Optimal Real-Space Renormalization-Group Transformations with Artificial Neural Networks

Jui-Hui Chung
Department of Physics
National Taiwan University
Taipei 10607, Taiwan

Ying-Jer Kao
Department of Physics
National Taiwan University
Taipei 10607, Taiwan
yjkao@phys.ntu.edu.tw

Abstract

We introduce a general method for optimizing real-space renormalization-group transformations to study the critical properties of a classical system. The scheme is based on minimizing the Kullback-Leibler divergence between the distribution of the system and the normalizing factor of the transformation parametrized by a restricted Boltzmann machine. We compute the thermal critical exponent of the two-dimensional Ising model using the trained optimal projector and obtain a very accurate thermal critical exponent $y_t = 1.0001(11)$ after the first step of the transformation.

1 Introduction

Deep learning (DL) [1] has yielded impressive results in difficult machine learning tasks and various fields of physics [2, 3]. Despite its success, theoretical understanding of the reason behind the surprising effectiveness of DL is still lacking. Although a connection between the renormalization group (RG) and the deep neural networks has been established [4], it is desirable to construct a scheme to enable learning in the RG procedure in order to extract useful information, such as the critical exponents.

Monte Carlo renormalization group (MCRG) [6] is a promising computational scheme for the real-space renormalization group (RSRG). The major source of systematic errors in the MCRG calculations is the lack of convergence due to slow approach to the fixed point. Attempts have been made to introduce variational parameters into the RG transformations with an optimal criterion to bring the fixed point closer to the nearest-neighbor model [7]. The interpretation of why such proposal works, however, remains controversial [8].

In this paper, we propose a general method for optimizing RSRG transformation through divergence minimization using neural network. In our approach, the projection operator is parametrized with a restricted Boltzmann machine and the parameters are chosen to minimize the Kullback-Leibler (KL) divergence between the distribution of the system and the normalizing factor of the projection operator.

2 Related Work

Metha and Schwab [4] established an exact mapping between the variational renormalization group [9] and the deep neural networks based on RBM. The authors then applied deep learning techniques to numerically coarse-grain the two-dimensional nearest-neighbor Ising model on a square lattice, and showed the scheme corresponds to implementing a coarse-graining scheme similar to block spin renormalization [9]. Therefore, they suggested there exists a connection between RG schemes and deep learning algorithms that minimize the Kullback-Leibler (KL) divergence.

Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.

On the other hand, Koch-Janusz and Ringel [10] claimed that training RBMs by minimizing the KL divergence does not perform RG. Instead, they proposed an information-theoretic characterization scheme that maximizes the real-space mutual information (RSMI), which is capable of generating samples of the coarse-grained system. After several RG transformations, they were able to extract the correlation length critical exponent $\nu = 1.0 \pm 0.15 (y_t = 1/\nu = 1.0 \pm 0.15)$. We note, however, although the RSMI algorithm was used to generate a sequence of configurations, they did not use the standard MCRG technique [6] to extract quantitative results from these configurations.

In our work, we demonstrate that applying divergence minimization in training RBM can generate an optimal RG transformation that filters out long-range fluctuations.

3 Optimal Criterion

A generic lattice-model Hamiltonian has the form

$$\mathcal{H}(\sigma) = \sum_{\alpha} K_\alpha S_\alpha(\sigma), \tag{1}$$

where the interactions S_α are combinations of the original spins σ labeled by α and the K_α are the corresponding coupling constants. Consider a generic RG transformation

$$\exp[\mathcal{H}'(\mu)] = \sum_{\sigma} \prod_{\alpha} \mathcal{P}_\alpha(\mu_\alpha, \sigma) \exp[\mathcal{H}(\sigma)], \tag{2}$$

with a parametrized projection operator of the form

$$\mathcal{P}_\alpha(\mu_\alpha, \sigma) = \frac{1}{Y_\alpha} \exp\left\{ \mu_\alpha \sum_i W_{i\alpha} \sigma_i \right\}, \tag{3}$$

where the normalization factor is

$$Y_\alpha = 2 \cosh\left\{ \sum_i W_{i\alpha} \sigma_i \right\}. \tag{4}$$

Here μ_α are the renormalized spins of the renormalized Hamiltonian $\mathcal{H}'(\mu)$. The optimal criterion for choosing the variational parameters $W_{i\alpha}$ will be described in the following section. In particular, if $W_{i\alpha}$ are set to infinity in a local block of spins and zero otherwise, we have the usual majority-rule transformation [11].

