

Equivariant Transformer is all you need

Akio Tomiya^{a,b,*} and Yuki Nagai^{c,d}

^a*Faculty of Technology and Science, International Professional University of Technology, 3-3-1, Umeda, Kita-ku, Osaka, 530-0001, Osaka, Japan*

^b*RIKEN/BNL Research center, Brookhaven National Laboratory, Upton, 11973, NY, USA*

^c*CCSE, Japan Atomic Energy Agency, 178-4-4, Wakashiba, Kashiwa, Chiba 277-0871, Japan*

^d*Mathematical Science Team, RIKEN Center for Advanced Intelligence Project (AIP), 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan*

E-mail: akio@yukawa.kyoto-u.ac.jp, nagai.yuki@jaea.go.jp

Machine learning, deep learning, has been accelerating computational physics, which has been used to simulate systems on a lattice. Equivariance is essential to simulate a physical system because it imposes a strong induction bias for the probability distribution described by a machine learning model. This reduces the risk of erroneous extrapolation that deviates from data symmetries and physical laws. However, imposing symmetry on the model sometimes occur a poor acceptance rate in self-learning Monte-Carlo (SLMC). On the other hand, Attention used in Transformers like GPT realizes a large model capacity. We introduce symmetry equivariant attention to SLMC. To evaluate our architecture, we apply it to our proposed new architecture on a spin-fermion model on a two-dimensional lattice. We find that it overcomes poor acceptance rates for linear models and observe the scaling law of the acceptance rate as in the large language models with Transformers.

*The 40th International Symposium on Lattice Field Theory (Lattice 2023)
July 31st - August 4th, 2023
Fermi National Accelerator Laboratory*

*Speaker

1. Introduction

Lattice QCD is essential for calculating quantum field expectations but struggles with the critical slowing down issue, reducing computational efficiency. Machine learning methods can efficiently handle this problem and work well with structured data like gauge configurations [1, 2].

Transformer is a neural network originally for dealing with natural languages, but it has been applied to various data [3–6]. The most crucial feature of Transformer is it can deal with global correlations, such as modifiers in natural language, which sometimes act from distant places.

In computational physics with machine learning, in particular equivariant neural networks enhance numerical calculations by capturing input data symmetries. This improves generalization on unfamiliar data and may boost computational efficiency. Data symmetry ensures alignment with physical laws like momentum conservation. Convolutional layers, for instance, exhibit equivariant properties to spatial translation. This reduces the risk of erroneous extrapolation that deviates from data symmetries and physical laws [7].

In this work, we develop an equivariant Transformer for physical systems. In quantum field theory, local action/Hamiltonians are typically considered. However, when fermions, described by Grassmann numbers, are integrated ahead of numerical simulations, the resulting effective action/Hamiltonian becomes non-local regarding bosons. We develop a neural network architecture which is capable with global correlations from fermions and symmetric under system's symmetries.

In this proof-of-principle study, we employ the *double exchange model* (DE) in two spacial dimensions. DE model is well established model in condensed matter physics and contains fermions and spatially fixed classical Heisenberg-spins. The model Hamiltonian is invariant under global $O(3)$ spin rotation. This model is similar to Yukawa system with fermions and three component scalars on the lattice in particle physics. To see more detail of this study, refer [8].

2. Concepts in Machine learning

2.1 Self-learning Monte-Carlo

We review concepts in machine learning to introduce our numerical calculation. The Self-learning Monte-Carlo (SLMC) is an exact Markov chain Monte-Carlo (MCMC) algorithm with an effective model [9]. In MCMC for a spin system, a spin configuration \mathcal{S} is distributed with a probability distribution $W(\mathcal{S})$. Samples from desired distribution are obtained after many steps. The detailed balance condition is a sufficient condition for convergence of MCMC, which is $W(\{\mathcal{S}\})T(\{\mathcal{S}'\}|\{\mathcal{S}\}) = W(\{\mathcal{S}'\})T(\{\mathcal{S}\}|\{\mathcal{S}'\})$, where $T(\{\mathcal{S}'\}|\{\mathcal{S}\})$ is the transition probability from a configuration $\{\mathcal{S}\}$ to another configuration $\{\mathcal{S}'\}$. If a probabilistic process described $T(\{\mathcal{S}'\}|\{\mathcal{S}\})$ with this condition, the obtained configurations distributed according to $W(\{\mathcal{S}\})$.

