Multilevel Sampling of Lattice Theories using Auto-regressive Based Models

ML meets LFT, Swansea University

Ankur Singha

ML group, BIFOLD TU Berlin

Collaborators: Elia Cellini, Kim Nicoli, Shinichi and Karl Jansen.

Introduction to Multillevel Sampling

- **State Space Decomposition:**
  - Let $\mathbf{X}$ denote the state space, which can be decomposed into multiple levels or scales:

$$\mathbf{X} = \bigcup_{k=1}^{L} \mathbf{X}_k$$

    where L is the number of scales, and $\mathbf{X}_k$ represents the state space at the k-th scale.
- **Multilevel Representation:**
  - Each state $\mathbf{x} \in \mathbf{X}$ can be represented hierarchically:

$$\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L)$$

    where $\mathbf{x}_k \in \mathbf{X}_k$. e.g. $(\mathbf{x}_1)$ is coarse lattice and $(\mathbf{x}_L)$ is the fine lattice.
- **Probability Distribution:**
  - Assume the target distribution $P(\mathbf{x})$ can be factorized hierarchically:

$$P(\mathbf{x}) = P(\mathbf{x}_1) \prod_{k=2}^{L} P(\mathbf{x}_k \mid \mathbf{x}_{1:k-1})$$

Multilevel Monte Carlo Sampling

- **Concept:**
  - Divide the problem into multiple hierarchical levels or scales.
  - Sample at each level, using results from the previous level.
- **Steps:**
  1. **Coarse-Grained Level:**
     - Sample the coarsest lattice from the lowest dimensional distribution.
     - Interactions depend on the marginalization process.
  2. **Intermediate Levels:**
     - Refine sampling by focusing on higher scales, informed by coarse samples.
     - Incorporate intermediate details and interactions.
  3. **Fine-Grained Level:**
     - Complete the sampling process by adding fine details.
     - Accurately capture target interactions at a detailed level.
- **Benefits:**
  - Reduces computational burden by focusing on relevant scales.
  - Enhances convergence and accuracy of Monte Carlo simulations.

Monte Carlo multilevel approach: [K. Jansen,2020]

The target density: $q(\varphi|k_f) = \frac{e^{-\beta H(\varphi|k_f)}}{Z}$ $;\varphi = [\varphi_f : a, \varphi_i : \sqrt{2}a, \varphi_c : 2a]$

Multilevel proposal:

$$q(\varphi; k_f, k_i, k_c) = q(\varphi_f|\varphi_f, \varphi_c; k_f)q(\varphi_i|\varphi_c; k_i)q(\varphi_c; k_c)$$

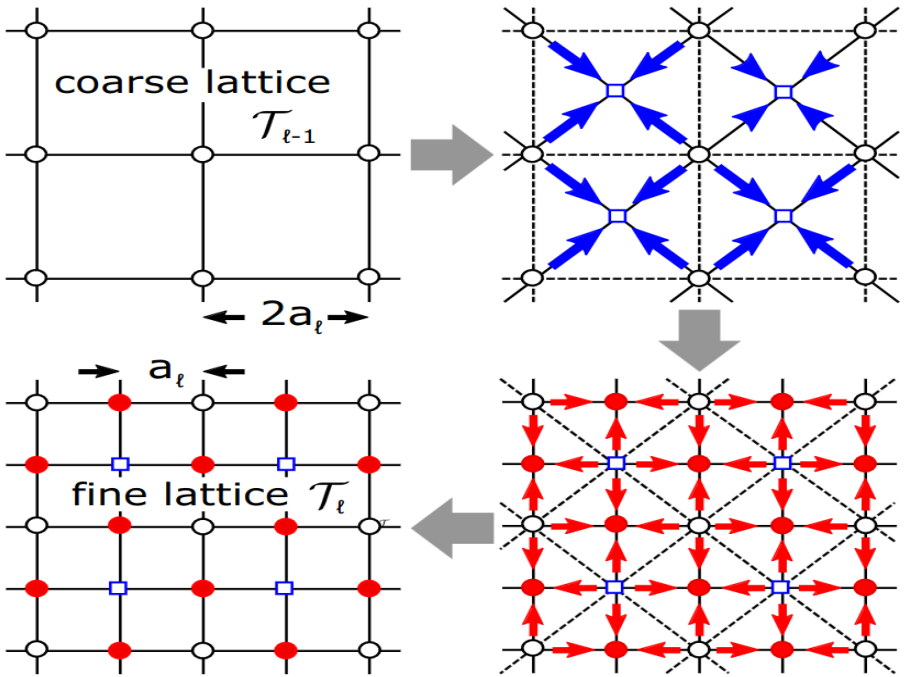


Figure: Taken from [K. Jansen,2020][1]

Traditional multilevel approach

- Sample the coarse variables φ_c from the distribution

$$\varphi_c \sim q(\varphi_c|k_c)$$

using Monte Carlo method from a distribution defined by the Renormalised Hamiltonian at $2a$.

- Sample the intermediate φ_i from $\varphi_i \sim q(\varphi_i|k_i, \varphi_c)$ from a renormilized Hamiltonian at $\sqrt{2}a$ and and fine variables $\varphi_f \sim q(\varphi_f|k_f, \varphi_i, \varphi_c)$ from the target Hamiltonian using Heatbath method.
- One we have $\varphi = [\varphi_f, \varphi_i, \varphi_c]$, do a Metropolish-Hasting:

$$\alpha = min\left(1, \frac{p(\varphi^t)}{p(\varphi^{t-1})} . \frac{q(\varphi^{t-1})}{q(\varphi^t)}\right)$$

Where, the new proposal at time t, $\varphi^t = [\varphi_f^t, \varphi_i^t, \varphi_c^t]$

Multilevel: Traditional Approach for Ising Model

Low acceptance

The overlap between the target and the proposal distribution is very low and we get a very low acceptance rate in the MH step [Schmidt, 1983], [Faraz, 1985].

