Restricted Boltzmann Machines and the Renormalization Group: Learning Relevant Information in Statistical Physics

Jesse Hoogland

Amsterdam University College

05-06-2019

Introduction

This talk will revolve around the intersection of:

- The Renormalization Group (RG) of Statistical Physics
- Deep Neural Networks (DNNs) in Machine Learning
- Information and Probability Theory

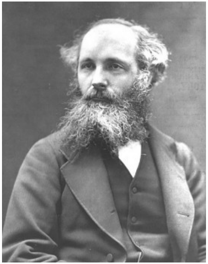
Introduction

- We derive a more *exact* correspondence between RG and RBMs than previous works.
- We provide a new implementation of an existing algorithm that learns *optimal* RG transformations and calculate the Ising model's correlation length critical exponent.
- We describe a generalization of this algorithm to *arbitrary* lattice systems.

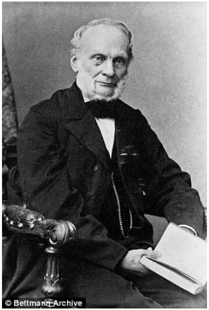
Outline

Introduction to Statistical Physics

(a) Ludwig Boltzmann [1]



(b) James Clerk Maxwell [2]



(c) Rudolf Clausius [3]

The Boltzmann Distribution

$$\boxed{P(\boldsymbol{s}) = \frac{1}{Z} e^{-\beta E(\boldsymbol{s})} \quad Z = \sum_{\boldsymbol{s}} e^{-\beta E(\boldsymbol{s})} \quad \beta = \frac{1}{kT}}$$

Ferromagnetism

Figure 2: Ferromagnetism, the process by which materials like iron form permanent magnets [4].

The Ising Model

$$\boldsymbol{s} = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_N \end{pmatrix} \quad s_i \in \{-1, 1\}$$

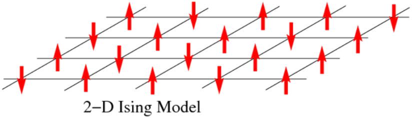


Figure 3: The Ising model [5].

The Ising Hamiltonian

$$\boxed{P(\boldsymbol{s}) = \frac{1}{Z} e^{-\beta E(\boldsymbol{s})} \quad Z = \sum_{\boldsymbol{s}} e^{-\beta E(\boldsymbol{s})}}$$

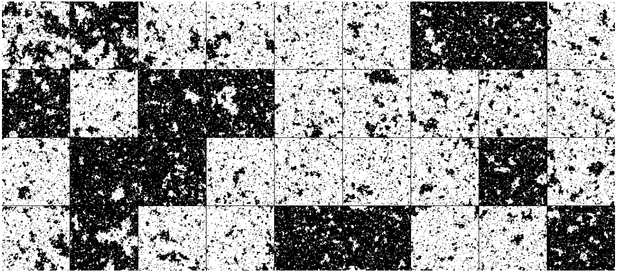
$$\boxed{E(\boldsymbol{s}) = -B \sum_{i} s_i - J \sum_{\langle i,j \rangle} s_i s_j}$$

1. B: The external magnetic field.
2. J: The interaction strength between neighboring spins.

Markov-Chain Monte Carlo (MCMC) Methods

- Perform ensemble average over a subset of (representative) samples.
- Relative probabilities are easier to evaluate than absolute probabilities.

$$P(\boldsymbol{s}) = \frac{1}{Z} e^{-\beta E(\boldsymbol{s})} \quad Z = \sum_{\boldsymbol{s}} e^{-\beta E(\boldsymbol{s})}$$



Mean-Field Theory (MFT)

- Approximates the value at each spin by an average over its neighbors.
- Yields interesting predictions about *critical behavior*.

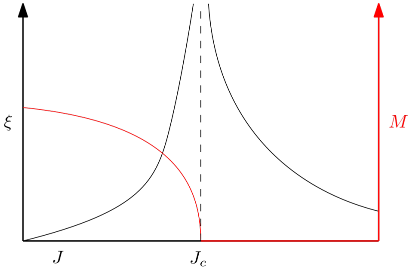


Figure 4: MFT predicts a *critical point*, below which the system *spontaneously magnetizes*. At the critical point, MFT predicts a divergence of the correlation length.

