Pathwise Information Theory in Two Dimensions

Jordan Snyder
Department of Mathematics, UC Davis
jasnyder@math.ucdavis.edu

12 June 2014

Abstract

We consider an approach to extending Information Theory to spatially extended systems based on computing information quantities along collections of self-avoiding walks on the underlying lattice. Several test cases are analyzed and techniques are proposed for extracting geometric information from pathwise information quantities. Finally, we consider some data sets of physical interest (2D Ising model) and discuss possible generalizations.

1 Introduction

Information Theory (IT), catalyzed by Claude Shannon's seminal 1948 paper, *A Mathematical Theory of Communication*, has seen enormous success over the years, and has given rise to the framework of ϵ-machines - minimal optimally predictive models of dynamical systems, constructed by partitioning the space of system histories into those which are causally distinct, obtaining a set of causal states.

One major drawback, however, is that classically, IT is developed in the setting of a one-dimensional data sequence. In applications this one dimension may be the time coordinate of some measurement of the system, or the spatial coordinate of a one-dimensional system. If we consider a system posessing more than on spatial dimension (as almost all physically interesting systems do), and wish to construct an ϵ-machine reproducing its dynamics, we are then forced to consider a time-series of measurements on the entire system. At best, we may expect to glean a temporal model (perhaps with very large alphabet size) that reproduces the dynamics of the system. However, such a model would have its spatial information lumped into "causal states", and we would then be faced with the daunting task of seeking spatial commonalities among the equivalent histories making up each causal state.

Hence we seek methods for applying information theory to spatially extended systems that retain detailed geometric information about how spatially separated portions of a system communicate with each other. As IT is already very well understood for one-dimensional sequences of data, we propose to bootstrap our way into higher dimensions by applying IT to a large collection of one-dimensional subsets of the system of interest. For conceptual clarity and ease of computation, we consider 2D systems living on a square lattice, and we consider self-avoiding walks on the square lattice as our family of candidate one-dimensional subsets.

2 Background

The setting for the most richly developed portion of information theory is a one-dimensional sequence of data. We consider some alphabet $\mathcal{A}$ and a bi-infinite sequence of random variables $\{X_t\}_{t\in\mathbb{Z}} \in \mathcal{A}^{\mathbb{Z}}$. Given such a

sequence we can compute such quantities as:

- Block Entropy $H(L)$

$$H(L) = \sum_{w^L} \Pr(w^L) \log_2(\Pr(w^L)) \tag{1}$$

  where $w^L \in \mathcal{A}^L$ denotes a word of length L.

- Entropy rate h_μ and myopic entropy rates $h_\mu(L)$

$$h_\mu = \lim_{L \to \infty} h_\mu(L) \tag{2}$$

$$h_\mu(L) = H(L+1) - H(L) \tag{3}$$

- Excess Entropy

$$E = I(X_{:0}; X_{0:}) \tag{4}$$

  where $I(\cdot;\cdot)$ denotes the mutual information between two probability distributions, $X_{:0}$ denotes the semi-infinite past, and $X_{0:}$ denotes the semi-infinite future. Excess entropy is, in heuristic terms, the amount of information that the past shares with the future.

In each of the above formulae, the essential one-dimensional nature of the underlying data ($\{X_t\}$) is manifest. As one might expect, there have been attempts to generalize these quantities to higher-dimensional data. For example, Lempel and Ziv, in their 1986 paper *Compression of Two-Dimensional Data* [2], consider a compression scheme based on a finite-state encoder that visits each site of a 2D lattice along a space-filling path, and define the *asymptotic compressibility* of an image, a quantity which shares the essential properties of a generalization of entropy rate (perhaps better named "entropy density"). However, this approach is limited in that its focus is simply reproduction of the original data set, rather than giving a causal description of the system's spatial structure.

Another approach developed by Feldman and Crutchfield in their 2002 paper *Structural Information in Two-Dimensional Patterns: Entropy Convergence and Excess Entropy*[3], considers a two-dimensional template to replace the one-dimensional "word" in the definitions above, and scales this template linearly to recreate the $L \to \infty$ limit in one dimension. The authors then show consistent results for this technique when applied to the Ising model under nearest-neighbor and next-nearest-neighbor interactions. While this approach is appealing, its main drawback is a lack of generality - optimality for the specific template considered is only shown for the Ising model, and to apply the same shape to other systems would be to include spurious assumptions about their spatial organization.

3 System

As stated, we consider systems on a two dimensional square lattice, in which each lattice site is assigned a state of either +1 or -1. A representative configuration, in which each site's state was chosen independently with equal probability, is shown in Figure 1.

