# Machine Learning and the Renormalization Group

Zhiru Liu*
Term Essay for PHYS 563†
University of Illinois at Urbana-Champaign

**Abstract**

In this essay, we reviewed the recent attempts on relating Machine Learning to Renormalization Group. Restricted Boltzmann Machines, a type of neural network, was shown to be connected to Variational Renormalization Group. A treatment of Principal Component Analysis analogous to momentum shell Renormalization Group uncovered a possible fixed point.

*zliu106@illinois.edu
†http://guava.physics.uiuc.edu/ nigel/courses/563/

# 1 Introduction

The resurgence and success in machine learning (ML) have started a trend to combine physics and ML in physics community. Some borrowed ideas from ML and modified them for physics contexts[1, 2], while others could not help but wonder: could physics help explain why ML works? Works thus emerged trying to examine ML from a Renormalization Group (RG) perspective[3, 4, 5, 6]. We will focus on two of the methods(model) studied, namely Restricted Boltzmann Machines (RBM) and Principle Components Analysis (PCS), which are connected to variational and momentum space RG respectively.

The essay is organized as follows: we first introduce RBM and PCA and how they are used in typical ML contexts; in section 3 we present the connections between these methods and RG; then we pose several questions and potential problems of the articles; lastly we briefly comment on the other two works.

# 2 A Crash Course in ML

## 2.1 Restricted Boltzmann Machines

One of the central goal of ML is to extract information from a collection of high dimensional data. A straightforward approach would be to model the data as the probability distribution of some high dimensional random vector. Among all the choices of probability distribution, RBM stands out due to its simplicity. Formally, RBM is specified with the following joint distribution function[7]:

$$P(\{v_i\},\{h_i\}) = \frac{e^{-\boldsymbol{E}(\{v_i\},\{h_i\})}}{\mathcal{Z}} \tag{1}$$

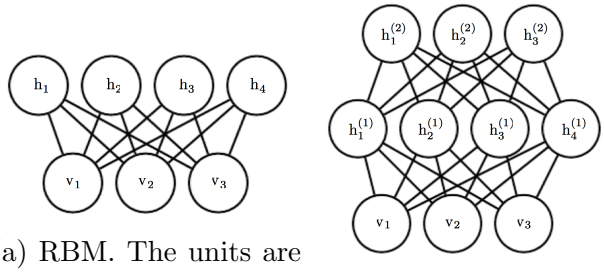
where

$$\boldsymbol{E}(\{v_i\},\{h_i\}) = \sum_i b_i h_i + \sum_i c_i v_i + \sum_{ij} v_i W_{ij} h_j \tag{2}$$

and

$$\mathcal{Z} = \mathrm{Tr}_{v_i,h_i}\, e^{-\boldsymbol{E}(\{v_i\},\{h_i\})} \tag{3}$$

In the above equations, $\{v_i\}$ represents the random vector that we observe, known as visible units, and $\{h_i\}$ is a auxiliary vector known as hidden units. All units take values $\pm 1$, like Ising spins. Because of the notions of energy




(a) RBM. The units are grouped into two layers.

(b) DBM. Created from two RBM

Figure 2: Graphical representation of RBM and DBM. Lines connecting two units represent the interactions $W_{ij}$.[7]

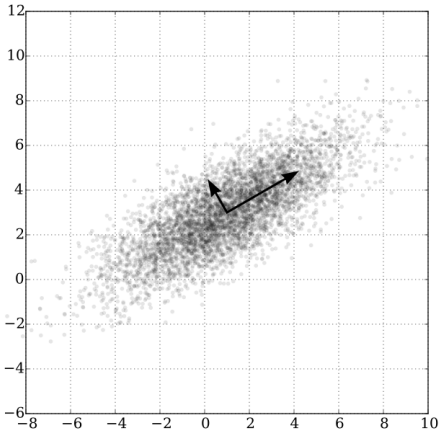



Figure 3: PCA of a multivariate Gaussian distribution. The arrows show two eigenmodes. (Author: Nicoguaro)

and partition function, RBM belongs to the class of energy-based model. Physically, this model describes the full interaction between the visible layer and the hidden layer. The model can be represent graphically (see fig.1a).

