Hierarchical autoregressive approach to two-dimensional statistical systems

Piotr Korcyl

UNIWERSYTET
JAGIELLOŃSKI
W KRAKOWIE

work one with P. Białas and T. Stebel
based on: PK *et al.*, Comput.Phys.Commun. 281 (2022) 108502,
PK *et al.*, 2111.10189

Workshop on *Efficient simulations on GPU hardware* at ETH Zurich,
24-27.10.2022

ML acceleration of LQFT simulations

Motivations

- in LQCD we are limited by the number of simulation points in the parameters space
- we would like to generate more ensembles with smaller autocorrelations
- ML may help $\rightarrow$
  - P. Shanahan Lattice2022 talk (arXiv:2208.03832)
  - D. Albandea Lattice2022 talk
  - K. Nicoli *et al.*, Phys.Rev.Lett. 126 (2021) 3, 032001
  - L. del Debbio PoS LATTICE2021 (2022) 059
  - and others

Two-dimensional statistical systems

- test the approach for systems with discrete degrees of freedom and try to scale up
- playground: Ising model
- Potts model with $Q = 12$ at the first-order phase transition

Variational Autoregressive Neural Network

Complete factorisation (VAN approach):

$$p(\mathrm{s}) = p(s^1)\prod_{i=2}^{N} p(s^i|s^1,\ldots,s^{i-1})$$

Conditional factorisation (hierarchical approach):

$$p(\mathrm{s}) = p(\mathrm{B}(\mathrm{s}))p(\mathrm{I}(\mathrm{s})|\mathrm{B}(\mathrm{s})) = p(\mathrm{B}(\mathrm{s}))\prod_{a=1}^{4} p(\mathrm{I}^a(\mathrm{s})|\mathrm{B}^a(\mathrm{s})),$$

where

$$p(\mathrm{B}(\mathrm{s})) = p(s_B^1)\prod_{i=2}^{N_B} p(s_B^i|s_B^1,s_B^2,\ldots,s_B^{i-1})$$

and

$$p(\mathrm{I}^a(\mathrm{s})|\mathrm{B}^a(\mathrm{s})) = \prod_{i=1}^{N_I} p(s_I^{a,i}|s_I^{a,1},s_I^{a,2},\ldots,s_I^{a,i-1};\mathrm{B}^a(\mathrm{s})).$$

Variational Autoregressive Neural Network

$$p(\mathrm{s}) \to q_\theta(\mathrm{s}) = q_\theta(s^1)\prod_{i=2}^{N} q_\theta(s^i|s^1,\ldots,s^{i-1})$$

Architecture

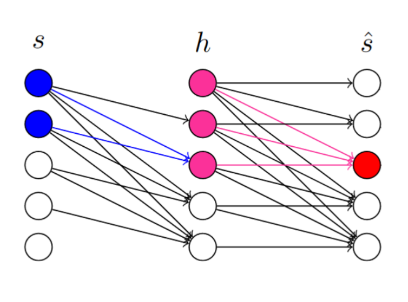


Figure taken from D. Wu, Phys. Rev. Lett. 122, 080602 (2019)

NMCMC

Acceptance probability:

$$\mathscr{P}(s_k \to s_{k+1}) = \min\left(1, \frac{p(s_{k+1})q_\theta(s_k)}{p(s_k)q_\theta(s_{k+1})}\right) = \min\left\{1, \frac{w(s_{k+1})}{w(s_k)}\right\}$$

K. Nicoli *et al.*, Phys.Rev.E 101 (2020) 2, 023304

Importance weights:

$$w(s) = \frac{p(s)}{q(s)} \quad \text{and} \quad w(s_1) \geq w(s_2) \geq \cdots \geq w(s_M)$$

Markov chain transition matrix eigenvalues:

$$\lambda_k = \begin{cases} 1 & \text{for } k = 0, \\ \sum_{i=k}^{M} q(s_i)\left(1 - \frac{w(s_i)}{w(s_k)}\right) & \text{for } 0 < k \leq M-1 \end{cases}$$