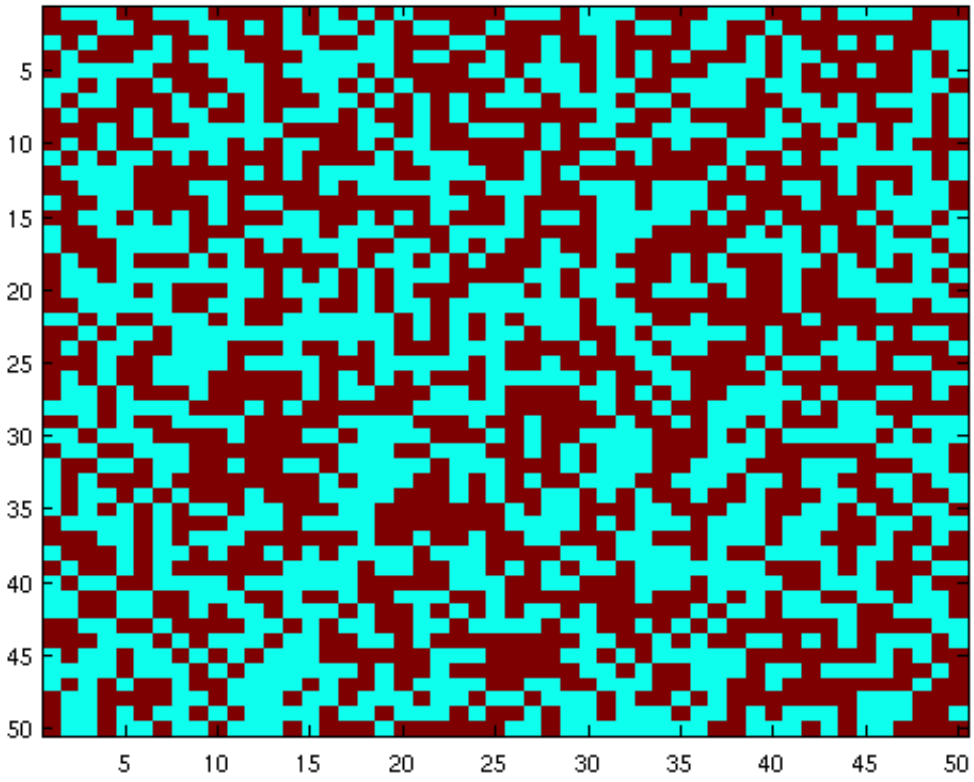

Figure 1: Random Configuration

Next we consider the set of candidate paths along which we will perform IT. I have chosen to consider the set of all self-avoiding walks beginning at a fixed base point. The reason I have included the self-avoiding assumption is that I will be looking at correlations along each path, and I don't want to introduce the possibility of spurious correlation. The reason I have included *only* this assumption is that I want to remain as agnostic as possible about the geometry of the system's communication with itself.

A further assumption is more implicit, and arises from the way in which I perfrom IT along paths. For a given path, I wish to find a *probability distribution* over words read along that path. For computational convenience, I treat *each site* in the lattice as a representative sample from a probability distribution, which is the one that I aim to study - this can be stated as an assumption of homogeneity. Hence, I perform computations over a *single configuration*, by reading a word along the given path for each possible choice of base point.

The configurations to which I have applied my methods are mostly contrived test cases to illuminate the meaning of my measurements. Additionally, I have considered some data sets that are meant to be representative of a configuration of a 2D spin system at the critical temperature under the Ising model.

4 Methods

I will now make explicit the processes I have outlined above.

4.1 Enumerating Walks

I have encoded self-avoiding walks as strings of the form, for instance, $'uSLR'$, where the first letter, u, d, l, r represents stepping first up, down, left, or right from the base point, respectively. The subsequent capital letters L, S, R represent turning left, going straight, or turning right, respectively. This enables me to recursively enumerate the paths of length n by choosing each path of length $n-1$ and appending each of L, S, and R. Once all the *potential* paths are enumerated, I prune the collection by imposing the requirement that the paths do not visit the same site twice.

I enumerated self-avoiding walks up to those that visit 9 lattice sites, of which there are a whopping 5916. I checked my work by checking that I had listed the appropriate number of walks of each length - these figures are known, and listed in several sources, e.g. [1]

4.2 Computing Information Quantities

The first step in doing information theory along a path is collecting a word distribution. First, we fix a path p and a configuration C, presented as a matrix. Next, I list out the coordinates of every site the path visits, in terms of the displacements from the base point. For example, the visited sites of the path $p = uSLR$ are $VS = \{(0,0), (-1,0), (-2,0), (-2,-1), (-3,-1)\}$, using the the matrix indexing notation native to MATLAB, in which the first coordinate increases downward and the second coordinate increases rightward.