One might wonder why we take the detour to introduce hidden units, instead of allowing interactions within the visible layer directly. One advantage of RBM is that it serves as a transformation of original data. Suppose we have found (or trained) the best parameters $(\{b_i\}, \{c_i\}, \{W_{ij}\})$ that capture the data distribution. Then equivalently we also obtained a distribution for the hidden units,

$$P(\{h_i\}) = \mathrm{Tr}_{v_i} P(\{v_i\}, \{h_i\}) = \mathrm{Tr}_{v_i} \frac{e^{-\boldsymbol{E}(\{v_i\},\{h_i\})}}{\mathcal{Z}} \equiv \frac{e^{-\boldsymbol{H}_h(\{h_i\})}}{\mathcal{Z}} \tag{4}$$

By sampling the distribution, for example, we obtain a new set of data which are transformed version of observed data. One can then feed the data into another RBM, and repeat the process. The resulting multi-layer, deep architecture model is known as Deep Boltzmann Machine (DBM) (See fig1b). After the above "preprocessing", one can apply other ML algorithms on the transformed data. As we will see, such transformation could be connected to a Variational RG transformation.

## 2.2 Principle Component Analysis

The other approach to analyze a high dimensional data set is to reduce its dimension. Following the conjecture that data might distribute along some hyperplane, we seek a linear projection that best retains the original information. PCA is a way of doing so.

Given $n$ samples of a $m$-dimensional vector $\boldsymbol{\phi}$, we can compute its covariance matrix $\boldsymbol{A}$, and then perform a eigendecomposition of the matrix, $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$. In this way, we have discomposed the variations of the data into modes that are independent at second order; the modes are called principal components (PCs). By applying the orthogonal matrix $\boldsymbol{Q}^T$ to $\boldsymbol{\phi}$, we transform it to $\tilde{\boldsymbol{\phi}}$ which resides in the space of eigenmodes. If we only choose the first $l$ eigenvectors (those with largest eigenvalues) as the basis of projection, the resulting $\tilde{\boldsymbol{\phi}}'$ ($l$-dimensional) will retain the information of the most important modes.

Let's consider a simple example of PCA. Suppose we are given some data generated by multivariate Guassian distribution, and we want to understand them in just one dimension. Then a quick PCA would reveal the two eigenmodes of our data, as shown in figure 3. Projecting the data onto the major axis of the ellipse, we would observe that the one dimension distribution obeys Gaussian distribution.

Suppose $\boldsymbol{\phi}$ has some complex distribution $P(\boldsymbol{\phi})$. Performing the projection actually is equivalent to marginalize the lowest eigenmodes in the distribution:

$$P(\tilde{\boldsymbol{\phi}}) = \prod_{i=l}^{n} \sum_{\phi_i} P(\{\widetilde{\phi_i}\}) \tag{5}$$

This "integrating out some modes" approach resembles momentum space RG we learned in class, and in later section we'll see the implications of this resemblance.

# 3 Connecting RG with RBM

## 3.1 Variational RG

We start by slightly modifying the notion of real space RG used in class. As usual, the Hamiltonian describing a spin system is specified by a set of

coupling parameters, $\boldsymbol{K} = \{K_s\}$,

$$\boldsymbol{H}(\{v_i\}) = \sum_i K_i v_i + \sum_{ij} K_{ij} v_i v_j + ... \tag{6}$$