J. S. Liu, *Metropolized independent sampling with comparisons to rejection sampling and importance sampling*, Statistics and Computing, 6 (1996) 113

ML acceleration of LQFT simulations

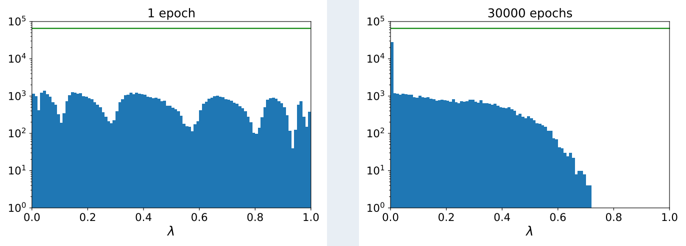
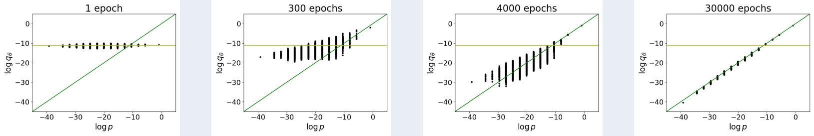


Figure: Histograms of $M-1$ eigenvalues of transition matrix, $\lambda_{k>0}$, for system 4×4. Training was performed using the D_{KL} loss function. Left figure is for initial state of network, right is for fully trained network. Green line denotes $M-1=2^{16}-1=65535$ value.

$\beta = 0.6$:



$\beta = 0.3$:

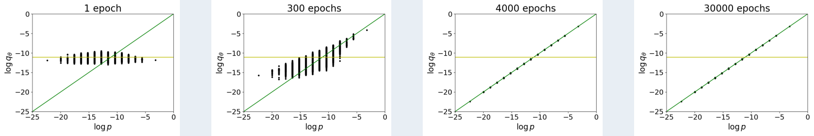


Figure: Dynamics of the neural network training using the KL divergence loss function. The yellow horizontal line shows a uniform probability distribution of $p(s) = 2^{-16}$.

ML acceleration of LQFT simulations

Factorization revisited

$$p(\mathrm{s}) = p(\mathrm{B}(\mathrm{s}))p(\mathrm{I}(\mathrm{s})|\mathrm{B}(\mathrm{s})) = p(\mathrm{B}(\mathrm{s}))\prod_{a=1}^{4} p(\mathrm{I}^a(\mathrm{s})|\mathrm{B}^a(\mathrm{s})),$$

Hierarchical approach

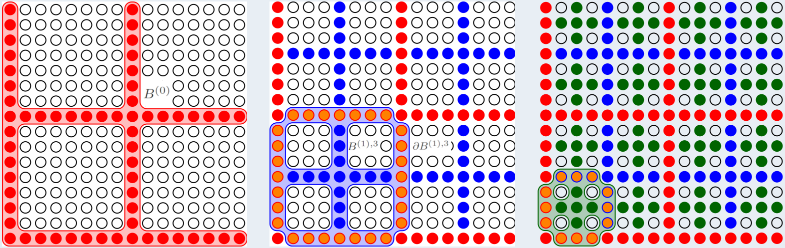


Figure: Example of hierarchical partitioning for $L = 16$.

ML acceleration of LQFT simulations

Hierarchical approach

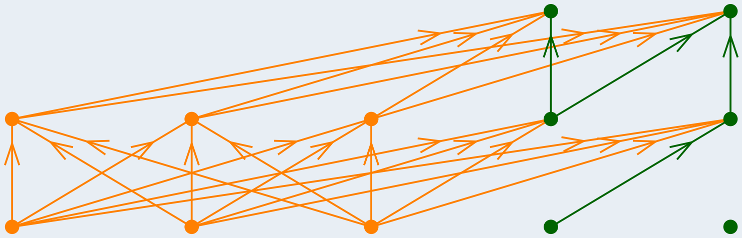


Figure: Scaled down representation of the architecture of the smallest neural network used to generate green sublattice.