In general Metropolis-Hastings (MH) algorithm, the transition probability is factorised in two sub-steps, $T(\{\mathcal{S}'\}|\{\mathcal{S}\}) = g(\{\mathcal{S}'\}|\{\mathcal{S}\})A(\{\mathcal{S}'\}, \{\mathcal{S}\})$, where the proposal distribution $g(\{\mathcal{S}'\}|\{\mathcal{S}\})$ is the conditional probability of proposing a configuration $\{\mathcal{S}'\}$ when a configuration $\{\mathcal{S}\}$ is given, and the acceptance ratio $A(\{\mathcal{S}'\}, \{\mathcal{S}\})$ is the probability to accept the proposed configuration $A(\{\mathcal{S}'\}, \{\mathcal{S}\})$. The Markov chain that has the desired distribution $W(\{\mathcal{S}\})$ is obtained when the



Figure 1: Schematic figure of SLMC. Blue thin arrows the Metropolis chain in the inner chain with W_{eff} . Red bold arrows indicate the MH test in the outer chain with W_{eff} and W as Eq. (2).

acceptance ratio is given as

$$A(\{\mathbf{S}'\}, \{\mathbf{S}\}) = \min \left(1, \frac{W(\{\mathbf{S}'\}) g(\{\mathbf{S}\}|\{\mathbf{S}'\})}{W(\{\mathbf{S}\}) g(\{\mathbf{S}'\}|\{\mathbf{S}\})} \right). \quad (1)$$

This is the general argument about the MH algorithm. The Metropolis test is a reduced version, $A(\{\mathbf{S}'\}, \{\mathbf{S}\}) = \min(1, W(\{\mathbf{S}'\})/W(\{\mathbf{S}\}))$ and this is obtained assuming the reversibility for $g(\{\mathbf{S}'\}|\{\mathbf{S}\}) = g(\{\mathbf{S}\}|\{\mathbf{S}'\})$.

2.1.1 Inner Markov chain in SLMC

SLMC is a nested MCMC. The simplest update is the local update, where a single site in the configuration is randomly selected and its spin orientation is changed. We perform the Metropolis accept/reject procedure following to the single-site-update with $W_{\text{eff}}(\{\mathbf{S}\})$, which is the Boltzmann weight with an effective model. Since the Metropolis test with the single cite update satisfies the detailed balance condition, it will converge into $W_{\text{eff}}(\{\mathbf{S}\})$. In SLMC, the effective model $W_{\text{eff}}(\{\mathbf{S}\})$ contains trainable parameters.

2.1.2 Outer Markov chain in SLMC

In SLMC, the inner Markov chain using an effective model proposes processes for the outer chain. A correction step ensures the target system's distribution. In the general MH algorithm, the inner Markov updates represent the $g(\cdot|\cdot)$ process in the outer chain, with the acceptance ratio designed to offset W_{eff} and distribute as $W(\cdot)$. The acceptance ratio in the SLMC is given as

$$A(\{\mathbf{S}'\}, \{\mathbf{S}\}) = \min \left(1, \frac{W(\{\mathbf{S}'\}) W_{\text{eff}}(\{\mathbf{S}\})}{W(\{\mathbf{S}\}) W_{\text{eff}}(\{\mathbf{S}'\})} \right), \quad (2)$$

where $W(\{\mathbf{S}\})$ is the probability weight for the target system. We remark that, the second factor in the second column in $\min(1, \cdot)$ is up-side-down of the weights for the inner chain. In summary, SLMC process can be express as Fig. 1.

2.2 Equivariance

The concept of equivariance plays a pivotal role in machine learning and physics, offering solutions to issues like model overfitting and the preservation of physical laws. Its importance stems from its capacity to embed symmetries directly into neural networks, ensuring that the learned model respects certain symmetries of the data. In neural networks, equivariance is usually achieved through weight sharing, reducing the number of irrelevant parameters. This is critical as an excess of parameters often leads to model overfitting, undermining the model's ability to generalize to new data.

In physics, ensuring numerical calculations align with physical laws can be achieved via equivariant neural networks. While penalties in the loss function were previously used to impose

physical laws [10], this method is not always reliable. A more dependable approach embeds these symmetries directly into the neural network architecture [7].