Mean-Field Theory Critical Exponents

Defining t is the *reduced temperature*: $t := (T - T_c)/T_c$, MFT predicts:

$$\boxed{\langle M \rangle|_{B=0} \sim |t|^{-1/2}}$$

$$\boxed{\xi \sim |t|^{-1/2}}$$

Mean-Field Theory Critical Exponents

Defining t is the *reduced temperature*: $t := (T - T_c)/T_c$, MFT predicts:

$$\boxed{\langle M \rangle|_{B=0} \xcancel{\propto} |t|^{-1/2}}$$

$$\boxed{\xi \xcancel{\propto} |t|^{-1/2}}$$

The Renormalization Group

Instead of computing $Z = \sum_{\boldsymbol{s}} e^{-\beta E(\boldsymbol{s})}$ explicitly, try to reexpress Z in a simpler form. We follow Cardy's derivations [6].

$$\boxed{\sum_{\boldsymbol{s}'} e^{-H'(\boldsymbol{s}')} = \sum_{\boldsymbol{s}} e^{-H(\boldsymbol{s})}}$$

For example, *decimation*:

$$e^{-H'(\boldsymbol{s}')} = \sum_{s_2, s_4, \dots, s_N} e^{-H(\boldsymbol{s})},$$

Majority-Rule Block-Spin Renormalization

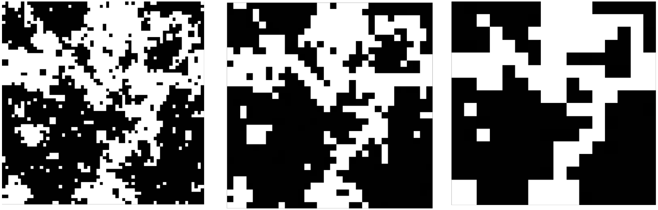


Figure 5: Three steps of majority-rule block-spin renormalization, preceding left to right (block size $b = 2$).

1. Divide configuration into j (3x3) "blocks," $\boldsymbol{v}^{(j)} = (v_1^{(j)}, \cdot, v_9^{(j)}))$.
2. For each block, create a new *coarse-grained* spin h_j, according to the *majority-rule*:

$$\boxed{h_j = \operatorname{sgn} \sum_{i=1}^{9} v_i}$$

3. Rescale our coarse-grained configuration to the original size.

General Theory of RG

Suppose $J' = \mathcal{R}(J)$, then a critical point is such that $J^* = \mathcal{R}(J^*)$. In its vicinity:

$$J' \approx \mathcal{R}(J^*) + \mathcal{R}'(J^*)(J - J^*) = J^* + b^y(J - J^*),$$

where b is the block size and

$$y \equiv \frac{\ln \mathcal{R}'(J^*)}{\ln b}.$$

Knowing that

$$\xi(J) \sim A(J - J^*)^{-\nu},$$

we determine:

$$\boxed{\nu = \frac{1}{y}}$$

Relevant Operators

$$\boxed{y \to \{y_i\}}$$

- $y_i > 0$: **relevant**, repeated RG iterations bring us away from fixed point value.
- $y_i < 0$: **irrelevant**, repeated RG iterations bring us closer to fixed point value.
- $y_i = 0$: **marginal**, linearized equations do not provide enough information.

The Consequences and Implementation of RG

Consequences:

- *RG-flow*: Critical exponents are expressed as derivatives of RG transformations.
- *Scaling relations*: we can express critical exponents in terms of one another.
- *Universality*: there are finitely many fixed points, and many microscopic theories are indistinguishable macroscopically.

In practice:

- $4 - \epsilon$ expansion.
- MCMC methods and finite-size scaling.
- Kadanoff's mnethod $e^{-H'(\boldsymbol{s}')} = \sum_{\boldsymbol{s}} e^{\boldsymbol{T}_\lambda(\boldsymbol{s}',\boldsymbol{s}) - H(\boldsymbol{s})}$.