After a RG transformation $\boldsymbol{R}$, the new effective Hamiltonian, $\boldsymbol{H'} = \boldsymbol{R}[\boldsymbol{H}]$, is specified by $\tilde{\boldsymbol{K}} = \{\tilde{K}_s\}$,

$$\boldsymbol{H'}(\{h_i\}) = \sum_i \tilde{K}_i h_i + \sum_{ij} \tilde{K}_{ij} h_i h_j + ... \tag{7}$$

And the two Hamiltonian is connected by requiring partition function to be the same:

$$\mathcal{Z} = \mathrm{Tr}_{v_i}\, e^{-\boldsymbol{H}} = \mathrm{Tr}_{h_i}\, e^{-\boldsymbol{H'}} \tag{8}$$

Now let's introduce an general function with arbitrary form, $\boldsymbol{T}_\lambda(\{v_i\}, \{h_i\})$, which depends on some parameter $\lambda$, and rewrite the effective Hamiltonian as

$$e^{-\boldsymbol{H'}(\{h_i\})} = \mathrm{Tr}_{v_i}\, e^{\boldsymbol{T}_\lambda(\{v_i\},\{h_i\}) - \boldsymbol{H}(\{v_i\})} \tag{9}$$

In the form, the RG transformation is entirely encoded in the function $\boldsymbol{T}_\lambda$. By choosing convenient forms of $\boldsymbol{T}_\lambda$ we are able to analyze the transformation, even when it's not exact (which means that equation 8 might not holds). We can also optimize the parameter $\lambda$ variationally to approximate some exact RG transformation. Therefore, function $\boldsymbol{T}_\lambda$ defines a Variational RG transformation.

## 3.2 The Mapping

The crucial observation made by authors of [3] is the following: if we interpret the spins before and after VRG transformation as visible and hidden units, then the energy function in equation 2 defines a VRG transformation through

$$\boldsymbol{E}(\{v_i\}, \{h_i\}) = -\boldsymbol{T}_\lambda(\{v_i\}, \{h_i\}) + \boldsymbol{H}(\{v_i\}) \tag{10}$$

Then, evidently equation 9 and 4 are describing the same transformation (One can check by dividing 9 by $\mathcal{Z}$).

We pause for a moment to clarify what we mean by transformation. We assumed that both $\{v_i\}$ and $\{h_i\}$ obey Boltzmann distribution, and thus they are accompanied by Hamiltonians $\boldsymbol{H}$ and $\boldsymbol{H'}$; further, Hamiltonians are

determined by coupling parameters $\boldsymbol{K}$ and $\boldsymbol{K'}$, which are infinite dimensional vectors. Therefore, by transformation we mean the mapping $\boldsymbol{K} \to \boldsymbol{K'}$.

This observation thus provides us an interpretation of what the aforementioned "preprocessing" does: it performs a VRG transformation on the observed data set. The more layers of RBM we stack together, the more rounds of VRG transformation we are performing. By using learning algorithms to optimize parameters $(\{b_i\}, \{c_i\}, \{W_{ij}\})$, we are effectively varying $\lambda$ for $\boldsymbol{T}_\lambda$.

### 3.3 Learning the Ising Model

The authors then conducted a numerical experiment where they feed two dimensional spin configurations, sampled from a Ising model near $T_c$, into a DBN. They plotted the learned parameter $W_{ij}$ to explicitly show how each hidden unit interact with the visible units, shown in fig.4. They claimed that the network learned block spin RG.

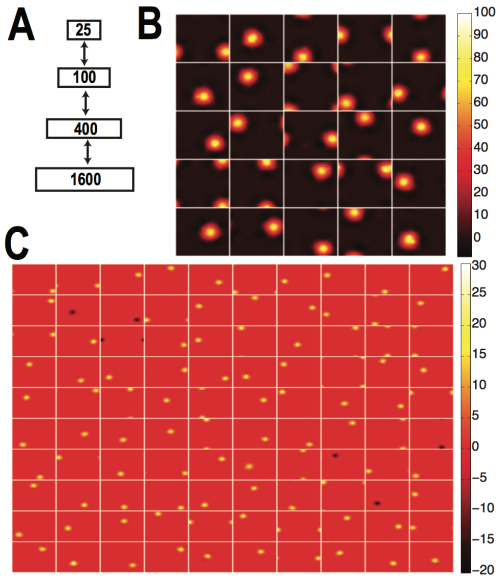



Figure 4: The parameters of learning Ising model. A: Archetecture of the network. B,C: Plots of $W_{ij}$

