^2$ (y-axis) calculated at a fraction (x-axis) of the states with the smallest amplitudes. Note that in the j_2/j_1 case the small amplitude states become much more important (the distribution of amplitudes becomes very wide).

Square lattice

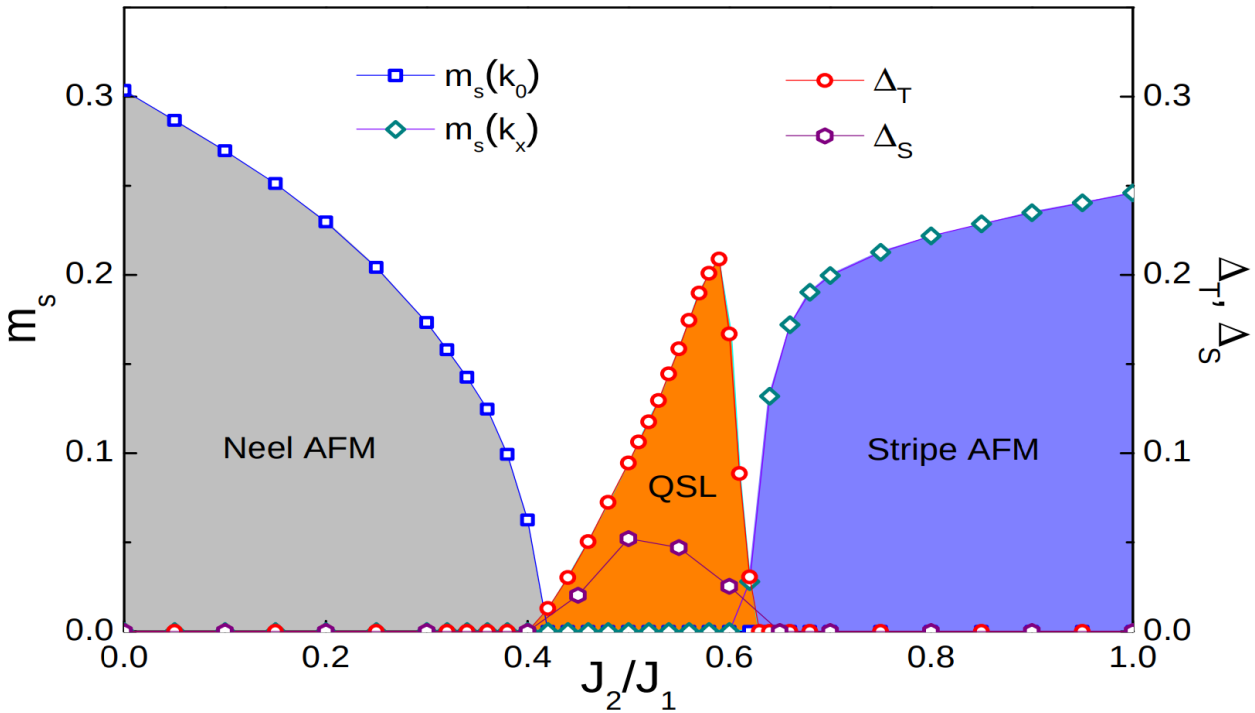


Figure: The ground state phase diagram for the spin 1/2 AFM Heisenberg $J1 - J2$ model on the square lattice, as determined by accurate DMRG calculations. Note that the region of our problems with generalization is precisely the QSL region on the phase diagram (complete disorder). AFM phases are predicted perfectly by the NN.