Partition Functions or Probability Distributions

Statistical Physics

$$H(\boldsymbol{s}) \to Z \to \mathcal{S} \tag{1}$$

$$\frac{P(\boldsymbol{s})}{P(\boldsymbol{s}')} \to \mathcal{S}_{data} \tag{2}$$

Machine Learning

$$\mathcal{S} \to P(\boldsymbol{s}) \tag{3}$$

$$\mathcal{S}_{data} \to P_{\theta}(\boldsymbol{s}) \tag{4}$$

Feed-Forward Neural Networks

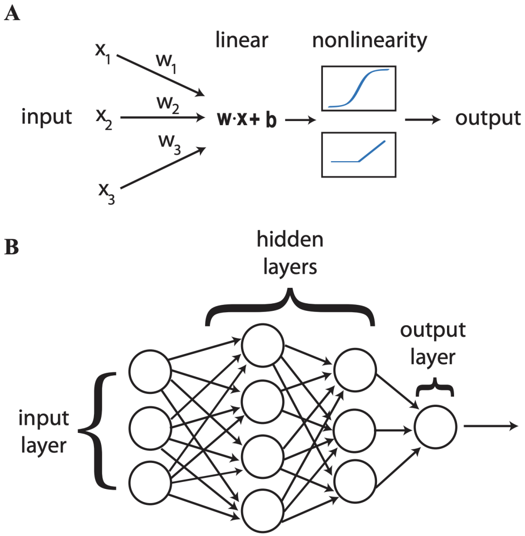


Figure 6: A neural network consists of alternating linear and non-linear transformations [7].

Restricted Boltzmann Machines (RBMs)

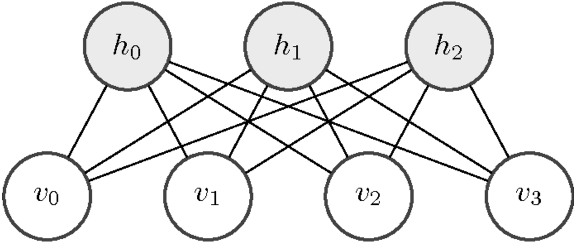


Figure 7: RBMs are a bidirectional neural network of binary-valued units [8].

$$\boxed{E_\theta(\boldsymbol{v}, \boldsymbol{h}) := -\sum_i a_i v_i - \sum_j b_j h_j - \sum_{ij} w_{ij} v_i h_j}$$

$$\boxed{P_\theta(\boldsymbol{v}, \boldsymbol{h}) := \frac{1}{Z} e^{-E_\theta(\boldsymbol{v}, \boldsymbol{h})} \quad Z := \sum_{\boldsymbol{v}', \boldsymbol{h}'} e^{E_\theta(\boldsymbol{v}', \boldsymbol{h}')}}$$

Phase Classifier

$$\boxed{P_\theta(\boldsymbol{h}|\boldsymbol{v}) = \prod_{j=1}^{M} \frac{1}{1 + e^{-h_j(\sum_i w_{ij} v_i + b_j)}}}$$

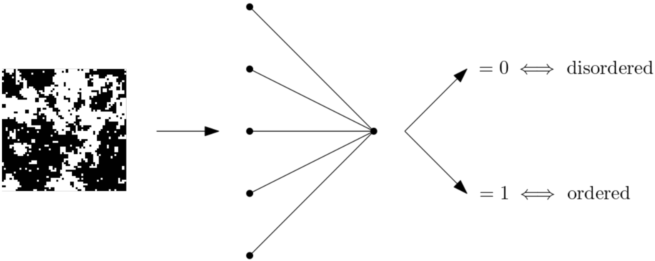


Figure 8: We can use $P_\theta(\boldsymbol{h}|\boldsymbol{v})$ as a phase classifer.

Gibbs Sampling

$$\boxed{P(\boldsymbol{v}) = \sum_{\boldsymbol{h}} P(\boldsymbol{v}, \boldsymbol{h})}$$

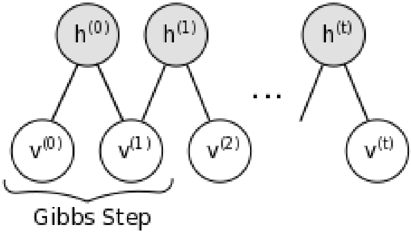


Figure 9: RBMs can implement a MCMC sampling technique known as *Gibbs sampling* [9].

RG = RBM?

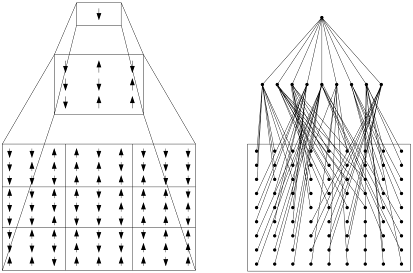


Figure 10: Two iterations of block renormalization and a deep Boltzmann machine of three layers.

An Exact Correspondence between Kadanoff's Variational RG and RBMs

$$\boldsymbol{T}_\lambda(\boldsymbol{v}, \boldsymbol{h}) = -\boldsymbol{E}_\theta(\boldsymbol{v}, \boldsymbol{h}) + \boldsymbol{H}(\boldsymbol{v}),$$

$$P_\theta(x) = P_{\text{true}}(x) \iff Z'_{\text{Kadanoff}} = Z$$

Do RBMs Learn Block Spin RG?