Poor proposals

The RG approach is not exact and hence the action derived at different levels is not the true action or the action of the true marginal distribution.

Observations:

- Parameterize proposal distribution $q(\varphi; k_f, k_i, k_c)$: Optimizing parameters (k_i, k_c) does not provide a significant impact on acceptance rate.
- The maximum acceptance rate from the proposal is less than 10% for 32×32.
- The assumption of the same order of interactions at other scales is not true anymore.

Variational Autoregressive Networks (VAN)

Variational Autoregressive Networks (VAN) model the joint probability of a lattice configuration φ using a product of conditional probabilities:

$$q_\theta(\varphi) = \prod_{i=1}^{N} q_\theta(\varphi_i|\varphi_1, \cdots, \varphi_{i-1})$$

Example

Ising spins:

$$q_\theta(s_i|s_{<i}) = \hat{s}_i \delta_{s_i,+1} + (1 - \hat{s}_i)\delta_{s_i,-1}$$

where,

$$\hat{s}_i = \sigma(g_\theta(s_{<i}))$$

with g_θ autoregressive netwok such as the Pixel CNN.

Autoregressive-based Multilevel: [K. Jansen,2020]

One block Multilevel proposal:

$$q(\varphi;\theta) = q(\varphi_f|\varphi_f,\varphi_c;\theta_f)q(\varphi_i|\varphi_c;\theta_i)q(\varphi_c;\theta_c)$$

Coarse Level: The coarse level distribution is modeled by a Variational Autoregressive Network (VAN).

$$\varphi_c \sim q(\varphi_c;\theta_c)$$

Interaction range

At the coarse level interactions range is not known, so a model cable of generating long range interactions is suitable.

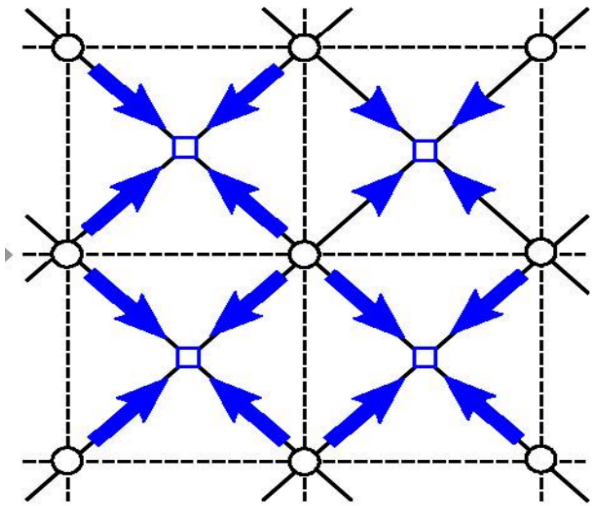
Training a VAN

The coarse distribution is quite low dimension; $2 \times 2; 4 \times 4$. Thus training is efficient and faster.

Autoregressive-based model:

Intermediate Level: The intermediate distribution is learned represented by a conditional auto-regressive model.

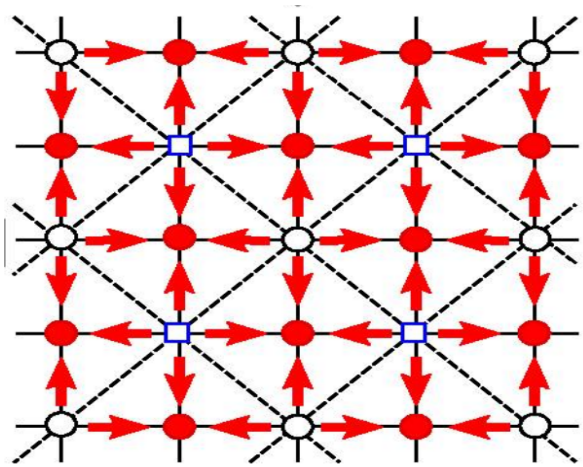
$$\varphi_i \sim q(\varphi_i; \varphi_c, \theta_i)$$



The interactions range is decide to be next nearest neighbour on the current sites. This is fixed by choosing suitable kernel.

Autoregressive-based model

Fine Level: The fine distribution sampled via Heatbath given the coarse and intermediate spins: $\varphi_f \sim q(\varphi_f; \varphi_i, \varphi_c)$



Heatbath

Since the target distribution is local, hence given $[\varphi_i, \varphi_c]$, the Heatbath sampling is more efficient to sample φ_f instead of the conditional auto-regressive model.

Multilevel Blocks

One Block

$N \times N \to 2N \times 2N$

One block distribution: $q(\varphi;\theta) = q(\varphi_f|\varphi_i,\varphi_c;\theta_f)q(\varphi_i|\varphi_c;\theta_i)q(\varphi_c;\theta_c)$

n blocks

$N \times N \to 2^n N \times 2^n N$

$$q(\varphi;\theta) = q(\varphi_f|\varphi_1\varphi_{2,\ldots,}\varphi_{2n-1};\theta_f)\prod_{i=1}^{2n-1} q(\varphi_i|\varphi_0\varphi_{1,\ldots,}\varphi_{i-1};\theta_i)q(\varphi_0;\theta_0)$$

Components

I) The coarse distribution is sampled via VAN II) All intermediate are sampled conditional VAN and III) The final level by Heatbath.