Scaling

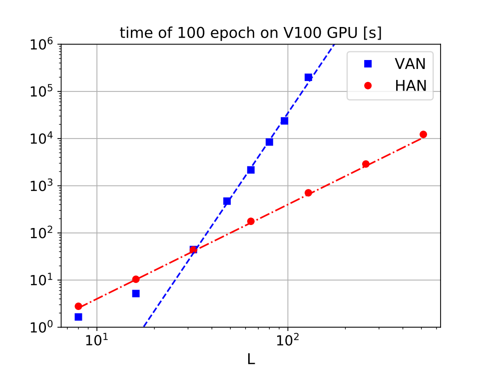
For each of the L^2 spins we need to calculate the probability, which costs approximately L^4 FLOPs, because its a matrix-vector multiplication of size L^2,

$$C_{\mathrm{VAN}} \sim L^2 \times L^4 = L^6.$$

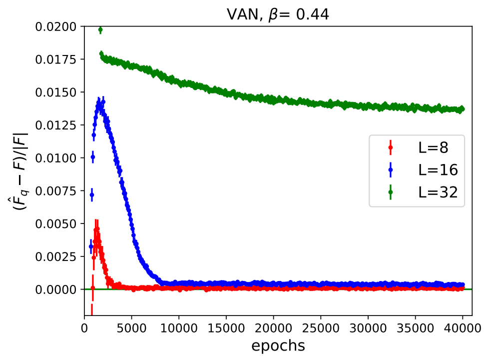
The largest lattice has $2L$ spins, hence we multiply a matrix-vector of size $2L$,

$$C_{\mathrm{HAN}} \sim 2L \times 4L^2 = 136L^3.$$

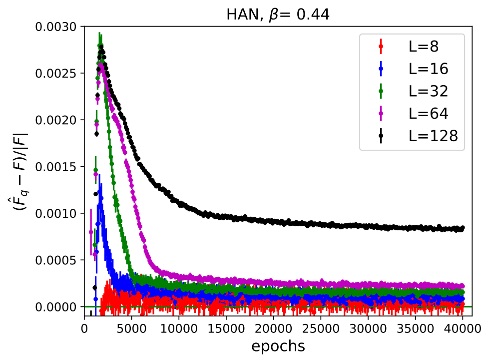
Results for the Ising model



Results for the Ising model



Results for the Ising model



ML acceleration of LQFT simulations

Potts model

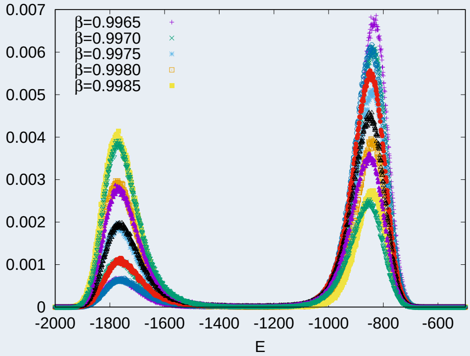
- $Q = 12$ state model
- first order phase transition at known β_c
- one-hot encoding increases the input/output by factor 12
- `softmax` layer at the output

Pretraining

- one can reuse the neural nets trained at smaller system size, different temperature
- only the two largest neural nets have to be trained from scratch

ML acceleration of LQFT simulations

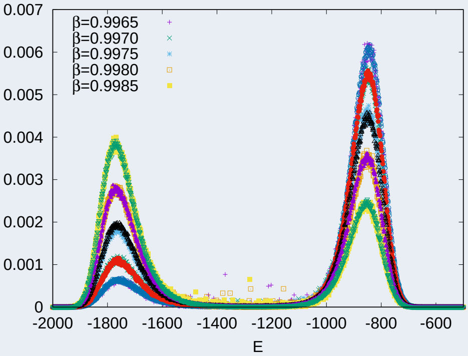
Histograms



Comparison of $p(E)$ and $q_\theta(E)$.