Next, for each site $s \in C$, I read a word $w = \{C(s+VS(i))\}_{i=1}^{n}$, where C is considered with periodic boundary conditions. As I read these words, I build a list of distinct words, and how many times each of them is read. Once I visit every site, I normalize the word counts so that they comprise a probability distribution, and compute the Shannon entropy of this distribution, which I denote $H(p)$.

Having tabulated a word distribution for the path p, I can now define a myopic entropy rate $h_\mu(p)$ and a myopic excess entropy $E(p)$, by very mild modifications of formulae (3) and (4). First, I denote by p' the path obtained from p by removing the last step, and define

$$h_\mu(p) = H(p) - H(p') \tag{5}$$

Next, I define sub-paths $p_1, p_2 \subset p$ to be the first and second halves of the path p, respectively. When p visits an odd number of sites, I assign one more site to p_1 than to p_2. I then define myopic excess entropy by the formula

$$E(p) = I(p_1; p_2) = H(p_1, p_2) - H(p_1) - H(p_2) \tag{6}$$

Note that in the above formula, $H(p_1, p_2) = H(p)$, so all that remains to compute are the marginal distributions $H(p_1)$ and $H(p_2)$, which can be straightforwardly obtained from the total word distribution. I describe examples that illuminate the physical meaning of $E(p)$ in section 5.1.1 below.

Finally, I define what has been my primary object of interest, the *mutual information spectrum*. This is simply the collection of $S(C, L) = \{E(p^L) | p^L \text{ a path of length } L\}$ for a given configuration C.

4.3 Unpacking Outputs

Given a mutual information spectrum $S(C, L)$, I wish to extract meanignful geometric information about how the system represented by C is organized. The basic strategy is to quantitatively investigate the relationship between $E(p)$ and the geometric features of the path p.

One way to do this is consider the connectedness of the path, denoted $c(p)$, which I define to be the number of pairs of sites along the path that are direct neighbors to each other (i.e., they differ by a unit vector $(\pm 1, 0)$ or $(0, \pm 1)$). Once this is computed, I can plot $E(p)$ against $c(p)$ to discern any correlation. A positive correlation between $E(p)$ and $c(p)$ can be taken to indicate that neighboring lattice site pairs influence each other more highly than non-neighboring pairs. An example of such an occurrence is given in the Results section.

Another plan of attack is to consider operations transforming one path into another, and measuring the effect of such a transformation on $E(p)$. One natural choice for a set of transformations is that induced by the action of the dihedral group D_4 on the square lattice - these transformations are simply reflections about each axis and each diagonal, and rotation by any multiple of 90 degrees. To do this, I computed a multiplication table, indicating the action of each element $f \in D_4$ on the set P of all paths, and denote this action by $f.p$. I can then perform a number of different computations, given a spectrum $S(C, L)$. For example, given an element $f \in D_4$, I define

$$\delta(f) := \langle |E(p) - E(f.p)| \rangle_P \tag{7}$$

Where $\langle \cdot \rangle_P$ denotes an expectation value taken over all $p \in P$. This quantity is simply the average of how much action on p by f changes $E(p)$. We may interpret $\delta(f) = 0$ as saying that information propagation within C is *invariant under* f.

This notion can be generalized to subgroups $G \leq D_4$ in the following way: I partition the paths P into orbits under action by G, the collection of which is denoted P/G. An orbit of a path p under G is defined as

$$G.p = \{g.p | g \in G\} \subseteq P$$

That orbit memebership is an equivalence relation, and hence partitions paths, is a standard fact from group theory. I can then consider, for each orbit $G.p$, the mutual information $E(p)$ read along each path $p' \in G.p$, and define the quantity

$$\sigma(G.p) = \left\langle (E(p) - \langle E \rangle_{G.p})^2 \right\rangle_{G.p}^{\frac{1}{2}} \tag{8}$$

where $\langle \cdot \rangle_{G.p}$ denotes an average over the orbit $G.p$. This is simply the standard deviation of $E(p)$ within an orbit $G.p$. Intuitively, this quantity captures the size of the effect of action by G on information $E(p)$ for a given starting path p. Finally, we can take an average of $\sigma(G.p)$ over orbits, defining

$$\Delta(G) = \langle \sigma(G.p) \rangle_{P/G}$$

This quantity represents the effect of action by G on information $E(p)$ averaged over all possible starting paths p. We may interpret $\Delta(G) = 0$ as saying that information propagates within the system in a way posessing *exactly* the symmetry of G. Importantly, $\Delta(G)$ gives us not only a yes-or-no answer to the question, "is information flow in this system symmetric under G?", but a number that can quantify *how* (a-)symmetric information flow is in a particular system. For instance, given two subgroups $H, G \leq D_4$, we can interpret $0 < \Delta(G) < \Delta(H)$ as saying that information flow is asymmetric under *both* G and H, but is closer to symmetric under G.