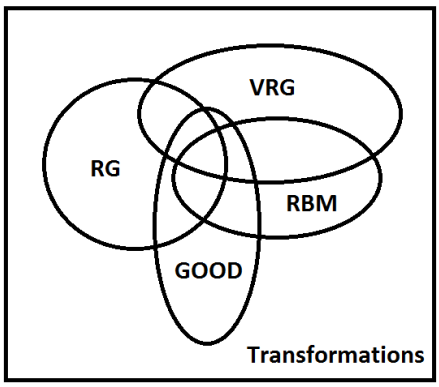



Figure 5: A Venn diagram illustrating relations between mentioned transformations.

# 4 Connecting RG with PCA

In this section we will present the treatment of PCA in [5], and show that the resemblance mentioned in equation 5 indeed can be exploited further using the techniques of momentum shell RG.

## 4.1 A Quick Review of Momentum Space RG

To illustrate the connection, we begin from a sketch of Momentum shell RG for Ising universality class[8]. Instead of starting with microscopic Hamiltonian in the real space RG approach, we start with a effective Hamiltonian in Landau theory:

$$-\boldsymbol{H}\{\boldsymbol{S}\} = \int d^d\boldsymbol{r}\Big[\frac{1}{2}\nabla^2\boldsymbol{S} + \frac{1}{2}r_0\boldsymbol{S}^2 + \frac{1}{4}u_0\boldsymbol{S}^4\Big] \tag{11}$$

where S is a function defined on $\boldsymbol{R}^d$. In momentum space, the Hamiltonian can be written as,

$$-\boldsymbol{H}\{\boldsymbol{S'}\} = \int_0^\Lambda \frac{d^d\boldsymbol{k}}{(2\pi)^d}\mathcal{L}\big[\boldsymbol{S'}(\boldsymbol{k})\big] \tag{12}$$

The idea is to integrate out the highest frequency modes in the range $\frac{\Lambda}{l} < |\boldsymbol{k}| < \Lambda$, which has the effect of coarse-graining the lattice by scale $l$. By rescaling the lattice back to scale $\Lambda$ and comparing the form of $\boldsymbol{H}$, we can obtain the RG recurrence relation of $r_0$ and $u_0$ after some nasty but systematic calculation. In particular, the non-Gaussian coupling constant $u_0$ has the following relation:

$$\frac{du_s}{ds} = (4-d)u_s - Au_s^2 \tag{13}$$

where A is some hard-to-calculate constant. From this one immediately find two fixed points, namely

$$u^* = 0, u^* = \frac{4-d}{A} \tag{14}$$

which correspond to Gaussian fixed point and Wilson-Fisher fixed point respectively.

## 4.2 A Momentum Flavor Treatment of PCA

Imagine that we have observation data for a $N$-dimensional random vector $\boldsymbol{\phi}$ (a collection of N random variables). In analogy to the proceeding section, we begin by modeling the distribution as

$$P(\boldsymbol{\phi}) = \frac{e^{-\boldsymbol{H}\{\boldsymbol{\phi}\}}}{\mathcal{Z}} \tag{15}$$

where $\mathcal{Z}$ is defined in the common sense and

$$\boldsymbol{H}\{\boldsymbol{\phi}\} = \sum_{ij} \phi_i K_{ij} \phi_j + g \sum_i \phi_i^4 + ... \tag{16}$$