3. Lattice setup

In this study, we employ semi-classical *double exchange* (DE) model in two dimension for a testbed [11–13]. It is a semi-classical system and has electrons and classical Heisenberg spins,

$$H = -t \sum_{\alpha, \langle i, j \rangle} (\hat{c}_{i\alpha}^\dagger \hat{c}_{j\alpha} + \text{h.c.}) + \frac{J}{2} \sum_i \mathbf{S}_i \cdot \hat{\boldsymbol{\sigma}}_i, \quad (3)$$

where \mathbf{S}_i is the classical Heisenberg spins on the i -th site, namely it is normalized O(3) scalar field. $\hat{c}_{i\alpha}^\dagger$ is the fermionic creation operator at the i -th site for fermion with spin $\alpha \in \{\uparrow, \downarrow\}$. A symbol $\langle i, j \rangle$ indicate the pairs of nearest neighbors. The interaction term with Pauli matrices are defined as $[\hat{\boldsymbol{\sigma}}_i]_\gamma \equiv \hat{c}_{i\alpha}^\dagger \sigma_{\alpha\beta}^\gamma \hat{c}_{i\beta}$ ($\gamma = x, y, z$). J is the interaction strength between the classical spins and the electrons and we consider the hopping constant t as the unit of energy. We adopt the periodic boundary condition on $N_x \times N_y$ site system. The total number of the sites is $N \equiv N_x N_y$. The Hamiltonian has O(3) rotational symmetry in the spin sector and discrete translational invariance. We want to calculate statistical expectation value with $\exp[-\beta H]$ for (3).

3.1 Equivariant Transformer for physical system

Here we introduce an equivariant Transformer for a physical spin system. Input of our Transformer is a spin configuration \mathbf{S} and output is modified spin configuration. In our formalism, query, key, and value are $N \times 3$ matrices and they are defined as

$$\mathbf{S}^Q \equiv \hat{W}^Q \mathbf{S}, \quad \mathbf{S}^K \equiv \hat{W}^K \mathbf{S}, \quad \mathbf{S}^V \equiv \hat{W}^V \mathbf{S}, \quad (4)$$

respectively. $S_{i\mu}^\alpha \equiv \sum_{\langle i, j \rangle_k} W_k^\alpha S_{j\mu}$, $W_k^\alpha \in \mathbb{R}$ ($\alpha = Q, K, V$ and $k = 0, 1, 2, \dots, \tilde{N}$). A symbol $\langle i, j \rangle_k$ picks up k -th nearest neighbors for sites i, j . This procedure can be regarded as a block spin transformation with $\tilde{N} + 1$ free parameters. In this work, we take $\tilde{N} = 6$.

Consequently, we introduce the following (equivariant) self-attention block as,

$$\text{SelfAttention}^{\text{spin}}(\mathbf{S}) = \check{M} \mathbf{S}^V, \quad (5)$$

where \check{M} is a $N \times N$ matrix. This \check{M} is defined by $[\check{M}]_{ij} = \sigma \left(\sum_{\mu=1}^3 S_{i\mu}^Q S_{j\mu}^K / \sqrt{3} \right)$, where i, j are indices for spatial position on the lattice. $\sigma(\cdot)$ is a nonlinear activation function. Intuitively, the argument $\sum_{\mu} S_{i\mu}^Q S_{j\mu}^K$ is a set of 2 point correlation functions of the blocked spin field from a point j to i and this is invariant under O(3) rotations since rotation matrices cancel out. In this study, we take an activation function $\sigma(\cdot)$ as ReLU function.

We construct the effective spin field that with multiple attention layers. Our neural neural network architecture is defined with a residual connection,

$$\mathbf{S}^{(l)} \equiv \mathcal{N} \left(\mathbf{S}^{(l-1)} + \text{SelfAttention}_{\boldsymbol{\theta}^{(l)}}^{\text{spin}}(\mathbf{S}^{(l-1)}) \right), \quad (6)$$

and $\mathbf{S}^{(0)} \equiv \mathbf{S}$ and $\mathbf{S}^{\text{eff}} \equiv \mathbf{S}^{(L)}$. l is an index for layers and $l = 1, 2, \dots, L$. $\boldsymbol{\theta}^{(l)}$ represents a set of network trainable parameters in l -th layer. $\mathcal{N}(\mathbf{S})$ normalizes the spin vector on each lattice