Possible generalizations of this procedure include computations similar to Δ, but for actions not obtained from subgroups of D_4. For instance, we might wish to consider the transformation that maps a path to one with a "kink" introduced at some point.

5 Results

The majority of the data I have collected are from relatively simple test cases, designed to illuminate the meaning of the various calculations I have performed and to ensure that all of my code works as I think it does. As mentioned previously, I also considered some data sets generated by an algorithm which simulates a 2D spin system at the critical temperature under the Ising model.

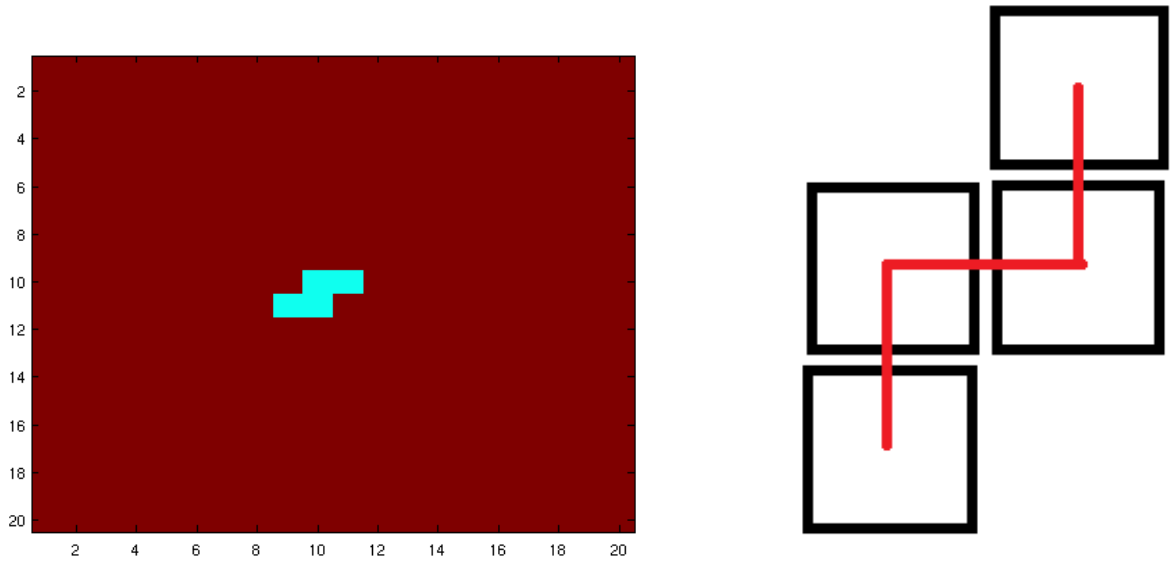

Figure 2: Notch Configuration and Optimal Path

5.1 Test Cases

5.1.1 Interpreting $E(p)$

The example I have used to gain insight into the physical significance of $E(p)$ is shown in the left of figure 2. It is simply a uniform background with a single notch shape in the middle, three sites wide and two sites high. I have computed $E(p)$ for each path that visits four sites, and found that $E(p)$ was largest for paths $'uRL'$ and $'dRL'$, both of which have the shape shown in the right of figure 2. I interpret this to mean that when evaluated over a given configuration, $E(p)$ is maximized along paths that frequently *cross* domain boundaries. With this in mind, we proceed to further examples.