Kagome lattice

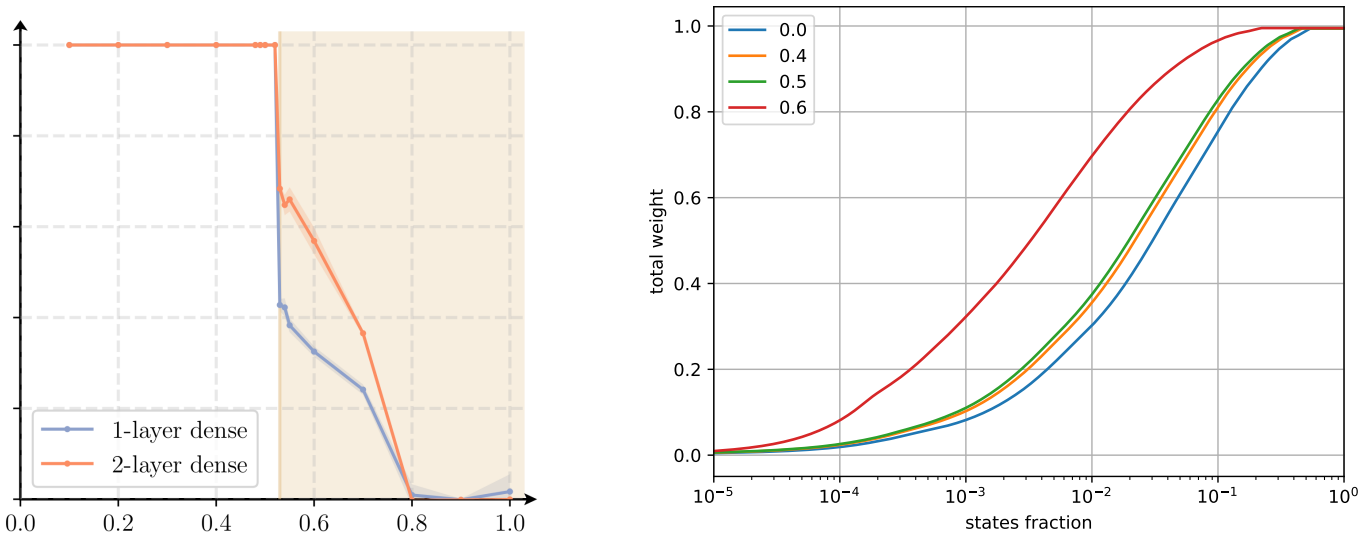


Figure: (Left): The kagome lattice dependence of overlap on the j_2/j_1 ratio at fixed $r_t = 0.01$. The frustrated region is shown in yellow. (Right): the cumulative amplitude $\sum_i |\psi_i|^2$ (y-axis) calculated at a fraction (x-axis) of the states with the smallest amplitudes.

