Renormalization Group and Information Theory

Kyle Reing
University of Southern California

April 18, 2018

USC Viterbi
School of Engineering

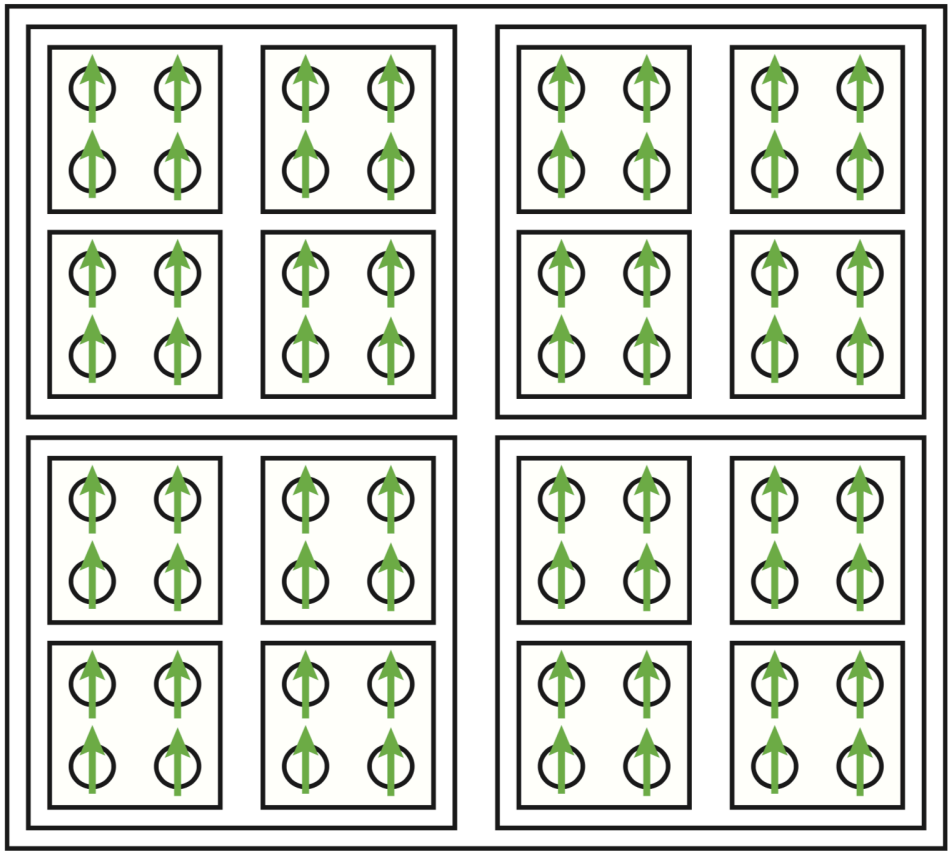
Information Sciences Institute

Overview

- Renormalization Group Overview
- Information Theoretic Preliminaries
- Real Space Mutual Information
- Better ways to do this?
- The ML and Physics Friendship

Renormalization Group

- Example: Block Spin Renormalization Group



Renormalization Group

- Describe system in terms of block variables (ex: average spin of the block)
- Use the 'same form' for the Hamiltonian, except defined over blocks, with different values for the parameters (ex: coupling strength)
- Change in parameters after renormalization defines a RG Flow
- Limit of flow often leads to fixed point characterizations (ex: Ising ferromagnetic/ paramagnetic)

Another Perspective

- Sufficient Statistic: A statistic (function) of the data that contains all of the information with respect to some model (ex: mean/ variance with Gaussians)
- Can think of RG as a (lossy) compression, for some course graining operation R, data X, and measure of 'macro scale properties' Y, want:

$$R^n = R \circ R \circ R \circ R \circ ...$$

$$I(Y : X) = I(Y : R^n \circ X)$$

Super Quick Info Theory Definitions

- (Shannon) Entropy
  - Measure of uncertainty in a random variable, how 'spread out' is the distribution?