This Hamiltonian based on Gaussian distribution with a $\phi^4$ correction is a minimal model, just like the above Landau effective Hamiltonian. In fact, equation 16 is nothing more than a discrete version of 11.

$$\int d^d \boldsymbol{r} \nabla^2 \boldsymbol{S} \to \sum_{ij} \phi_i K_{ij} \phi_j \tag{17}$$

The principal components (eigenmodes of $\boldsymbol{K}$) correspond to momentum (Fourier modes). With that in mind, we proceed just as in momentum RG. First transform into eigenmodes space,

$$\sum_{ij} \phi_i K_{ij} \phi_j = \sum_{\mu} \lambda_\mu \tilde{\phi}_\mu^2 \tag{18}$$

and introduce a distribution for eigenvalues

$$\rho(\lambda) = \frac{1}{N} \sum_{\mu} \delta(\lambda - \lambda_\mu) \tag{19}$$

We can then convert the sum into integral

$$\sum_{\mu} \lambda_\mu = \int_0^\Lambda d\lambda \rho(\lambda) \lambda \tag{20}$$

In this form, our minimal model is of course very similar to the previous section. The authors of [5] continue their calculation under the limit of small $\tilde{g}$, the $\phi^4$ coupling constant, and arrived at the following recurrence relation:

$$\tilde{g} \to \tilde{g} - B\tilde{g}^2 \tag{21}$$

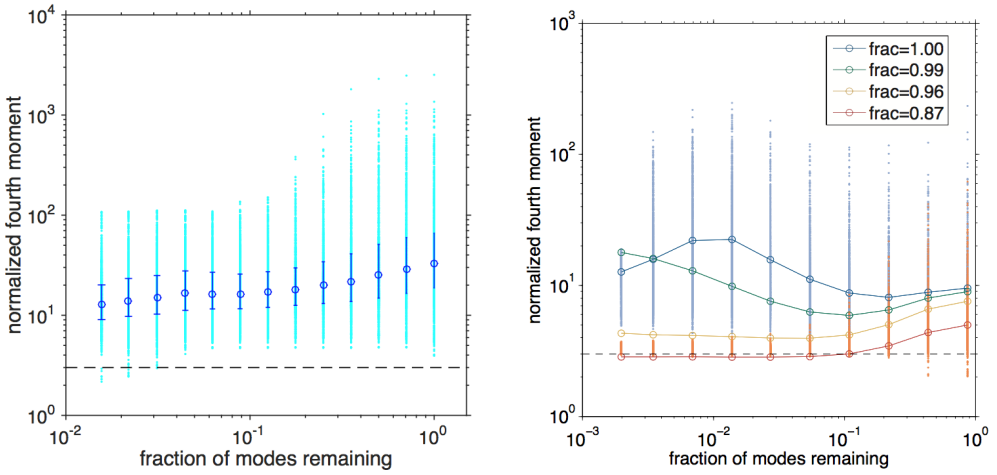
where B is some constant that depends on the distribution of eigenvalues of $\boldsymbol{K}$. Remarkably, this indicates the existence of two fixed points. One is the familiar $\tilde{g} = 0$, i.e. Gaussian fixed point. The nontrivial one is analogous to the Wilson-Fisher fixed point. In particular, if $\rho(\lambda) \sim \lambda^{(\alpha-1)}$

$$\tilde{g}^* \propto \alpha - 2 \tag{22}$$

## 4.3 Pinning the Fixed Points in Data

The authors tried to illustrate how the fixed points play their roles by analyzing data from neural activity and stock market. In the figures 6a and 6b, they use the normalized fourth moments of the data, $(\langle\phi_i^4\rangle/\langle\phi_i^2\rangle^2)$, as an indicator of the effect of $\phi^4$ coupling.