Triangle lattice

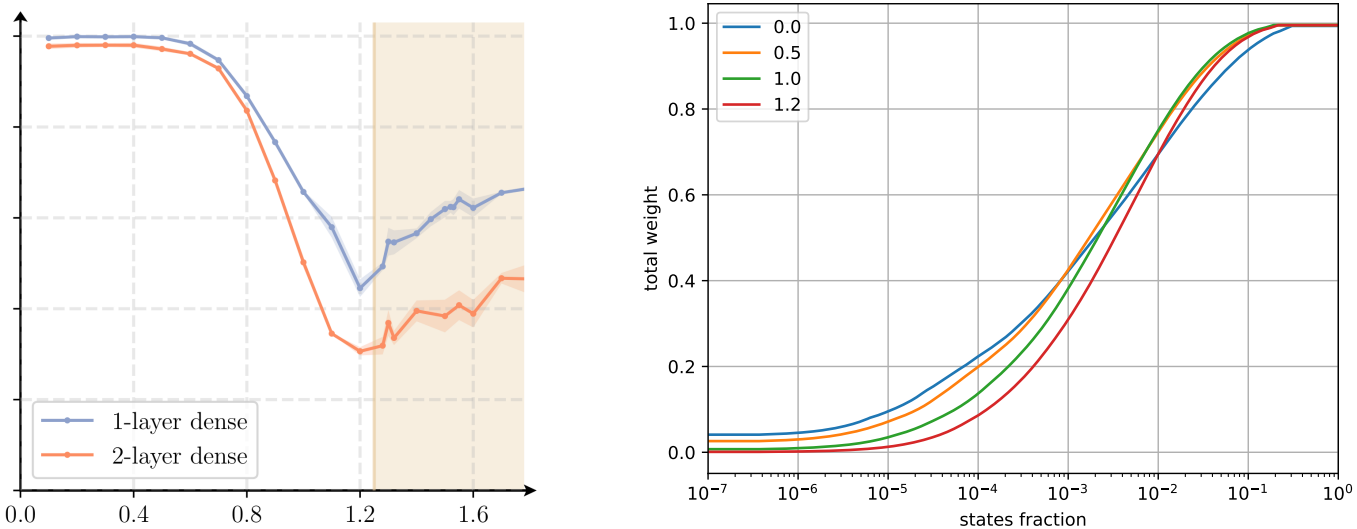


Figure: (Left): The triangle lattice dependence of overlap on the j_2/j_1 ratio at fixed $r_t = 0.01$. The frustrated region is shown in yellow. (Right): the cumulative amplitude $\sum_i |\psi_i|^2$ (y-axis) calculated at a fraction (x-axis) of the states with the smallest amplitudes.

Square lattice

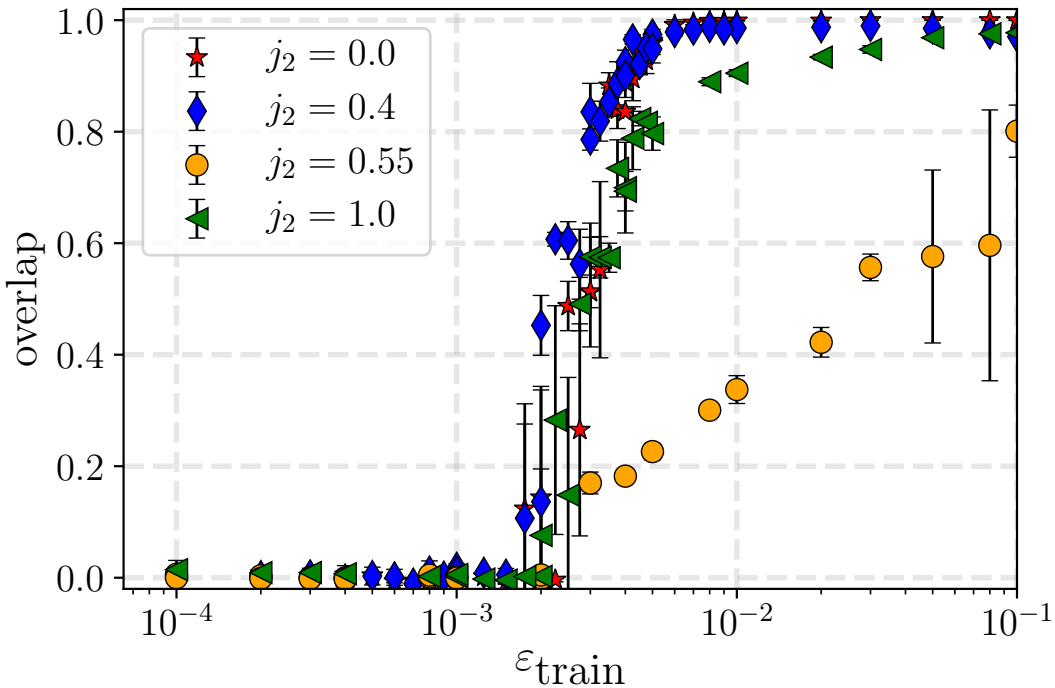


Figure: The square lattice dependence of the generalization capabiltiy on the r_t for various values of j_2/j_1. Note that the generalization capability has a phase-transition-like behavior overlap $= \sqrt{r_t - r_t^c}$.