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$ **Discrete**

$$= -\int_{x \in X} p(x) \log p(x)$$ **Continuous**

- Mutual Information
  - Measure of similarity between two random variables, reduction in uncertainty when we know another variable

$$I(X:Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(y)p(x)}$$

$$= H(X) - H(X|Y)$$

Learning Relevant Degrees of Freedom?

- For a given model, RG procedure is often explicitly defined (what are the fast/ slow degrees of freedom, what is the course graining procedure, how many steps to take, what is the relevant macro scale property?)

- What if we wanted to learn the relevant degrees of freedom and course graining automatically in a data driven way?

nature physics

ARTICLES

https://doi.org/10.1038/s41567-018-0081-4

Mutual information, neural networks and the renormalization group

Maciej Koch-Janusz[1*] and Zohar Ringel[2]

Physical systems differing in their microscopic details often display strikingly similar behaviour when probed at macroscopic scales. Those universal properties, largely determining their physical characteristics, are revealed by the powerful renormalization group (RG) procedure, which systematically retains 'slow' degrees of freedom and integrates out the rest. However, the important degrees of freedom may be difficult to identify. Here we demonstrate a machine-learning algorithm capable of identifying the relevant degrees of freedom and executing RG steps iteratively without any prior knowledge about the system. We introduce an artificial neural network based on a model-independent, information-theoretic characterization of a real-space RG procedure, which performs this task. We apply the algorithm to classical statistical physics problems in one and two dimensions. We demonstrate RG flow and extract the Ising critical exponent. Our results demonstrate that machine-learning techniques can extract abstract physical concepts and consequently become an integral part of theory- and model-building.

Machine learning has been captivating public attention lately due to groundbreaking advances in automated translation, image and speech recognition[1], game-playing[2] and achieving super-human performance in tasks in which humans excelled while more traditional algorithmic approaches struggled[3]. The applications of those techniques in physics are very recent, initially leveraging the trademark prowess of machine learning in classification and pattern recognition and applying them to classify phases of matter[4–8], study amorphous materials[9,10], or exploiting the neural networks' potential as efficient nonlinear approximators of arbitrary functions[11,12] to introduce a new numerical simulation method for quantum systems[13,14]. However, the exciting possibility of employing machine learning not as a numerical simulator, or a hypothesis tester, but as an integral part of the physical reasoning process is still largely unexplored and, given the staggering pace of progress in the field of artificial intelligence, of fundamental importance and promise.

The renormalization group (RG) approach has been one of the conceptually most profound tools of theoretical physics since its inception. It underlies the seminal work on critical phenomena[15], and the discovery of asymptotic freedom in quantum chromodynamics[16], and of the Kosterlitz–Thouless phase transition[17,18]. The RG is not a monolith, but rather a conceptual framework comprising different techniques: real-space RG[19], functional RG[20] and density matrix RG[21], among others. While all of those schemes differ quite substantially in their details, style and applicability, there is an underlying physical intuition that encompasses all of them—the essence of RG lies in identifying the 'relevant' degrees of freedom and integrating out the 'irrelevant' ones iteratively, thereby arriving at a universal, low-energy effective theory. However potent the RG idea, those relevant degrees of freedom need to be identified first[22,23]. This is often a challenging conceptual step, particularly for strongly a Boltzmann distribution; no further knowledge about the microscopic details of the system is provided. The internal parameters of the network, which ultimately encode the degrees of freedom of interest at each step, are optimized ('learned', in neural network parlance) by a training algorithm based on evaluating real-space mutual information (RSMI) between spatially separated regions. We validate our approach by studying the Ising and dimer models of classical statistical physics in two dimensions. We obtain the RG flow and extract the Ising critical exponent. The robustness of the RSMI algorithm to physically irrelevant noise is demonstrated.