For the first data set, the fourth moment remained nearly unchanged and far away from Gaussian (dashed line) as the lowest eigenmodes were thrown away, suggesting that the neural system might be close to a critical point. For the second set, they found that the largest eigenmodes determines where the fourth moment flows to. For example, if the largest 10 percent were removed, the fourth moment quickly would flow to that value for Gaussian. However, if all modes are kept, the flow is more complex and non trivial.




(a) Neural Activity (b) Stock Market

Figure 6: Analysis of data with PCA. The circles represent median of the normalized fourth moments. The entire spectrum is plotted for each fraction.

# 5 Discussion

Thus far, we have done a coarse-grained summary of the connections. We will begin to pose some points not addressed in the papers.

## 5.1 VRG-RBM

**Transformations are not created equal** We first wish to point out that VRG transformations, RBMs, RG transformations and "good" transformations are all different sets (a schematic Venn diagram might help 5). Apparently, RBMs/VRG are variational, and thus different from RG. RBMs are different from VRG transformations because the optimization goal of RBM is:

$$P(\{v_i\}) = \frac{e^{-\boldsymbol{H}_{v|\lambda}(\{v_i\})}}{\mathcal{Z}} = \mathrm{Tr}_{h_i} \frac{e^{-\boldsymbol{E}_\lambda(\{v_i\},\{h_i\})}}{\mathcal{Z}} \tag{23}$$

yet the goal of VRG is:

$$\mathcal{Z} = \mathrm{Tr}_{h_i}\, e^{-\boldsymbol{H}'_{h|\lambda}\{h_i\}} = \mathrm{Tr}_{h_i,v_i}\, e^{-\boldsymbol{E}_\lambda(\{v_i\},\{h_i\})} \tag{24}$$

In words, one tried to fit the microscopic distribution exactly, while the other wished to maintain the partition function, or macroscopic observable (free energy).

The true problem is, RG/VRG/RBM are all too general. RG might not be able to performed easily, which is why we introduce VRG. We could lift the restriction of RBM, and allow interaction within hidden layer and visible layer (Boltzmann Machine), and it still could be mapped to VRG. We could write all sorts of crazy $\boldsymbol{T}_\lambda$ with no practical use: remember, we still need to trace out all $\{v_i\}$.

RG approach displays its power when a "good" transformation yields a simple or tractable $\boldsymbol{K'} = \boldsymbol{R}[\boldsymbol{K}]$. And then we could linearize the recurrence relation and analyze the RG flow, without actually tracing all degrees of freedom. However, the mapping uncovered by the authors of [3] does not show that RBM transformations are "good" or not, and the powerful tools in RG cannot be applied. [6] provides an amusing example: let $\{h_i\}$ be completely independent from $\{v_i\}$, and the transformation is still a VRG. To summarize, merely bearing the name of RG is not worthy of excitement.

**Block Spin or correlation?** The numerical experiment, however, is interesting in the sense that it hints the RBM might learned a good RG transformation: block spin. Yet two things remain unclear. First, the goal of RBM is completely different from goal of RG, so why should we attribute the interaction to RG but not the distribution itself? As a alternate, could the blocks just stands for correlations between spins? It's possible that the block in $\{W_{ij}\}$ is just saying units in a block tends to align, instead of actively coarse-graining the configuration. The second question can be tested by changing the temperature of the model, and see if the block size change. We are interested to do so in the future.

## 5.2 PCA-RG

**Most Important?** In the treatment of momentum RG, we integrate out the shell of highest momentum, so that the microscopic variation are averaged out. In this sense, the highest momentum modes are least important for the calculation of partition function. On the contrary, in PCA, the highest modes are the most important ones. Thus, we cannot apply the intuition that changing scale of $\Lambda$ is integrating out unimportant information in PCA.

**The legitimacy of approximation** The calculations of fixed point, though not shown in this essay, depends heavily on the assumption that $g$ is small. Analogy in momentum RG is the assumption that $\epsilon = 4 - d$ is small. Notice in the latter case, the action to take $\epsilon = 1$ is justified through techniques from asymptotic analysis of the perturbation series. However, there's no calculation of the magnitude of $g$ in real data, and no justification for cases when $g$ is not small. This puts the consistency of the calculation in question.

