Renormalization Group and Information Theory

Kyle Reing
University of Southern California

April 18, 2018
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 coarse graining operation R, data X, and measure of ‘macro scale properties’ Y, want:

\[ R^n = R \circ R \circ R \circ R \circ \ldots \]

\[ I(Y : X) = I(Y : R^n \circ X) \]
Super Quick Info Theory Definitions

• (Shannon) Entropy
  • Measure of uncertainty in a random variable, how ‘spread out’ is the distribution?

\[
H(X) = - \sum_{x \in X} p(x) \log p(x) \quad \text{Discrete}
\]

\[
= - \int_{x \in X} p(x) \log p(x) \quad \text{Continuous}
\]

• 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?
Mutual information, neural networks and the renormalization group

Maciej Koch-Janusz and Zohar Ringle

Physical systems differing in their microscopic details often display strikingly similar behaviour when probed at macroscopic scales. These 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, game playing and achieving super-human performance in tasks in which humans excelled while more traditional algorithmic approaches struggled. The applications of these 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[1], study amorphous materials[2,3], or exploiting the neural networks’ potential as efficient nonlinear approximators of arbitrary functions[4,5] to introduce a new numerical simulation method for quantum systems[6,7]. 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, and the discovery of asymptotic freedom in quantum chromodynamics, and of the Kosterlitz–Thouless phase transition. The RG is not a monolith, but rather a conceptual framework comprising different techniques: real-space RG, functional RG and density matrix RG, 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. This is often a challenging conceptual step, particularly for strongly Boltzmann distributions. 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, 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[8-10]. To this end, we consider a small ‘visible’ spatial area $V$, which together with its environment $\mathcal{E}$ forms the system $X$, and we define a particular conditional probability distribution $P_{\mathcal{E}}(\mathcal{H}|V)$, which describes how the relevant degrees of freedom $\mathcal{H}$ (labeled hidden) in $V$ 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 $P_{\mathcal{E}}(\mathcal{H}|V)$ 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)

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
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](https://github.com/gregversteeg)

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

• Physics helps ML, ML helps Physics?