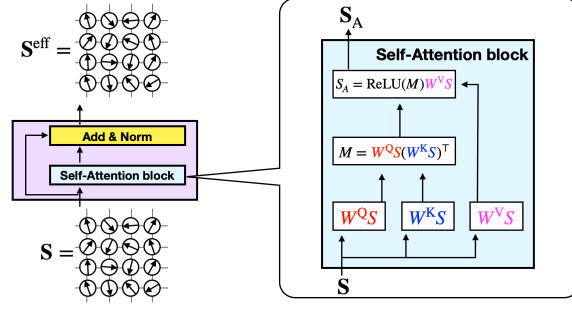


Figure 2: (Left) Effective spin construction using the Transformer with an Attention block. Yellow is defined by Eq. (6); purple is the attention block. (Right) Blue represents the attention block (see main text).

site, $\mathcal{N}(S_i) = S_i / \|S_i\|$. We call this network architecture the equivariant Transformer, which is schematically visualized in Fig. 2. We remark that if all weights W_k^α are 0 in l -th block, the self-attention block (indicated by a purple block in Fig. 2) of l works as an identity operation and it does nothing since the second term in the argument in (6) is zero (See [8] for details).

The long-range correlation in the DE model with SLMC is partially considered in the linear effective model in the literature [12, 14]. In this work, we replace a bare spin operator S_i by the effective spin operator from the Transformer as,

$$H_{\text{eff}}[S] = - \sum_{\langle i,j \rangle_n} J_n^{\text{eff}} S_i^{\text{eff}} \cdot S_j^{\text{eff}} + E_0, \quad (7)$$

where S_i^{eff} is an output the Transformer at site i . A symbol $\langle i,j \rangle_n$ represents the n -th nearest neighbor pair. The effective spin S^{eff} is a function of S and in total, H_{eff} is a function of S . This effective Hamiltonian contains a number of parameters in the Transformer. In SLMC, J_n^{eff} , E_0 and parameters in S_i^{eff} is determined by using AdamW and to increase acceptance ratio.

In our calculation, we apply the MH test (2) using the effective model (7) and the Hamiltonian (3) in a form where the fermions is traced out by the exact diagonalization.

4. Results

Here we show our results. First, we show results for physical observables. Our results show that it is consistent with exact results. As depicted in Fig. 3, the Self-Learning Monte Carlo (SLMC) method with effective models accurately replicates results from the original theory, exhibiting anti-ferromagnetic order at lower temperatures.

The acceptance rate with the number of layers is shown in Fig. 4, left panel. Using effective models trained at $T = 0.05t$ on a 6×6 lattice, we set $N_{\text{MC}}^{\text{original}} = 3 \times 10^4$ and $N_{\text{MC}}^{\text{eff}} = 100$. The linear model SLMC has an acceptance ratio of 21% due to the omission of long-range spin-spin interaction. As observed, the acceptance ratio improves with an increase in Attention layers.

Finally, we show results for the scaling law of the loss function in Fig. 4 (right panel). The value of loss function is estimated from acceptance [15]. It is known that, large language models with Transformer show a power-type scaling law; Model performance increases in depending on the size of the input data and the number of parameters in the model [16]. We find that, our model

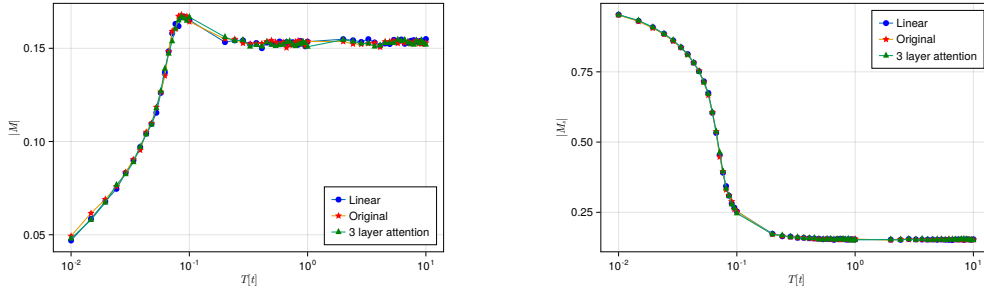


Figure 3: Magnetization (left panel) and staggered magnetization (right panel) for 6×6 lattice sites. For each temperature, we generate 2×10^5 samples using exact diagonalization (red stars), 5000 samples using SLMC with the linear model (green triangles) and the effective model with attention blocks (blue circles).

with equivariant Transformer shows the scaling law. There are no direct relation from our model to large language models and the origin of the scaling law has to be studied in another work.