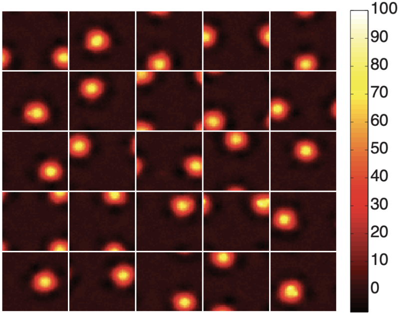


Figure 11: The receptive fields of hidden units in DBMs [10]

.

Surprisingly, this local block spin structure emerges from the training process, suggesting the [deep neural network] is self-organizing to implement block spin renormalization [10].

Do RBMs Learn Block Spin RG?

General block spin tranformations are **NOT** all appropriate RG procedures.

Extra conditions on RG Transformations

[T]he usefulness (and practicality) of the RG procedure depends on choosing [the transformation] ... such that the effective Hamiltonian... remains as short range as possible. [11]

$$\boxed{\sum_{\boldsymbol{s}'} e^{-H'(\boldsymbol{s}')} = \sum_{\boldsymbol{s}} e^{-H(\boldsymbol{s})}}$$

$$H(\boldsymbol{s}) = -\sum_{i} K_i^{(1)} s_i - \sum_{\langle i,j \rangle} K_{ij}^{(2)} s_i s_j - \sum_{\langle\langle i,j \rangle\rangle} K_{ij}^{(3)} s_i s_j \ldots$$

Misconceptions

- Locality
- Translation Invariance
- Physically-Relevant Information

1. Locality

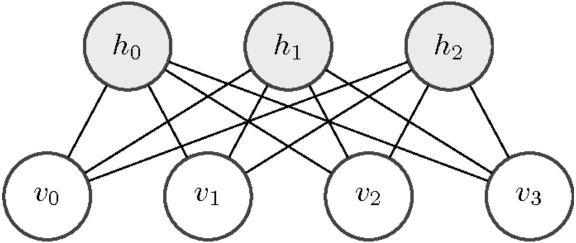


Figure 12: RBMs are invariant under any permutation of the hidden layer [8]

.

2. Translational Invariance

We should apply the same transformation to each block of spins. Consider the following transformations:

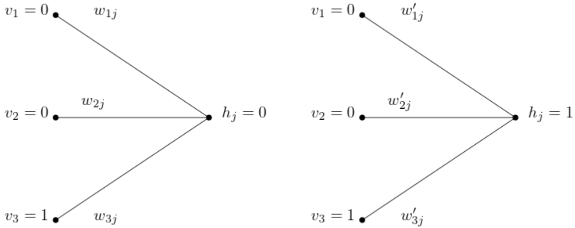


Figure 13: Translation invariance is not respected by compression RBMs, but it is respected by block-spin RG.

Convolutional architecture

We can recover these conditions using a convolutional architecture:

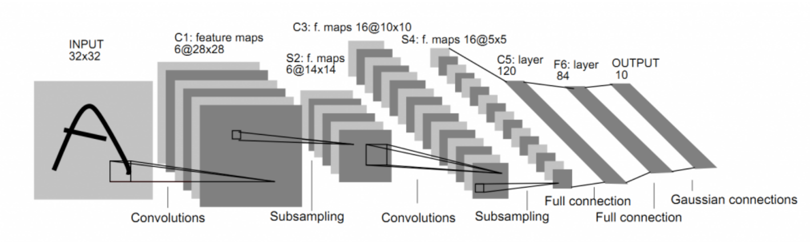


Figure 14: LeNet-5, an example of a convolutional neural network used for digit recognition and a seminal architecture [12].