**A possible connection to bigger world** Recently, people have found that many system, modeled by a large number of parameters, are insensitive to the changes in parameter space in a wide range of directions[9]. This property is termed "slopiness". This property can be quantified by the spectrum of eigenvalues of Fisher information matrix (FIM), and in many cases, the spectrum distributes uniformly, $\rho(\lambda) \sim 1$, which corresponds to $\alpha = 1$ in the relation preceding equation 22.

Since the critical dimension of Ising universality class is fundamentally due to the symmetry requirement, we wonder whether the fixed point in

eq.22 hints certain underlying symmetry for the "sloppy universality" class, if it exists in those data sets. Curiously, in a recent work[10], it is shown that upon compression of parameter space, conventional statistical physics models develop sloppiness. This discovery again seems to suggest a connection between RG and this new universality class.

We end the discussion about PCA and sloppy model by pointing out a connection between FIM and covariance matrix: the inverse of FIM is used as a Maximum Likelihood estimator for covariance matrix[11].

### 5.3 Other Attempts

**RG and MERA** In [4], the author attempted to relate a RG-based numerical method to a proposed deep architecture. However, the numerical method, Multiscale Entanglement Renormalization Ansatz (MERA) is used in quantum system, and is less like RG learned in class. In addition, the proposed network is not used in practice. Therefore, we choose not to discuss this paper with more detail.

**RG and Deep Learning in General** [6] spends lots of pages trying to argue the necessity of deepness in networks from information theory. Yet, its connection to RG stopped at analogy level and appeared less concrete than other selected works. Their comments on [3] in this paper did provide valuable insights and help us clarify the ideas of [3].

## 6 Conclusion

There's been a surge of paper in the interface between machine learning and physics, not restricted to RG. We are optimistic about the future of this interdisciplinary trend, and we are excited to see new physics and applications emerge from the frontier.

## Acknowledgement

The author is grateful to PHYS 563, without which [3, 4, 5, 6, 8] could not be understood. This essay was entirely supported by author's kind parents.

# References


[1] Giuseppe Carleo and Matthias Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355(6325):602–606, feb 2017.

[2] Evert P. L. van Nieuwenburg, Ye-Hua Liu, and Sebastian D. Huber. Learning phase transitions by confusion. *Nature Physics*, 13(5):435–439, feb 2017.

[3] Pankaj Mehta and David J. Schwab. An exact mapping between the variational renormalization group and deep learning, 2014.

[4] Cédric Bény. Deep learning and the renormalization group, 2013.

[5] Serena Bradde and William Bialek. PCA meets RG. *Journal of Statistical Physics*, 167(3-4):462–475, mar 2017.

[6] Henry W. Lin, Max Tegmark, and David Rolnick. Why does deep and cheap learning work so well?, 2016.

[7] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[8] Nigel Goldenfeld. *Lectures On Phase Transitions And The Renormalization Group (Frontiers in Physics)*. Addison-Wesley, 1992.

[9] Joshua J. Waterfall, Fergal P. Casey, Ryan N. Gutenkunst, Kevin S. Brown, Christopher R. Myers, Piet W. Brouwer, Veit Elser, and James P. Sethna. Sloppy-model universality class and the vandermonde matrix. *Physical Review Letters*, 97(15), oct 2006.

[10] B. B. Machta, R. Chachra, M. K. Transtrum, and J. P. Sethna. Parameter space compression underlies emergent theories and predictive models. *Science*, 342(6158):604–607, oct 2013.

[11] Markus Abt and William J Welch. Fisher information and maximum-likelihood estimation of covariance parameters in gaussian stochastic processes. *Canadian Journal of Statistics*, 26(1):127–137, 1998.