To determine the critical exponents, we need to calculate the derivatives of the transformation,

$$T^{(n+1)}_{\alpha\beta} \equiv \frac{\partial K^{(n+1)}_\alpha}{\partial K^{(n)}_\beta}, \tag{5}$$

which is given by the solution of the linear equation [6]

$$\frac{\partial \langle S^{(n+1)}_\gamma \rangle}{\partial K^{(n)}_\beta} = \sum_\alpha \frac{\partial \langle S^{(n+1)}_\gamma \rangle}{\partial K^{(n+1)}_\alpha} \frac{\partial K^{(n+1)}_\alpha}{\partial K^{(n)}_\beta}. \tag{6}$$

Here $\langle S^{(n)}_\gamma \rangle$ is the expectation of the spin combinations at the n-th RG iterations.

To motivate the optimal criterion for choosing the variational parameters, we shall recall the heuristic argument of why DL [1] works so well. It is believed that a RBM [12],

$$p(\mu, \sigma) = \frac{1}{Z} \exp\left\{ \sum_{i\alpha} W_{i\alpha} \sigma_i \mu_\alpha \right\}, \tag{7}$$

parametrized by weights $W_{i\alpha}$ with hidden variables μ_α and visible variables σ_i, works to extract feature distribution $p'(\mu)$ from the data distribution $p(\sigma)$ through

$$p'(\mu) = \sum_\sigma p(\mu|\sigma) p(\sigma), \tag{8}$$

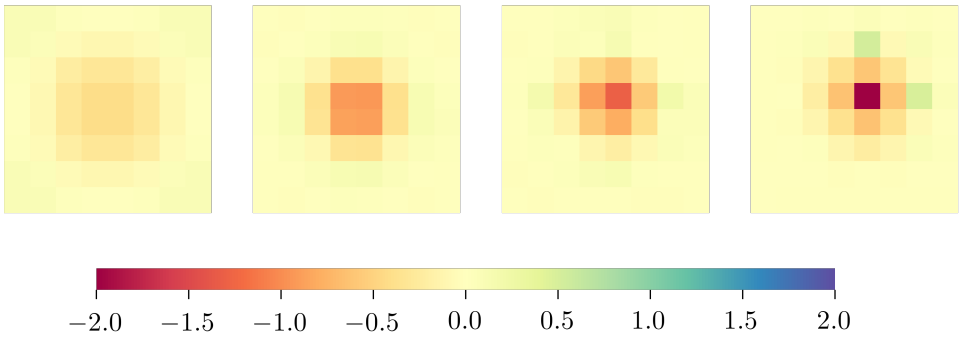

Figure 1: Machine representation of the optimal projection operator for a 2D Ising model. We show the feature maps for the 32×32 Ising model at the critical temperature. From left to right we show the development of the feature map as the training progress (feature maps shown are at the training epoch of $1, 3, 5$ and 50). The feature maps act as effective filters on the spin configurations, capturing the most important correlations.

where $p(\mu|\sigma) \equiv p(\sigma, \mu)/\sum_\mu p(\sigma, \mu)$ is the conditional distribution of the hidden variables given the values of the visible variables. In this scheme, the RBM parameters are chosen to minimize the Kullbach-Leibler (KL) divergence between the data distribution $p(\sigma)$ and the marginal distribution $\sum_\mu p(\mu, \sigma)$

$$D\left(p(\sigma)\middle\|\sum_\mu p(\mu,\sigma)\right), \tag{9}$$

where the KL divergence is defined as $D(p\|q) = \sum_\sigma p(\sigma)\log[p(\sigma)/q(\sigma)]$ for two discrete distribution $p(\sigma)$ and $q(\sigma)$. The KL divergence $D(p\|q)$ is always greater or equal to zero. The equality holds when the two distributions are the same, i.e., $p(\sigma) = q(\sigma)$ for all values of σ. The hope is that the machine will use the hidden variables to extract meaningful features from the data [13].