Objective Function

Training can be proceeded by minimizing the Kullback-Leibler (KL) divergence as:

$$\mathrm{D}_{\mathrm{KL}}(q_\theta \parallel p) = \sum_{\varphi} q_\theta(\varphi) \ln \left(\frac{q_\theta(\varphi)}{p(\varphi)} \right) = \beta(F_q - F)$$

This is equivalent as minimizing the variational free energy:

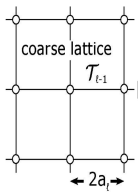
$$F_q = \sum_{\varphi} q_\theta(\varphi) \Big[\beta \mathcal{H}(\varphi) + \ln q_\theta(\varphi) \Big] \tag{1}$$

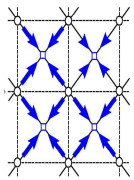
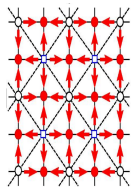
We don't train this multilevel model directly at fine level model. We start from the coarse level and move block-wise to the fine level.

Block wise Training procedure

Example: $L_c = 2 \times 2 \rightarrow 64 \times 64$. We have to train 6 multilevel blocks.

- $L_c = 2 \times 2 \rightarrow 4 \times 4$:
  The first the coarse level is trained independently using VAN. Then train C-VAN for intermediates (blue) and complete intermediates (red).



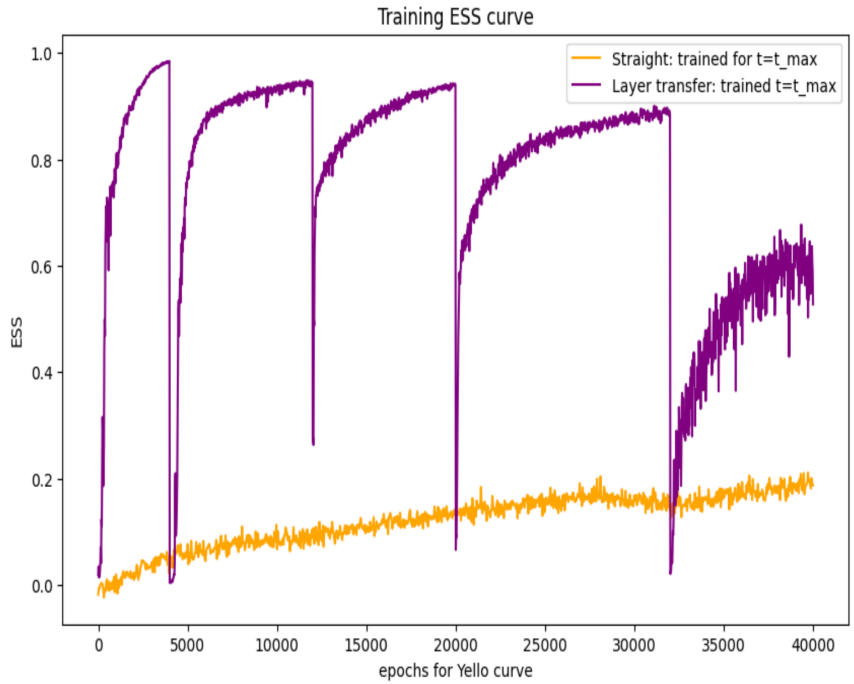



- $L_c = 4 \times 4 \rightarrow 8 \times 8$: While training the 2nd block we transfer the weights from the last block. This gives a good starting point for the training.

...continues until the last last block.
In the last block the fine distribution does not need initialization.

Training ESS Curve

We train the model using both the standard method and block-wise weight transfer, ensuring the same computational time for each.



result:

We find that the block transfer method significantly reduces training time and converges faster.

Hierarchical Autoregressive Network (HAN)

HAN: (P. Bialasl, 2022)

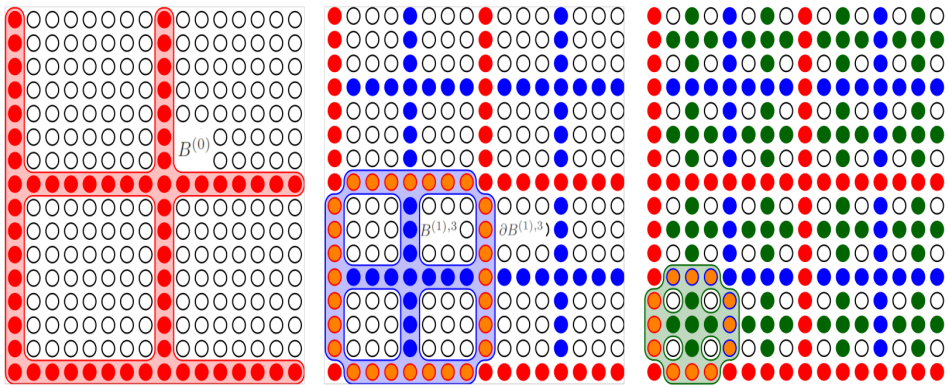


Figure 1: Example of hierarchical partitioning for $L = 16$. On the first level, the red boundary $B^{(0)}$ highlighted on the left panel of the figure is generated with one neural network $\mathcal{N}^0$. At the second level of hierarchy, one neural network of smaller size $\mathcal{N}_1$ is used to consecutively fix four sets of boundaries shown in blue. The example of $B^{(1),3}$ is highlighted in the middle panel. The surrounding spins $\partial B^{(1),3}$ are shown in orange. At the third level of hierarchy, one neural network of even smaller size $\mathcal{N}_2$ is used to consecutively fix sixteen sets of boundary spins marked in green. The example of $B^{(2),15}$ is highlighted in green on the right panel. The remaining empty spins corresponding to $I^k \equiv B^{(3),k}$, $k = 1, \ldots, 64$ have all the neighbours fixed and therefore can be generated from a local Boltzmann distribution with the heatbath algorithm.