5.1.2 Checkerboards

First, I consider a checkerboard configuration with a period of 1 in both directions (figure 3). This pattern is entirely predictable - given a particular path shape, reading a single symbol is enough to know exactly what symbols will be read along the rest of the path. Hence, we expect that $E(p)$ with be precisely 1 for all paths, and indeed, this is what we find

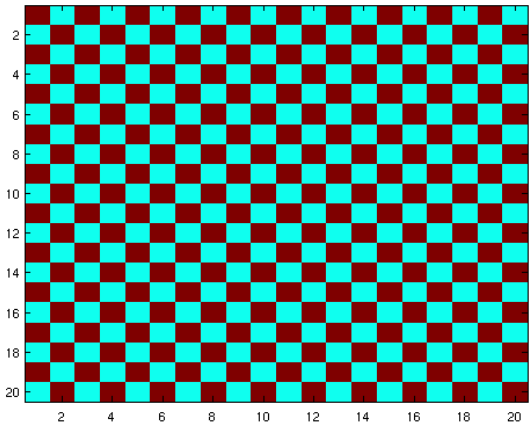

Figure 3: 1x1 Checkerboard

Next, we consider checkerboard patterns with larger periods. Figure 4 displays mutual information spectra for period-2 and period-3 square checkerboard patterns, while figure 5 shows the mutual information spectrum for a period-10 square checkerboard. In each case, the spectrum has been sorted in order of increasing $E(p)$. The spectra for periods 2 and 3 have been computed for all paths visiting 4 sites (of which there are 36), while the spectrum for period 10 has been computed for paths visiting 8 sites (of which there are 2172)

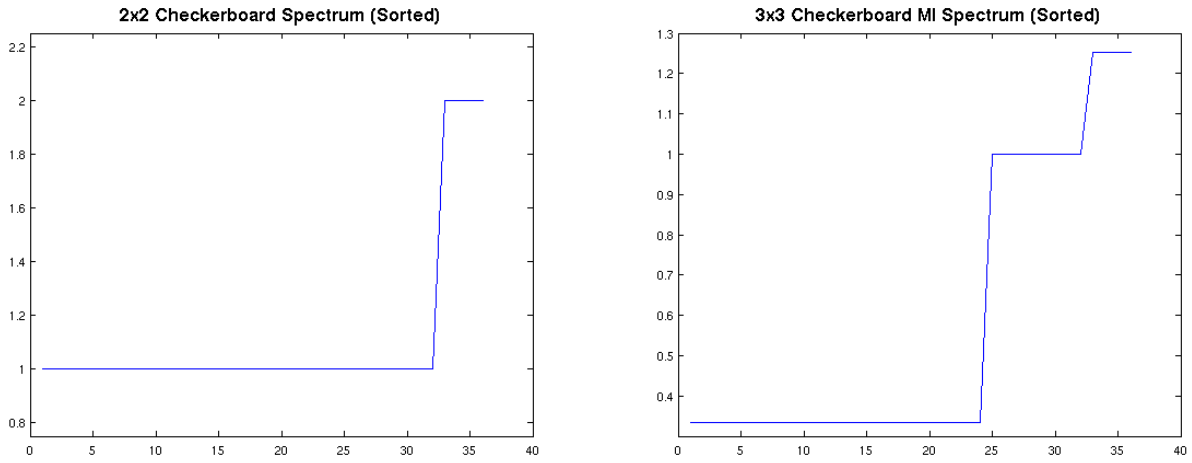

Figure 4: 2x2 and 3x3 Checkerboard Spectra

We observe readily that even for very simple patterns, such as a checkerboard, the mutual information spectrum can posess a very subtle structure.

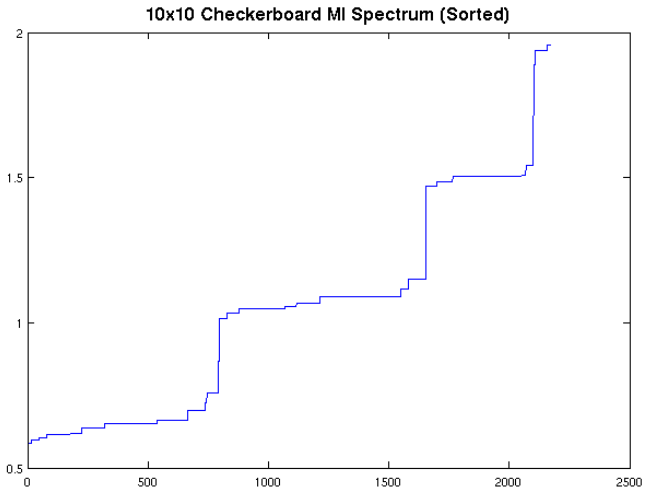

Figure 5: 10x10 Checkerboard Spectrum

5.1.3 Probing Symmetries

Several test cases I have considered are designed to test my ideas about quantifying symmetry. First, there is the slash, which I have considered on its own as well as with the states of 30 randomly chosen sites flipped, to investigate the robustness of my information measures. These are shown in figure 6.