For this reason, we *identify* the feature-extracting conditional distribution $p(\mu|\sigma)$ with our parametrized projection operator $\mathcal{P}(\mu, \sigma) = \prod_\alpha \mathcal{P}_\alpha(\mu_\alpha, \sigma)$. A sufficient condition is to identify the log-linear part $\sum_{i\alpha} W_{i\alpha}\sigma_i\mu_\alpha$ in the RBM to that in the parametrized projection operator, and associate the hidden and visible variables with the renormalized and original spins, respectively, and correspond the marginal distribution $\sum_\mu p(\mu, \sigma)$ to the normalizing factor $\prod_\alpha Y_\alpha$. In analogy to the RBM unsupervised learning, we propose an optimal criterion by minimizing the KL-divergence between the distribution of the system and the normalizing factor of the projection operator,

$$D\left(\frac{1}{Z}\exp[\mathcal{H}(\sigma)]\middle\|\prod_\alpha Y_\alpha\right). \tag{10}$$

The optimization problem is achieved in the stochastic setting where we use machine learning and contrastive divergence algorithms.

4 Results

To validate our scheme, we consider the problem of finding the thermal critical exponent of the Ising model. The Hamiltonian is

$$\mathcal{H}(\sigma) = K_{\text{nn}}S_{\text{nn}} = K_{\text{nn}}\sum_{\langle ij\rangle}\sigma_i\sigma_j, \tag{11}$$

where $\sigma_i = +1$ or -1 and K_{nn} is the nearest-neighbor coupling. In the following we consider a two dimensional lattice of size 32×32 with periodic boundary conditions.

We prepare a data set with 10^4 binary spin configurations sampled at the critical temperature. We update the parameters with contrastive divergence CD_3 and with an adaptive variant of stochastic gradient descent method called ADAM [14]. The learning rate is initially set to $\eta = 10^{-3}$ and decays during learning. A square root decay is applied to the initial learning rate to reach a finial value of

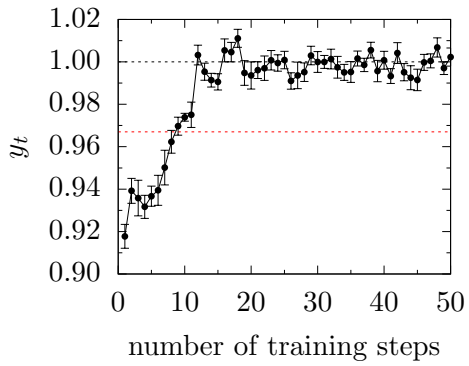

Figure 2: Thermal critical exponents y_t over training steps on a 32×32 lattice for the first step of renormalization. The red (lower) dotted line is the corresponding result $y_t = 0.967(3)$ for the majority-rule transformation [11] and the black (upper) dotted line is the exact value of $y_t = 1$.

Table 1: MCRG estimates for the thermal critical exponents of the 2D Ising model from a simulation on a 32×32 lattice using our optimal projection operator as compared to the majority-rule transformation. N_r is the number of RG steps and N_c is the number of couplings included in the analysis.

N_r	N_c	optimal	majority	N_r	N_c	optimal	majority
1	1	0.9217(09)	0.904(1)	2	1	0.9281(08)	0.953(2)
	2	0.9910(08)	0.966(2)		2	0.9875(08)	0.998(2)
	3	0.9971(09)	0.968(2)		3	0.9977(08)	1.000(2)
	4	1.0004(09)	0.968(2)		4	1.0020(10)	0.998(2)
	5	1.0005(10)	0.968(3)		5	1.0032(11)	0.997(2)
	6	1.0009(10)	0.968(3)		6	1.0018(10)	0.997(2)
	7	1.0001(11)	0.967(3)		7	1.0016(10)	0.997(3)

10^{-4} at the 25th epoch and the rate stays constant for the rest of 25 epochs. The minibatch contains 10 samples and the parameters are initialized uniformly around zero. Seven coupling terms were chosen according to Ref. [11] for the MCRG analysis.