Figure: The Hierarchical Autoregressive Network (HAN) Approach.

ESS: HAN vs Multilevel

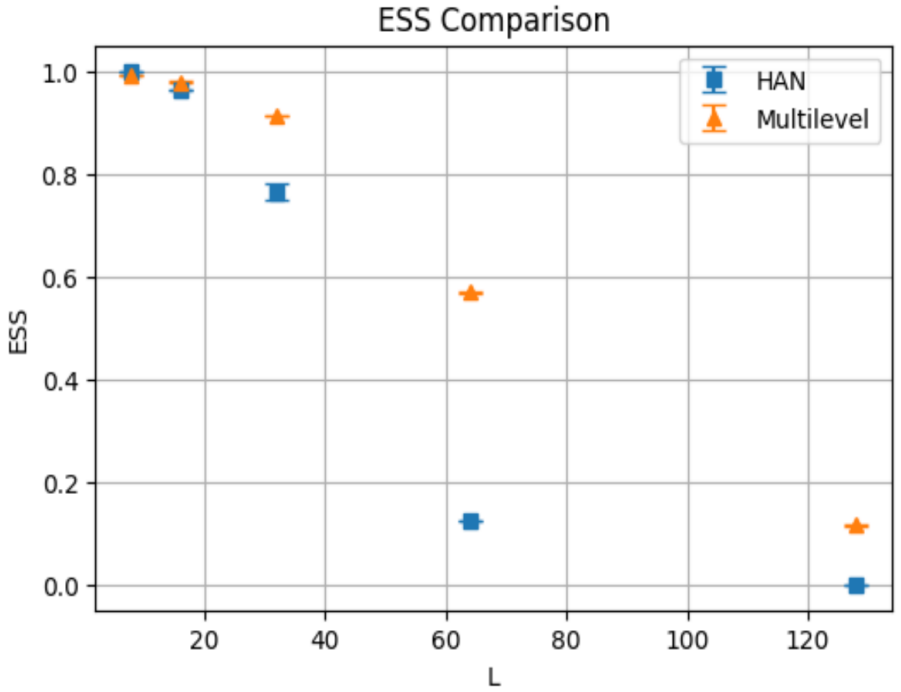
The Importance weight: $w = p(\varphi)/q_\theta(\varphi)$;

Estimator: $\hat{w} = \frac{1}{\hat{Z}} \frac{\exp\left(-\beta H(\varphi)\right)}{q_\theta(\varphi)}$

$$\hat{Z} = \frac{1}{N} \sum_i \frac{\exp\left(-\beta H(\varphi_i)\right)}{q_\theta(\varphi_i)}$$

$$\text{ESS} = \frac{1}{\mathbb{E}_q[w^2(\varphi)]}$$

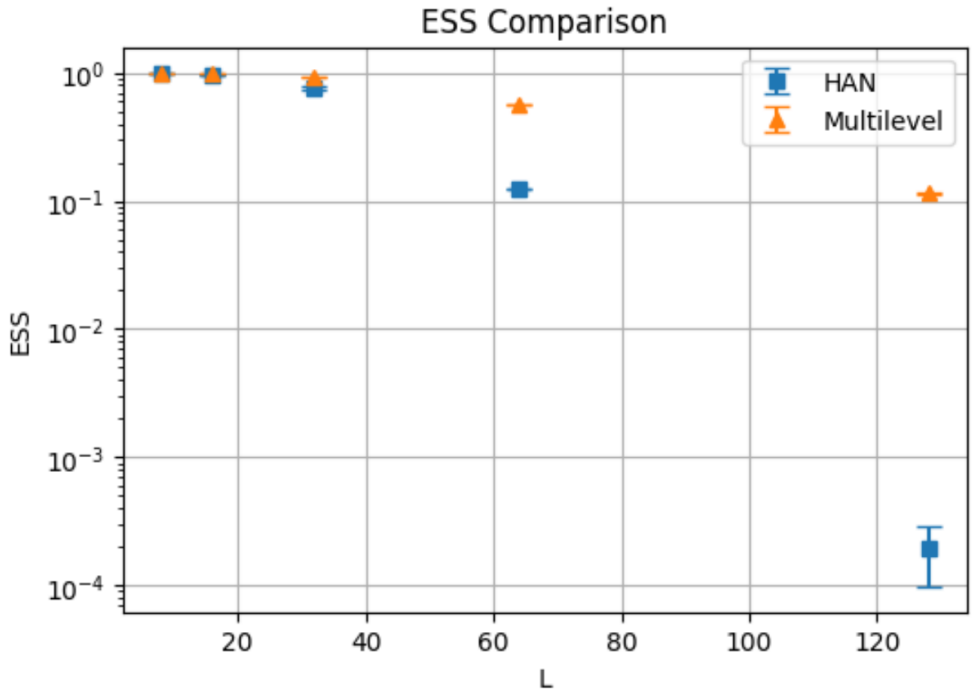
(Hackett, 2021; Vaitl, 2022)



Result

The ESS for HAN is significantly lower compared to the multilevel method for 64×64 and 128×128.