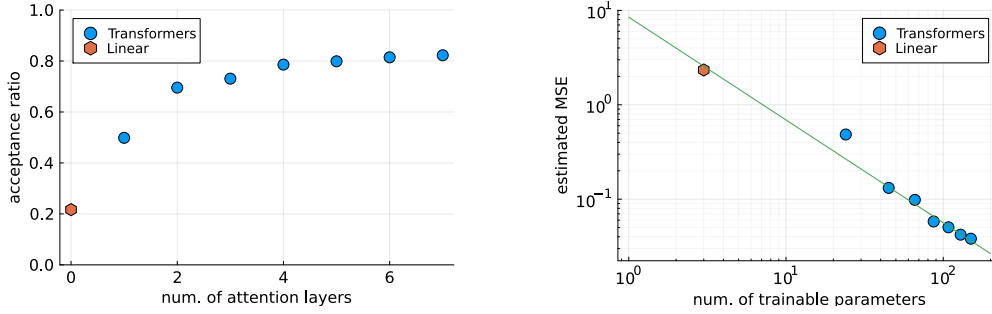


Figure 4: (Left) Acceptance ratio in SLMC: blue square for the linear model and red circles for attention models with $L = 1, 2, \dots, 6$. (Right) MSE vs. trainable parameters: blue square represents the linear model; red circles for models with $L = 1, 2, \dots, 6$. Fitting excludes the linear model and $L = 1$.

Acknowledgments

The work of A.T. was partially by JSPS KAKENHI Grant Numbers 20K14479, 22H05112, and 22H05111. Y.N. was partially supported by JSPS KAKENHI Grant Numbers 22K12052, 22K03539, 22H05111 and 22H05114. The calculations were partially performed using the super-computing system HPE SGI8600 at the Japan Atomic Energy Agency. This work was partially supported by MEXT as “Program for Promoting Researches on the Supercomputer Fugaku” (Grant Number JPMXP1020230411, JPMXP1020230409).

References

- [1] Kyle Cranmer, Gurtej Kanwar, Sébastien Racanière, Danilo J. Rezende, and Phiala E. Shanahan. Advances in machine-learning-based sampling motivated by lattice quantum chromodynamics. *Nature Rev. Phys.*, 5(9):526–535, 2023.

- [2] Yuki Nagai and Akio Tomiya. Gauge covariant neural network for 4 dimensional non-abelian gauge theory. March 2021.
- [3] J. Jumper, R. Evans, A. Pritzel, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589, 2021.
- [4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.
- [5] Philipp Thölke and Gianni De Fabritiis. TorchMD-NET: Equivariant transformers for neural network based molecular potentials. February 2022.
- [6] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.*, 13(1):2453, May 2022.
- [7] Masanobu Horie. *E(n)-Equivariant Graph Neural Networks Emulating Mesh-Discretized Physics*. Doctor of philosophy in engineering, University of Tsukuba, March 2023.
- [8] Yuki Nagai and Akio Tomiya. Self-learning monte carlo with equivariant transformer, 2023.
- [9] Junwei Liu, Yang Qi, Zi Yang Meng, and Liang Fu. Self-learning monte carlo method. *Physical Review B*, 95(4), jan 2017.
- [10] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, et al. Physics-informed machine learning. *Nature Reviews Physics*, 3:422–440, June 2021. Accepted: 31 March 2021, Published: 24 May 2021.
- [11] Kipton Barros and Yasuyuki Kato. Efficient langevin simulation of coupled classical fields and fermions. *Phys. Rev. B*, 88(23):235101, December 2013.
- [12] Junwei Liu, Huitao Shen, Yang Qi, Zi Yang Meng, and Liang Fu. Self-learning monte carlo method and cumulative update in fermion systems. *Phys. Rev. B*, 95(24):241104, June 2017.
- [13] Georgios Stratis, Phillip Weinberg, Tales Imbiriba, Pau Closas, and Adrian E Feiguin. Sample generation for the spin-fermion model using neural networks. *Phys. Rev. B*, 106(20):205112, November 2022.
- [14] Hidehiko Kohshiro and Yuki Nagai. Effective Ruderman–Kittel–Kasuya–Yosida-like interaction in diluted double-exchange model: Self-learning monte carlo approach. *J. Phys. Soc. Jpn.*, 90(3):034711, March 2021.
- [15] Huitao Shen, Junwei Liu, and Liang Fu. Self-learning monte carlo with deep neural networks. *Phys. Rev. B*, 97(20):205140, May 2018.
- [16] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.