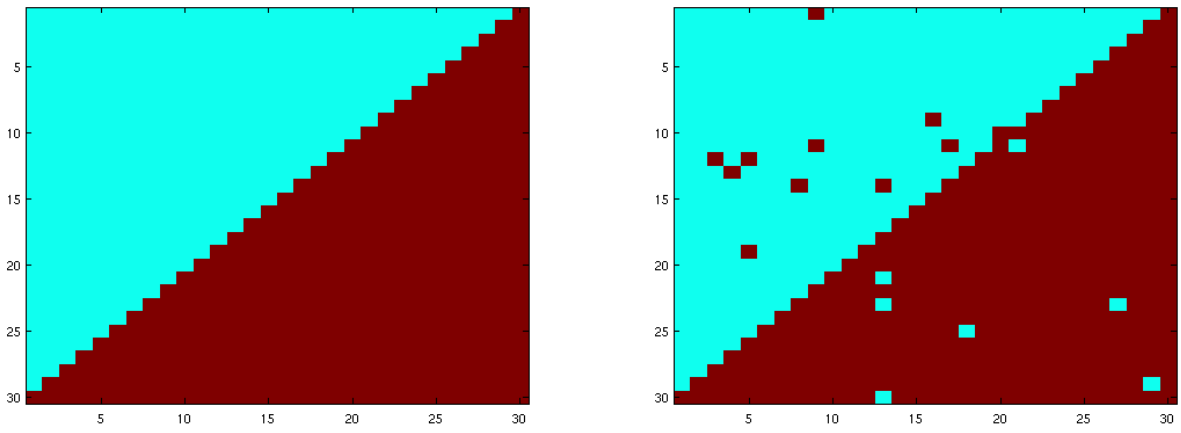

Figure 6: Slash and Noisy Slash

I considered this shape because it is invariant under a *proper* subgroup of D_4. That is, it is (up to sign flip) invariant under reflection about each diagonal, as well as rotation by 180 degrees, but not under action by any

other element of D_4. My goal was to see this fact arise by computing $\Delta(G)$ for the various subgroups of D_4. Figure 7 shows a bar chart of $\Delta(G)$ for G being the subgroup generated by each of the seven nontrivial elements of D_4. They are, respectively, reflection across $x = y$, reflection across $x = -y$, reflection across the y-axis, reflection across the x-axis, and rotation by 90, 270, and 180 degrees. Note that x and y again refer to matrix indices as used in MATLAB - x increases downward and y increases rightward.

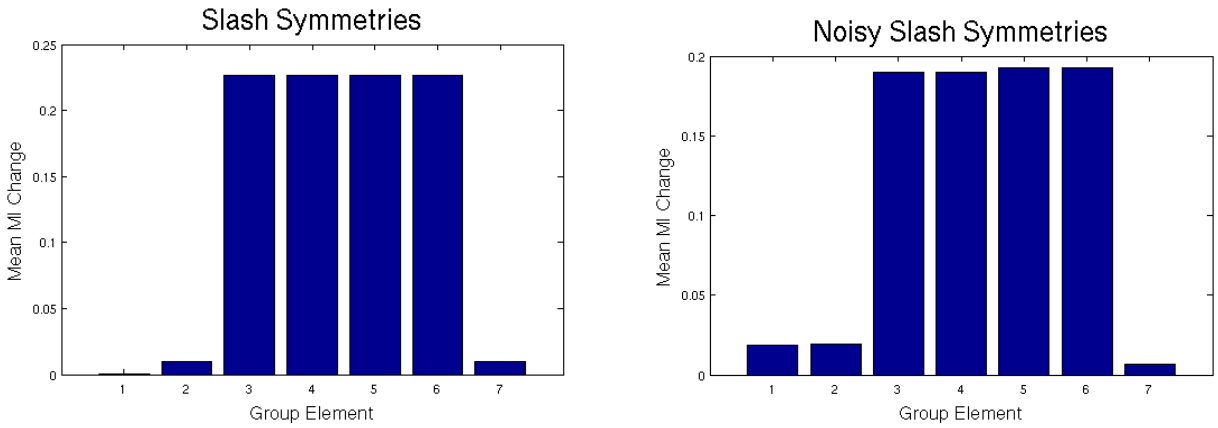

Figure 7: $\Delta(G)$ for cyclic subgroups of D_4

We can see immediately that pathwise mutual information is indeed (nearly) invariant under each diagonal reflection, and under 180 degree rotation, while not so under action by the other elements of D_4. Moreover, we can observe the *same* qualitative features in $\Delta(G)$ for the noisy slash pattern, but with larger variation under the actions that leave the noiseless shape invariant, and smaller variation under the actions that do not leave the noiseless shape invariant. This demonstrates the possiblity of using quantities such as $\Delta(G)$ to quantify the (a-)symmetry of two-dimensional patterns.