ESS (log-scale) : HAN vs Multilevel



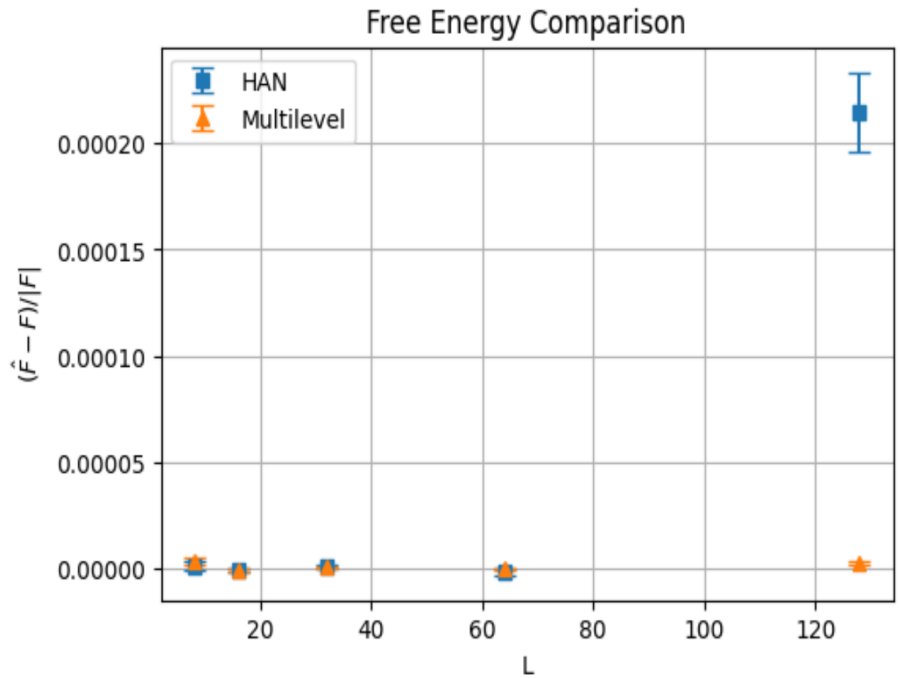
result:

For the 128×128 grid, the ESS for HAN is three orders of magnitude lower than that of the multilevel method

Free Energy: HAN vs Multilevel

The Free Energy:

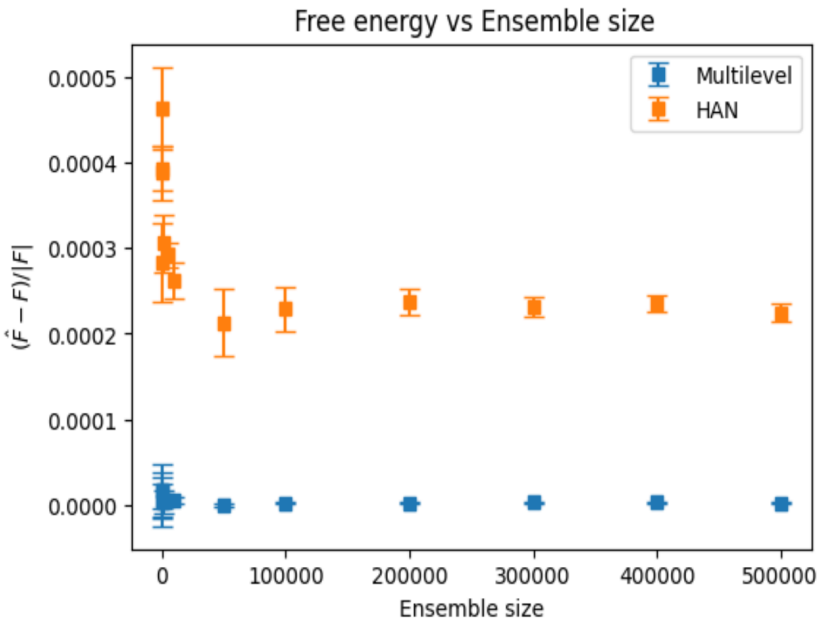
$$\hat{F} = -\frac{1}{\beta} \log \hat{Z}$$

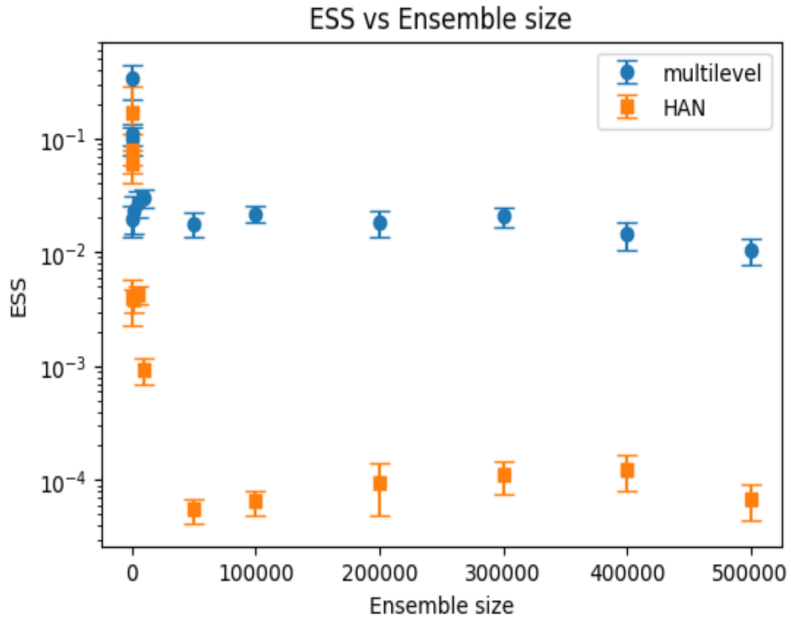


result:

The Free Energy is biased for HAN 128×128

ESS and Free Energy vs Sample size



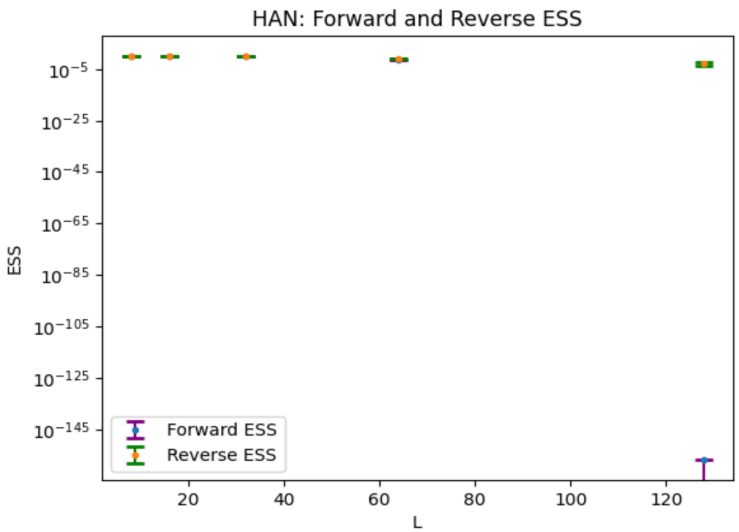


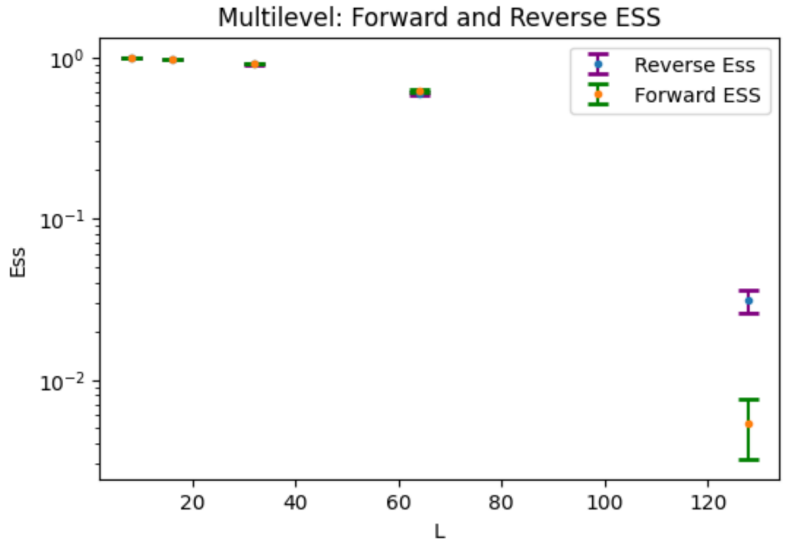
result:

- The bias in Free Energy cannot be reduced even with up to half a million samples.
- The ESS initially starts at similar values for both methods but decreases for HAN as the sample size increases.

Investigating the HAN bias: Forward ESS

We investigated the bias and low ESS for HAN by examining the ESS of both Forward and Reverse estimators. $\text{ESS} = \frac{1}{\mathbb{E}_p[w(\varphi)]}$ (Hackett, 2021; Vaitl, 2022)