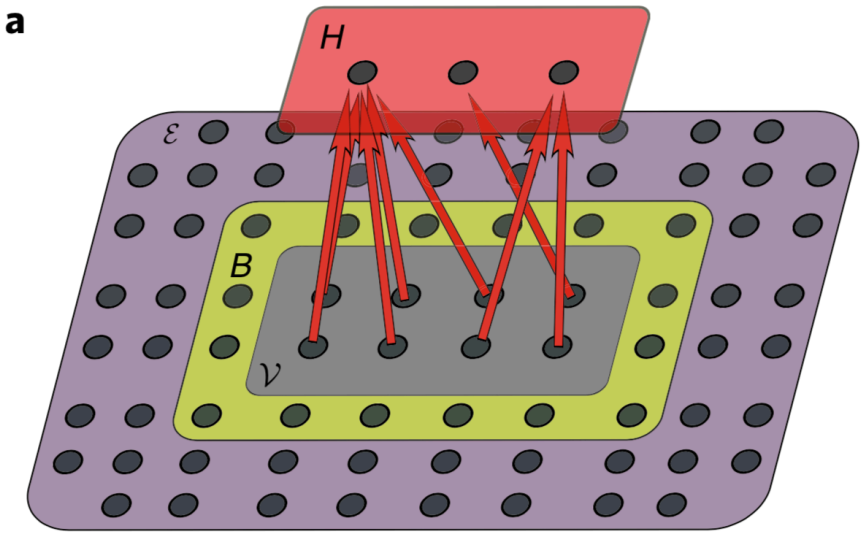
The identification of the important degrees of freedom, and the ability to execute a real-space RG procedure[19], has not only quantitative but also conceptual significance: it allows one to gain insights into the correct way of thinking about the problem at hand, raising the prospect that machine-learning techniques may augment the scientific inquiry in a fundamental fashion.

The RSMI algorithm

Before going into more detail, let us provide a bird's eye view of our method and results. We begin by phrasing the problem in probabilistic/information-theoretic terms, a language also used in refs [26–30]. To this end, we consider a small 'visible' spatial area $\mathcal{V}$, which together with its environment $\mathcal{E}$ forms the system $\mathcal{X}$, and we define a particular conditional probability distribution $P_\Lambda(\mathcal{H}|\mathcal{V})$, which describes how the relevant degrees of freedom $\mathcal{H}$ ('dubbed hiddens') in $\mathcal{V}$ depend on both $\mathcal{V}$ and $\mathcal{E}$. We then show that the sought-after conditional probability distribution is found by an algorithm maximizing an information-theoretic quantity, the mutual information, and that this algorithm lends itself to a natural implementation using artificial neural networks. We describe how RG is practically performed by coarse-graining with respect to $P_\Lambda(\mathcal{H}|\mathcal{V})$ and iterating

Real Space Mutual Information

- Algorithm:
  - Estimate data marginals using RBM's and contrastive divergence training
  - Use these data estimates to optimize another RBM between 'hidden' degrees of freedom (a function of some small subset of sites) and the environment (the remaining sites) through gradient descent



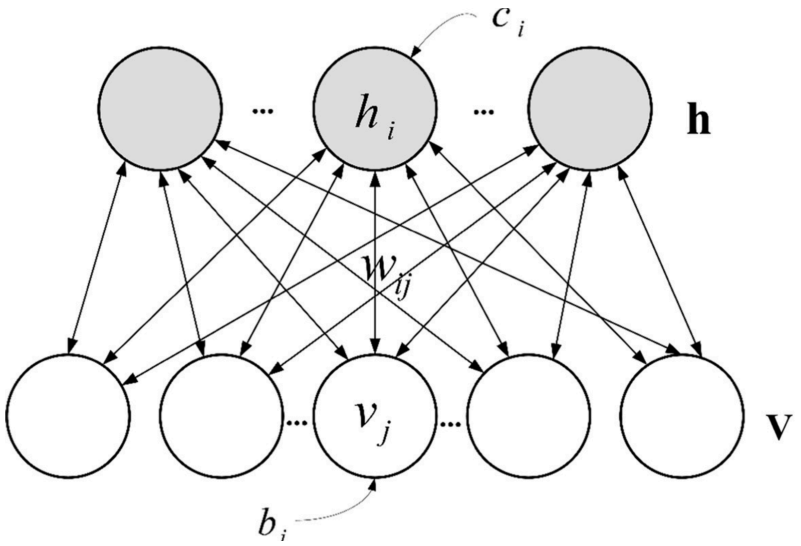
Restricted Boltzmann Machines

- Hamiltonian (Energy Function) given by:

$$H(\{v_i\},\{h_j\}) = \sum_j \beta_j h_j + \sum_{ij} v_i \lambda_{ij} h_j + \sum_j \alpha_i v_i$$

- Boltzmann distribution:

$$P(\{v_i\},\{h_j\}) = \frac{1}{Z} e^{-H(\{v_i\},\{h_j\})}$$



Contrastive Divergence

- Speed up sampling from Restricted Boltzmann Machines
- Want P(X), run a Markov Chain to convergence using Gibbs Sampling
- Gibbs Sampling in RBM's:
  - Fix value of hidden, sample marginal probability of a single spin conditioned on all other spins
  - Fix value of observed, sample marginal probability of hidden conditioned on other hidden
  - Repeat many times
- Contrastive Divergence:
  - Initialize Markov Chain close to target distribution, not randomly
  - Samples of P(X) not gathered after chain convergence, gathered after k steps of Gibbs sampling (often k=1)

Real Space Mutual Information

- Intuition behind I(H:E) as a measure of macro scale information:
  - Hidden variables H should contain information in V that overlaps with E
  - If E contained all the information in V, H would try to copy V
  - By restricting the support/ number of variables in H, can ensure that V can never be fully copied, introducing some potential information loss each renormalization step

Possible Issues

- Each renormalization step is computationally expensive (relies on many Monte Carlo samples to construct data distribution, and maximizes I(H:E) at every step for some partition)

- Depending on the information present in the subset V, a renormalization step may result in undesirable loss of information about macro scale

Possible Alternatives

- Shameless plug: Multivariate Mutual Information
  - Mutual Information: How much information is shared between two random variables?
  - Multivariate Mutual Information: How much information is shared between a group of random variables?

- Information among a group of random variables can be shared in different ways! (Redundantly, Synergistically)

X_1	X_2	X_3
0	0	0
1	1	1
1	1	1
0	0	0

A B	O
0 0	0
0 1	1
1 0	1
1 1	0

Redundancy as Macro Scale

- One interpretation of macro scale properties of a system corresponds to information in a system that is consistently repeated
- Extracting / preserving redundant information (if present) would have a similar effect as a large number of renormalization course graining steps
- There are measures of multivariate mutual information that are maximized for redundancy (total correlation)

$$TC(X) = \sum_{x \in X} p(x) \log \frac{p(x)}{\prod_i p(x_i)}$$
$$= \sum_i H(X_i) - H(X)$$
$$= KL[p(x) || \prod_i p(x_i)]$$