In Fig. 1 we show the optimal machine structure of the projection operator learnt on the Ising model with a filter size of 8×8 and with imposed translational symmetries. We find that the filter learns localized feature which is in agreement with the conventional wisdom that renormalized spins and original spins which are close to one another should be coupled more strongly than others [15]. For example, an extreme case of a localized machine is that of a 2×2 filter with infinite weights, which is equivalent to the typical majority-rule transformation [11]. The behavior of our learnt machine, on the other hand, also shows non-local interactions.

In Fig. 2 we use our optimal machine to calculate the thermal critical exponent for the 32×32 Ising model as a function of the training steps for the first step of renormalization transformation. The data used to compute the thermal exponenent is different from that used in training and consists of 5×10^4 samples. The exponent converges to the exact value (black dashed horizontal line) upon increasing training steps. The most striking result is that although the projection operators are *learnt* without any *prior* knowledge of the system, they are able to generate a renormalization transformation such that the exponent approaches very close to the exact value after the first step of the transformation.

As the data in Table 1 indicates, the optimization performs rather well. The first RG iteration generates an exponent of $y_t = 1.0001(11)$ which is within the statistical error of the exact value, and the second iteration generates $y_t = 1.0016(10)$ which are close to the the exact value. The data consists of 10^6 samples which is drawn independently from the training data set. It is surprising that the machine trained on such small training data of only 10^4 examples is able to *generalize* well and compute statistics based on a data set of 10^6 samples. Table 1 also contains values computed with majority-rule

transformation for comparison [11], giving $y_t = 0.967(3)$ and $y_t = 0.997(3)$ at the first and second RG iteration respectively.

5 Conclusions

We have parametrized our projection operator as an RBM to perform Monte Carlo renormalization group. The optimal criterion for choosing the parameters is proposed to minimize the KL-divergence between the physical distributions and the normalizing factors of the projection operators. It is shown that the set-up is completely equivalent to learning with an RBM.

Spin samples of 2D Ising model produced by Monte Carlo were used to train the machine. Once trained, the machine is used to perform Monte Carlo renormalization group analysis and evaluate critical exponents. We show that the trained projection operator is optimal in that it faithfully reproduces the known exact thermal critical exponent within statistical error at the first step of renormalization transformation.

Our results demonstrate that the divergence minimization criterion may produce optimal convergence in the Monte Carlo renormalization group and may serve as a tool for more challenging problem such as three-dimensional Ising model where the approach to the fixed point upon renormalization is known to be slow. Furthermore, our work may provide a statistical-mechanical point of view to the question of why DL works so well.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, Nature **521**, 436 (2015).

[2] G. Carleo and M. Troyer, Science **355**, 602 (2017).

[3] J. Carrasquilla and R. G. Melko, Nature Physics **13**, 431 (2017).

[4] P. Mehta and D. J. Schwab, arXiv preprint arXiv:1410.3831 (2014).

[5] K. G. Wilson, Physical Review B **4**, 3174 (1971).

[6] R. H. Swendsen, Physical Review Letters **42**, 859 (1979).

[7] R. H. Swendsen, Physical Review Letters **52**, 2321 (1984).

[8] M. E. Fisher and M. Randeria, Physical Review Letters **56**, 2332 (1986).

[9] L. P. Kadanoff, A. Houghton, and M. C. Yalabik, Journal of Statistical Physics **14**, 171 (1976).

[10] M. Koch-Janusz and Z. Ringel, Nature Physics **14**, 578 (2018).

[11] R. H. Swendsen, in *Topics in Current Physics* (Springer Berlin Heidelberg, 1982) pp. 57–86.

[12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning* (MIT press, 2016).

[13] A. Krizhevsky, G. Hinton, *et al.*, *Learning multiple layers of features from tiny images*, Tech. Rep. (Citeseer, 2009).

[14] D. P. Kingma and J. Ba, arXiv preprint arXiv:1412.6980 (2014).

[15] H. Hilhorst, M. Schick, and J. van Leeuwen, Physical Review B **19**, 2749 (1979).