5.1.4 (Potentially) Physically Interesting Data

Finally, I will describe some of the data to which I have applied my techniques that are of some physical interest. My initial inspiration for this project was to create new tools to investigate phenomena in spatially extended, interacting systems. A very important example of this is, of course, the ferromagnetic phase transition in the 2D Ising model.

To this end I have implemented two versions of the Wolff cluster algoritm, which produces configurations representative of the Ising model near the critical temperature. The reason I have two versions is because the algorithm requires a definition of what it means for two sites to be "neighbors". In one version, I used the Moore neighborhood, which, on a 2D square lattice, includes all eight sites which share at least a vertex with a given site. In the other version, I have used the von Neumann neighborhood, which includes only those four sites which share an edge with a given site. Representative configurations from each simulation are shown in Figure 8.

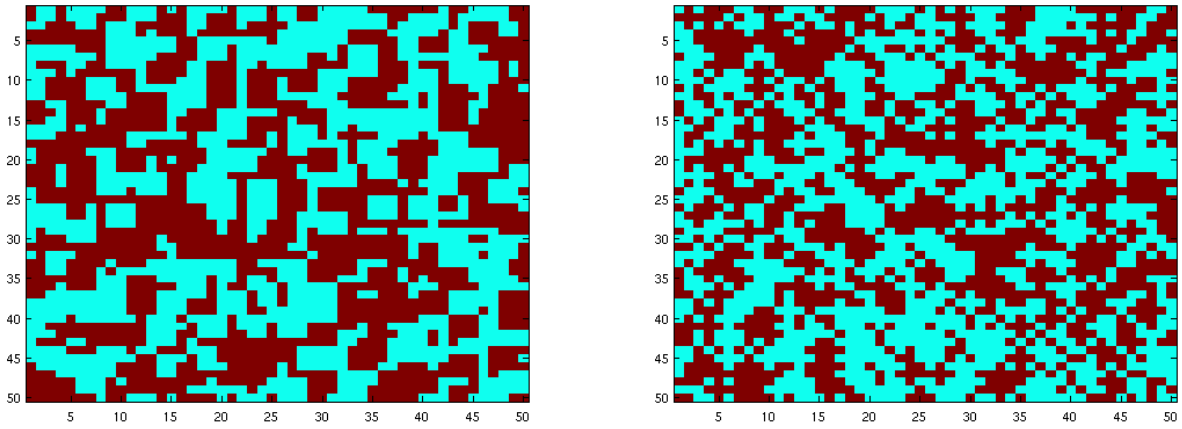

Figure 8: Ising Model Configurations using Moore and von Neumann neighborhoods (resp.)

For each of these configurations, I have computed the mutual information spectrum (figure 9).

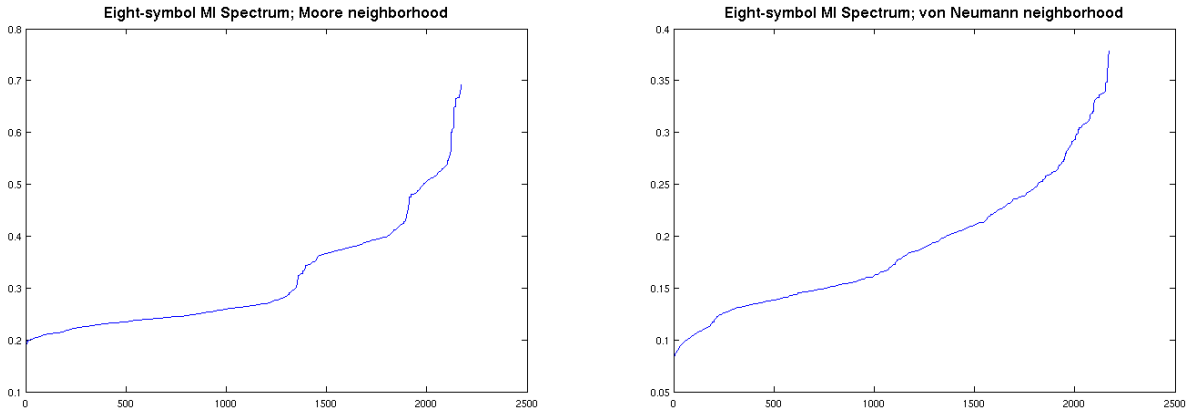

Figure 9: Ising Model Data Mutual Information Spectra

While we can tell from the spectra that there are some informational differences between Ising model data generated using Moore and von Neumann neighborhoods, it is supremely unclear what they are.