Kagome lattice

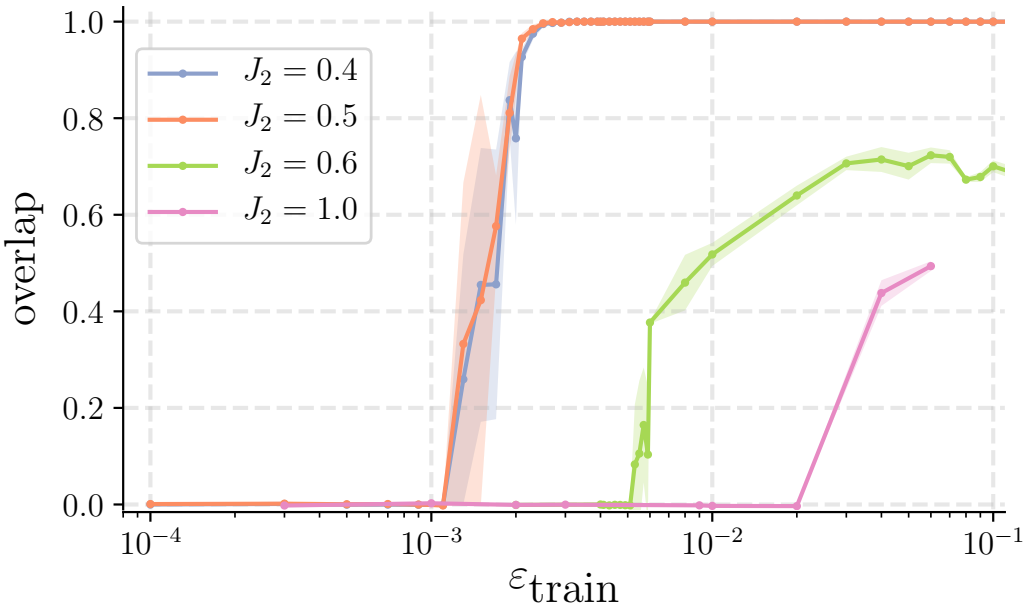


Figure: The kagome lattice dependence of the generalization capabiltiy on the r_t for various values of j_2/j_1. Note that the generalization capability has a phase-transition-like behavior overlap $= \sqrt{r_t - r_t^c}$.