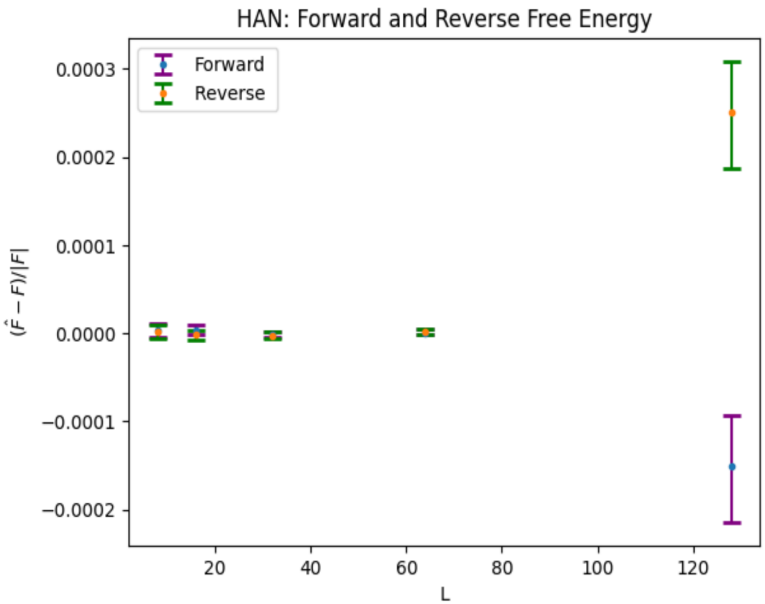


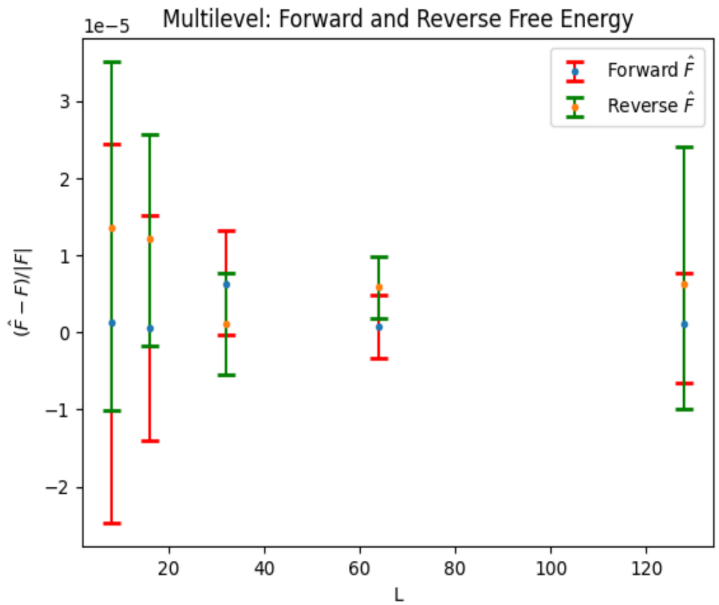
result

The Forward ESS is zero for 128×128, indicating that the model samples do not adequately represent the target distribution. This suggests a likelihood of effective mode dropping in the model.

Compatibility of F/R Free Energy: HAN vs Multilevel

$$\hat{F}_{Fw} = -\frac{1}{\beta}\log(\hat{Z}_p); \hat{F}_{Rv} = -\frac{1}{\beta}\log(\hat{Z}_q)$$



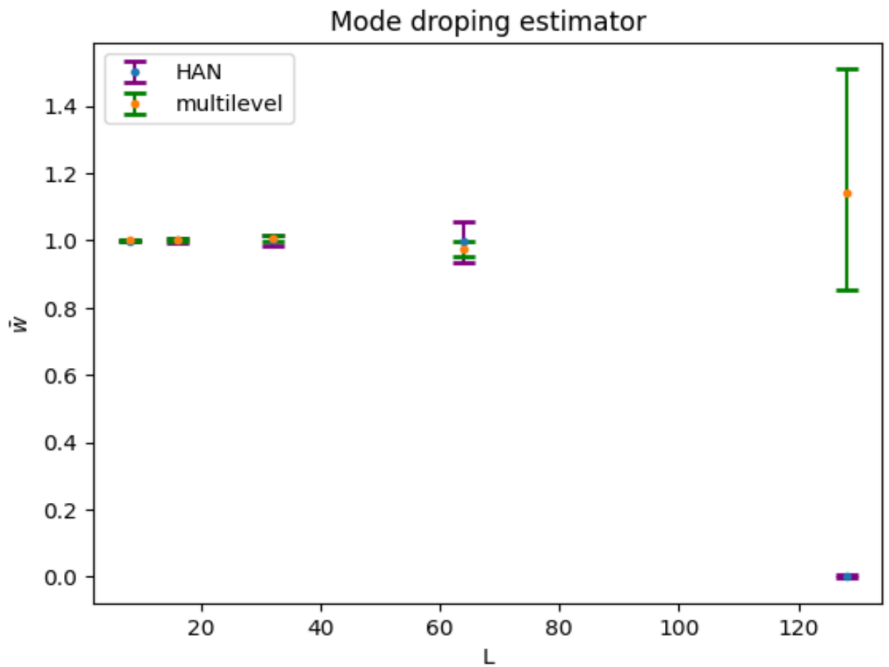


result:

The Forward and Reverse Free Energy are consistent for the multilevel approach, but this is not the case for HAN.

Mode dropping estimator

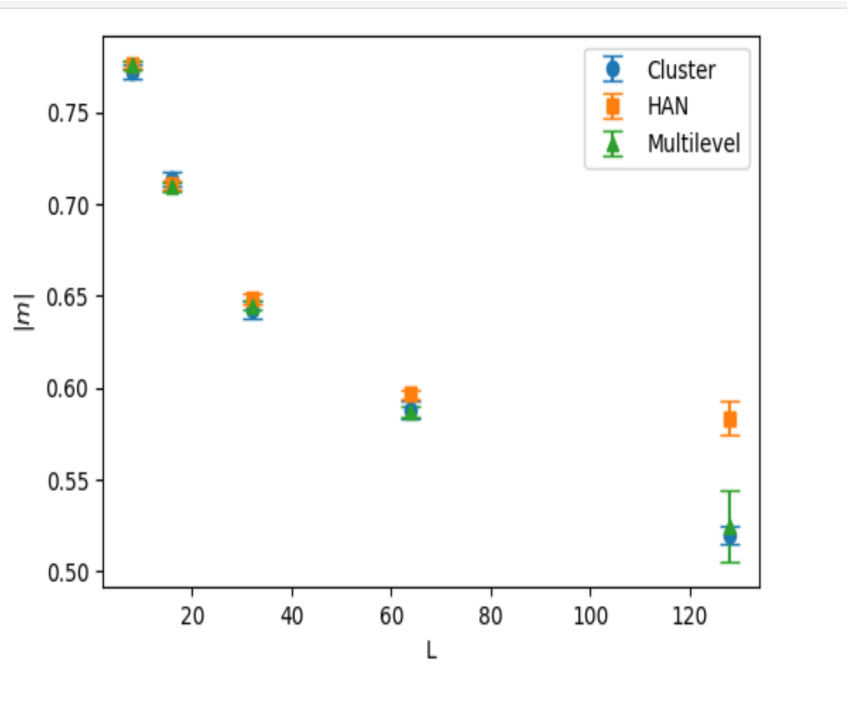
We define a mode dropping estimator: $\bar{w} \approx \frac{1}{\hat{Z}_p}\left(\frac{1}{N}\sum_i \frac{\exp\left(-\beta H(\varphi_i)\right)}{q_\theta(\varphi_i)}\right)$ (Nicoli, 2023)