To probe somewhat further, I have plotted $E(p)$ vs $c(p)$ as described in section 4.3. The results are shown in figure 10.

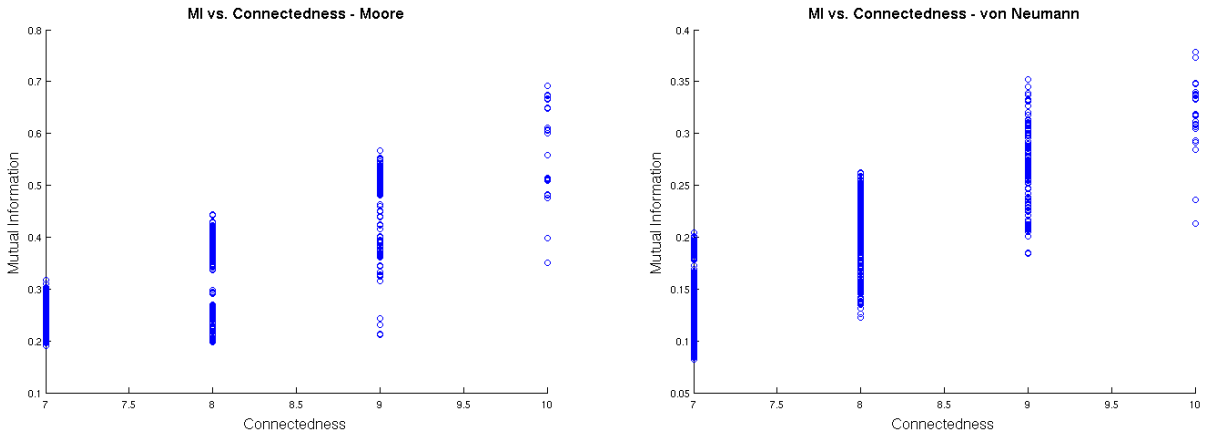

Figure 10: Mutual Information vs. Connectedness for Ising model data

One feature to notice is that correlation between $E(p)$ and $c(p)$ appears tighter under the von Neumann neighborhood than the Moore neighborhood. One interpretation of this is that when using the Moore neighborhood, correlations are present between sites that, when visited by a given path, do not contribute to connectedness (i.e. those that share only a vertex). This observation provides hope that these methods may be able to extract, from a configuration generated by some local dynamic, the geometric properties of that local dynamic.

6 Conclusion and Potential Future Work

We have developed a computational framework for extending information theory to systems posessing more than one dimension by exhaustively sampling along paths. We have seen that these methods can detect and quantify the symmetry of a two-dimensional configuration, and can provide hints at the nature of a local dynamic used to generate a randomly-seeded configuration.

Several assumptions have been built into the preceeding development. Chief among them is the assumption of homogeneity underlying the validity of sampling words across a single configuration. Potential applications of IT to $\geq$ 2D systems include partitioning space into domains of statistical homogeneity and, for example, studying their boundaries. One possible plan of attack for using pathwise IT for this application is parsing a configuration into many chunks, computing mutual information spectra by sampling across each chunk, and comparing spectra between chunks. This would give a potentially instructive way to measure variations across large regions of information flow in a system.

Other, more straighforward generalizations of this work exist. State alphabets other than $\{\pm 1\}$ require no modification of my existing code. A less trivial, but still straightforward generalization is to lattices in more than two dimensions. The main difficulty in this case will be the growth in the number of self-avoiding walks to consider. Asymptotically, the number of self-avoiding walks on a (hyper-)cubic lattice of dimension d goes as $(\mu_d)^n$, where the so-called "connective constant" μ_d has values of approximately $2.62, 4.57, 6.74, 8.83$, and 10.87 for dimensions $2, 3, 4, 5$, and 6 respectively. Another subtlety introduced when considering a higher-dimensional lattice is that the symmetry group that acts most naturally on paths is much larger than D_4. For instance, the corresponding group for a 3D cubic lattice is the octahedral group, which has order 48, in contrast to D_4's order of 8.

References

[1] A. R. Conway et al., Algebraic techniques for enumerating self-avoiding walks ..., J. Phys A 26 (1993) 1519-1534.

[2] Lempel, Abraham, and Jacob Ziv. "Compression of two-dimensional data." Information Theory, IEEE Transactions on 32.1 (1986): 2-8.

[3] Feldman, David P., and James P. Crutchfield. "Structural information in two-dimensional patterns: Entropy convergence and excess entropy." Physical Review E 67.5 (2003): 051104.