Triangle lattice

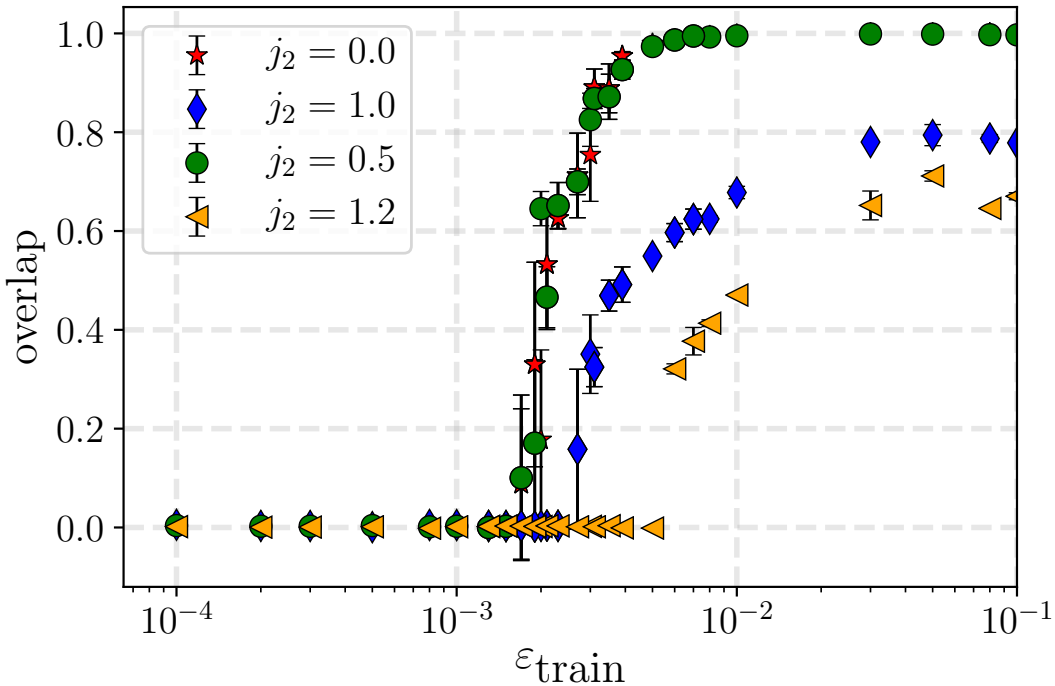


Figure: The triangle lattice dependence of the generalization capabiltiy on the r_t for various values of j_2/j_1. Note that the generalization capability has a phase-transition-like behavior overlap $= \sqrt{r_t - r_t^c}$.

Overlap versus phase

If we fix sign structure instead and fit the amplitudes, the problem is pronounced too, but not that much.

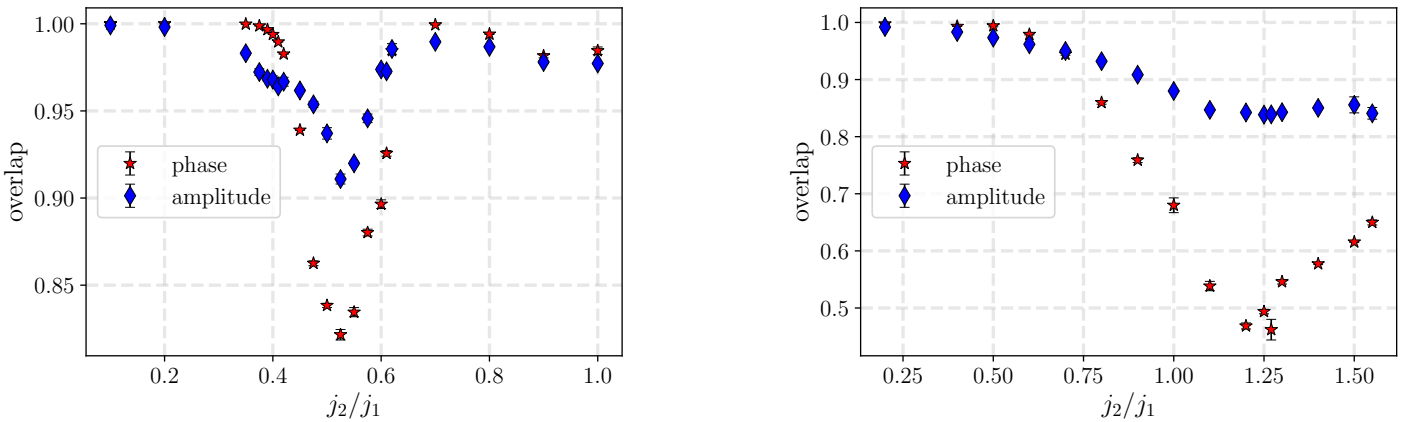


Figure: The square (left) and triangle (right) comparisons of the overlap obtained by sign structure fitting and amplitude fitting. Note that the problems with the amplitudes structure are much less severe.

Conclusions

- The C&T method is strong if the NN can extract the ground state structure from a very small subset of the states;
- in the AFM phases (crystal) this generalization is easy; however, in the quantum spin liquid phase this is much harder;
- the convolutional NN with the hard-coded translational symmetry showed very good results for the square lattice;
- the generalization drastically depends on the training size ratio r_t; while in the crystal phase $r_t \sim 10^{-3}$ is enough, it grows to 2×10^{-2} in the frustrated phase; on the 10×10 spin system studied by C&T, 10^{-2} spins would still mean 10^{27} samples required! So absence of frustrations is crucial;
- the sign problem is also present; only bipartite lattices can be studied reliably;