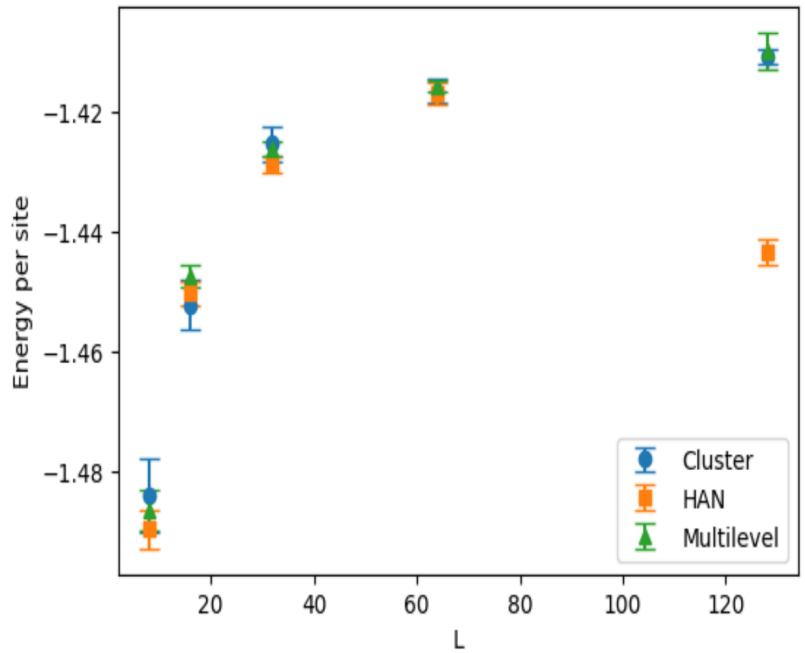


result:

From the mode dropping estimator it is clear that the HAN model smaller effective support than the target.

Other Observables: HAN vs Multilevel

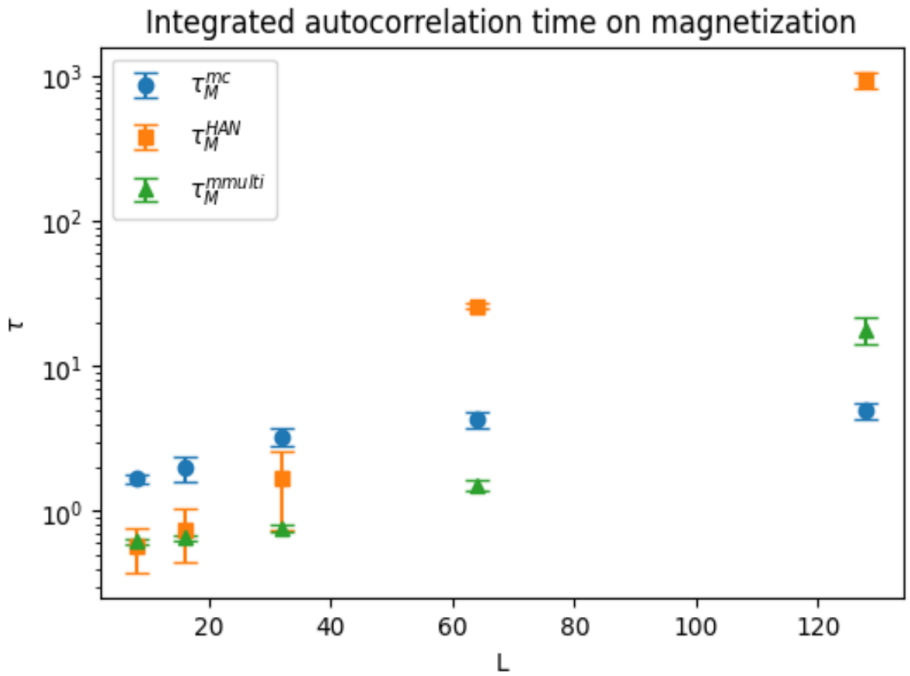




result:

Magnetization and Eenergy are also biased for 128×128.

Autocorrelation Comparison



result:

The integrated autocorrelation time for multilevel is much lower then HAN specially for larger lattices.

Summary and Conclusion

- We build a multilevel sampling method for Ising model following the traditional approach incorporating autoregressive models.
- We use a training strategy inspired by RG where fine level distributions are initialised with coarser level trained models.
- We find that the baseline method HAN model's ESS declines as the lattice size increases.
- The autocorrelation for multilevel is roughly 200 times smaller than HAN.
- VAN struggles to scale up, whereas HAN, although efficient in training, suffers from decreased overall performance. Multilevel sampling addresses and overcomes these issues.

Thank You!

bibliography

[1] Karl Jansen, Eike H. Müller, and Robert Scheichl. “Multilevel Monte Carlo algorithm for quantum mechanics on a lattice”. In: *Phys. Rev. D* 102 (11 Dec. 2020), p. 114512. DOI: 10.1103/PhysRevD.102.114512. URL: https://link.aps.org/doi/10.1103/PhysRevD.102.114512.