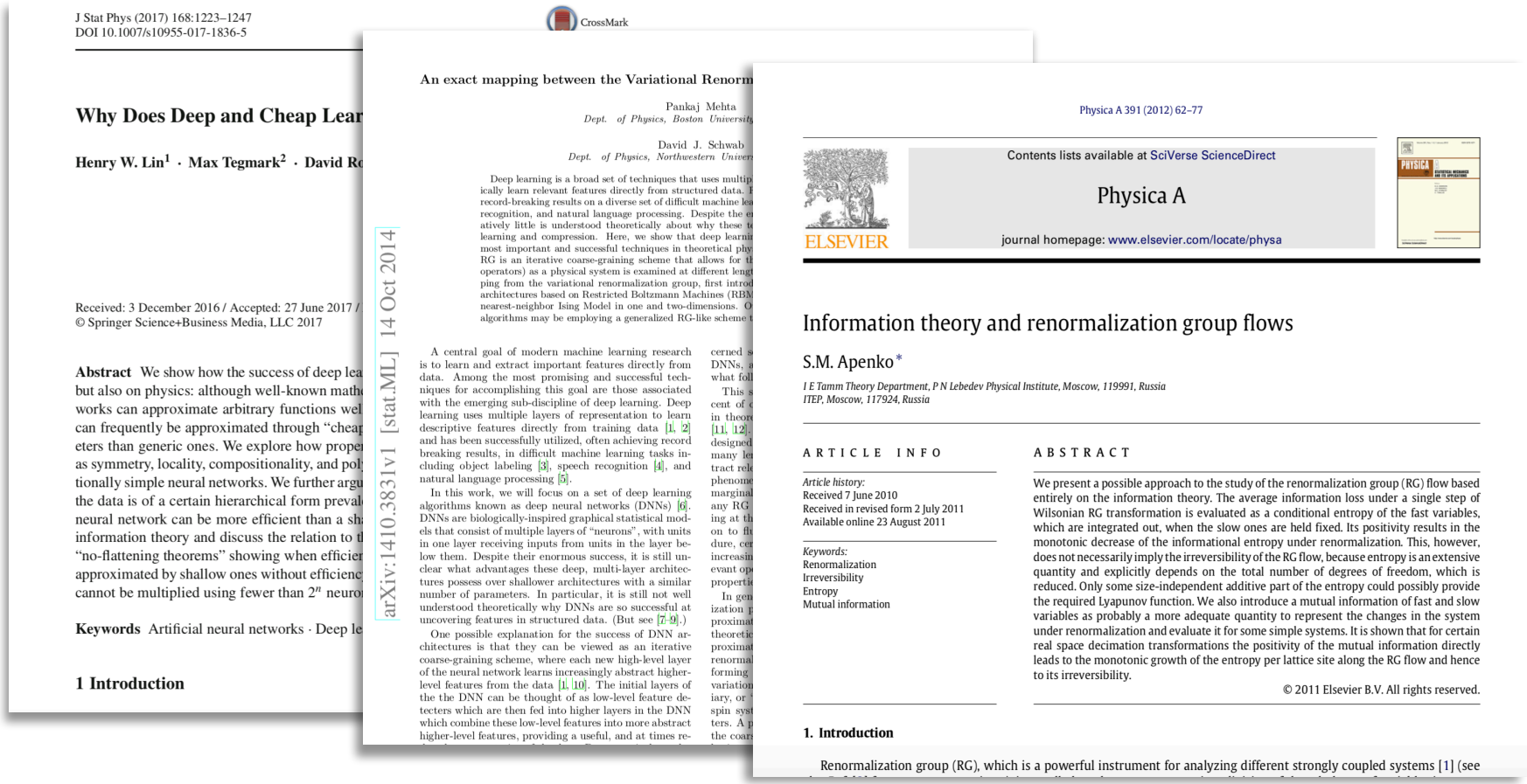
Redundancy as Macro Scale

- There are efficient algorithms (O(n) time !) for extracting redundancy in a similar fashion as Real Space Mutual Information
  - https://github.com/gregversteeg

- Next step: run it and see if it works!

Final Note

- Physics helps ML, ML helps Physics?



J Stat Phys (2017) 168:1223–1247
DOI 10.1007/s10955-017-1836-5

Why Does Deep and Cheap Lear

Henry W. Lin[1] · Max Tegmark[2] · David Rol

Received: 3 December 2016 / Accepted: 27 June 2017 /
© Springer Science+Business Media, LLC 2017

Abstract We show how the success of deep lea
but also on physics: although well-known mathe
works can approximate arbitrary functions wel
can frequently be approximated through "cheap
eters than generic ones. We explore how proper
as symmetry, locality, compositionality, and pol
tionally simple neural networks. We further argu
the data is of a certain hierarchical form preval
neural network can be more efficient than a sh
information theory and discuss the relation to t
"no-flattening theorems" showing when efficien
approximated by shallow ones without efficienc
cannot be multiplied using fewer than 2^n neuro