3. Long-Distance Information

To train compression RBMs, we use the Kullback-Leibler Divergence

$$\boxed{D_{KL}(P_{data}(\boldsymbol{x})||P_\theta(\boldsymbol{x})) = \sum_{x \in \mathcal{X}_{data}} P_{data}(\boldsymbol{x}) \ln\left(\frac{P_{data}(\boldsymbol{x})}{P_\theta(\boldsymbol{x})}\right)}$$

Information Theory and Relevant Information

Relevant information is the information contained in one signal x about another y. This is quantified with the mutual information:

$$\boxed{I(\boldsymbol{x};\boldsymbol{y}) \;= \sum_{\boldsymbol{x},\boldsymbol{y}} P(\boldsymbol{x},\boldsymbol{y}) \log\left(\frac{P(\boldsymbol{x},\boldsymbol{y})}{P(\boldsymbol{x})P(\boldsymbol{y})}\right)}$$

The Real-Space Mutual Information (RSMI) Maximization Algorithm

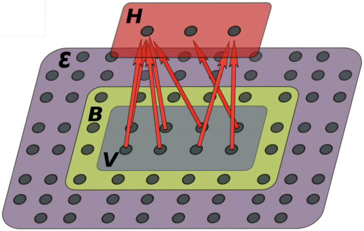


Figure 15: We partition the system into a visible block, buffer zone, and environmental area.[11]

$$\boxed{I(\boldsymbol{h};\boldsymbol{e}) = \sum_{\boldsymbol{h},\boldsymbol{e}} P(\boldsymbol{h},\boldsymbol{e}) \log\left(\frac{P(\boldsymbol{h},\boldsymbol{e})}{P(\boldsymbol{h})P(\boldsymbol{e})}\right)}$$

A Recalculation of the Correlation-Length Critical Exponent

$$\boxed{\nu \approx 0.79 \pm 0.39}$$

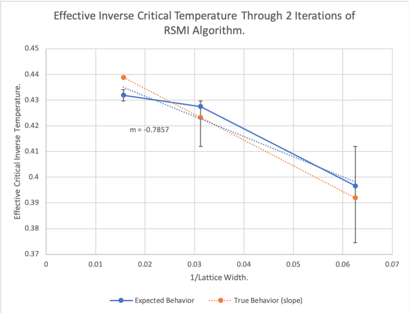


Figure 16: Finite-size collapse curve for our implementation of the RSMI algorithm.

Generalization to n-Spin and $O(n)$ Models

Models	Symmetry of order parameter	α	β	γ	δ	ν	η
$2-d$ Ising	2-component scalar	0 (log)	1/8	7/4	15	1	1/4
$3-d$ Ising	2-component scalar	0.10	0.33	1.24	4.8	0.63	0.04
$2-d$ Potts, $q=3$	q-component scalar	1/3	1/9	13/9	14	5/6	4/15
$3-d$ X-Y	$2-d$ vector	0.01	0.34	1.30	4.8	0.66	0.04
$3-d$ Heisenberg	$3-d$ vector	-0.12	0.36	1.39	4.8	0.71	0.04
Mean field		0 (dis)	1/2	1	3	1/2	0

Figure 17: Universality classes for different numbers of dimensions d and spin components n [13].

Discussion and Conclusions

- Information theory as conceptual framework for comparing ML and RG and devising *optimal* procedures: e.g. the RSMI algorithm.
- Symmetries of our systems as restricting allowed RG and ML transformations and enabling understanding of "black box" neural networks: e.g. convolutional architectures.
- (Unsupervised) Machine learning as guiding the "physical reasoning process," going beyond data analysis alone: e.g. calculating critical exponents [11].

References

[1] JT. Top 10 scientists who committed suicide, oct 2007.

[2] Who2 Biographies. James clerk maxwell biography.

[3] Sarah Griffiths. From child prodigies to playwrights, the world's 40 smartest people of all time revealed, Aug 2016.

[4] Eurico Zimbres FGEL/UERJ. Magnetez, 2005.

[5] JGTechSol. Optical computing, Feb 2017.

[6] John Cardy. *Scaling and Renormaliztion in Statistical Physics*. Cambridge University Press, Cambridge, United Kingdom, 1996.

[7] Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre G. R. Day, Clint Richardson, Charles K. Fisher, and David J. Schwab. A high-bias, low-variance introduction to machine learning for physicists. *arXiv e-prints*, page arXiv:1803.08823, Mar 2018.

[8] Marc-Alexandre Côté and Hugo Larochelle. An infinite restricted boltzmann machine. *Neural Computation*, 28:1265–1288, 2016.

[9] Theano Development Team. Restricted boltzmann machines, 2013.

[10] Pankaj Mehta and David J. Schwab. An exact mapping between the variational renormalization group and deep learning. *arXiv e-prints*, page arXiv:1410.3831, Oct 2014.

[11] Maciej Koch-Janusz and Zohar Ringel. Mutual information, neural networks and the renormalization group. *Nature Physics*, 14:578–582, Jun 2018.

[12] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[13] Sitangshu B. Santra. Advanced statistical mechanics - models and universality, November 2013.