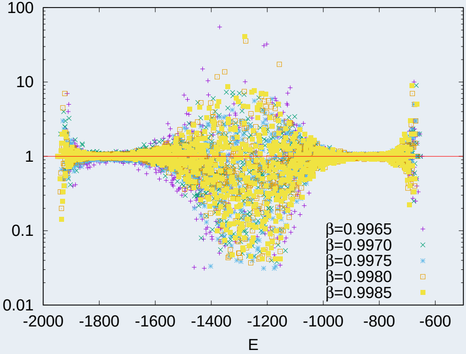
ML acceleration of LQFT simulations

Histograms



Comparison of $p(E)$ and $q_\theta(E)$ after accept/reject step.

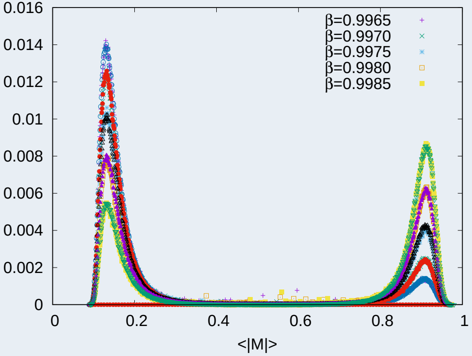
Histograms



Comparison of $p(E)$ and $q_\theta(E)$ after accept/reject step.

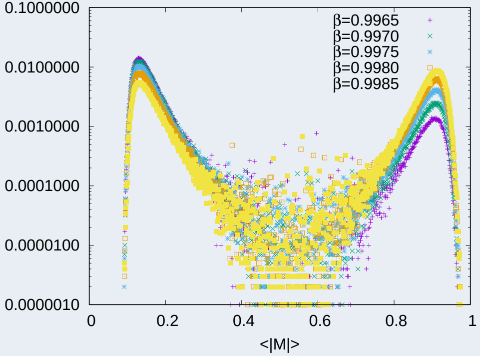
ML acceleration of LQFT simulations

Histograms



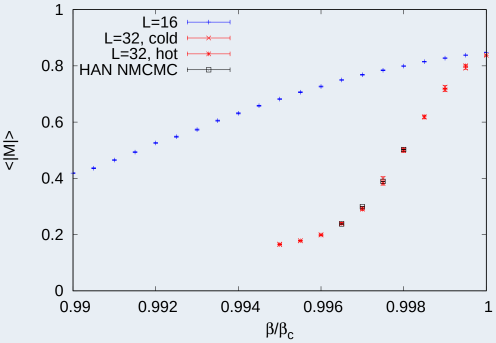
Comparison of $p(M)$ and $q_\theta(M)$ after accept/reject step.

Histograms



Comparison of $p(M)$ and $q_\theta(M)$ after accept/reject step.

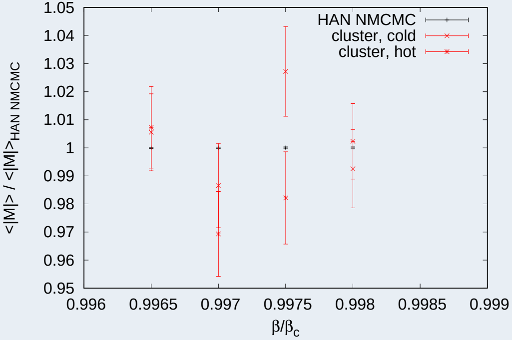
Magnetization



Comparison of $\langle M \rangle$ between cluster and NMCMC algorithms.

ML acceleration of LQFT simulations

Magnetization



Ratio of $\langle M \rangle$ between cluster and NMCMC algorithms using 10^6 configurations.

Conclusions

Summary

- we introduced a hierarchical approach based on VAN
- it allows to train efficiently systems of much larger number of degrees of freedom
- Potts model with $Q = 12$: $16 \times 16 \approx 2$h, $32 \times 32 \approx 48$h, attempting to train 64×64
- training and simulations work even at the first order phase transition

Outlook

- implemented and tested with conditional normalizing flows for ϕ^4 but lower performance than the original approach

Thank you very much for your attention