Keywords Artificial neural networks · Deep le

1 Introduction

CrossMark

An exact mapping between the Variational Renorm

Pankaj Mehta
Dept. of Physics, Boston University

David J. Schwab
Dept. of Physics, Northwestern Univer

Deep learning is a broad set of techniques that uses multip
ically learn relevant features directly from structured data.
record-breaking results on a diverse set of difficult machine lea
recognition, and natural language processing. Despite the e
atively little is understood theoretically about why these te
learning and compression. Here, we show that deep learni
most important and successful techniques in theoretical phy
RG is an iterative coarse-graining scheme that allows for th
operators) as a physical system is examined at different lengt
ping from the variational renormalization group, first introd
architectures based on Restricted Boltzmann Machines (RBM
nearest-neighbor Ising Model in one and two-dimensions. O
algorithms may be employing a generalized RG-like scheme t

A central goal of modern machine learning research is to learn and extract important features directly from data. Among the most promising and successful techniques for accomplishing this goal are those associated with the emerging sub-discipline of deep learning. Deep learning uses multiple layers of representation to learn descriptive features directly from training data [1, 2] and has been successfully utilized, often achieving record breaking results, in difficult machine learning tasks including object labeling [3], speech recognition [4], and natural language processing [5].

In this work, we will focus on a set of deep learning algorithms known as deep neural networks (DNNs) [6]. DNNs are biologically-inspired graphical statistical models that consist of multiple layers of "neurons", with units in one layer receiving inputs from units in the layer below them. Despite their enormous success, it is still unclear what advantages these deep, multi-layer architectures possess over shallower architectures with a similar number of parameters. In particular, it is still not well understood theoretically why DNNs are so successful at uncovering features in structured data. (But see [7–9].)

One possible explanation for the success of DNN architectures is that they can be viewed as an iterative coarse-graining scheme, where each new high-level layer of the neural network learns increasingly abstract higher-level features from the data [1, 10]. The initial layers of the the DNN can be thought of as low-level feature detecters which are then fed into higher layers in the DNN which combine these low-level features into more abstract higher-level features, providing a useful, and at times re-

cerned s
DNNs,
what foll
This s
cent of
in theor
[11, 12].
designed
many le
tract rele
phenome
marginal
any RG
ing at th
on to fi
dure, cer
increasin
evant op
properti
In gen
ization p
proximat
theoretic
proximat
renormal
forming
variation
iary, or
spin syst
ters. A p
the coars

arXiv:1410.3831v1 [stat.ML] 14 Oct 2014

Physica A 391 (2012) 62–77

Contents lists available at SciVerse ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

ELSEVIER

Information theory and renormalization group flows

S.M. Apenko*

I E Tamm Theory Department, P N Lebedev Physical Institute, Moscow, 119991, Russia
ITEP, Moscow, 117924, Russia

ARTICLE INFO

Article history:
Received 7 June 2010
Received in revised form 2 July 2011
Available online 23 August 2011

Keywords:
Renormalization
Irreversibility
Entropy
Mutual information

ABSTRACT

We present a possible approach to the study of the renormalization group (RG) flow based entirely on the information theory. The average information loss under a single step of Wilsonian RG transformation is evaluated as a conditional entropy of the fast variables, which are integrated out, when the slow ones are held fixed. Its positivity results in the monotonic decrease of the informational entropy under renormalization. This, however, does not necessarily imply the irreversibility of the RG flow, because entropy is an extensive quantity and explicitly depends on the total number of degrees of freedom, which is reduced. Only some size-independent additive part of the entropy could possibly provide the required Lyapunov function. We also introduce a mutual information of fast and slow variables as probably a more adequate quantity to represent the changes in the system under renormalization and evaluate it for some simple systems. It is shown that for certain real space decimation transformations the positivity of the mutual information directly leads to the monotonic growth of the entropy per lattice site along the RG flow and hence to its irreversibility.

1. Introduction

Renormalization group (RG), which is a powerful instrument for analyzing different strongly coupled systems [1] (see