A Few Experiments in 2-D Information Theory

Ander Aguirre
Department of Mathematics, UC Davis
aaguirre@ucdavis.edu

14 June 2018

Abstract

We consider the problem of generalizing information theoretic measures to finite lattices in 2 dimensions. In order to do so, we first generate 1-dimensional paths along these and compute well known quantities such as mutual information. Additionally, we study the relevance of the scale in which we choose to perform such measurements.

Introduction

Since the advent of Information Theory, the celebrated formula $H = -\sum p \log p$ has given rise to several related measures that help characterize discrete time stochastic processes.In this setting, we can develop quite an accurate picture of the underlying structure in the data generated by a stochastic process $\{X_n\}_{n\in\mathbb{Z}}$ emitting symbols in an alphabet $\mathcal{A}$.

A natural question is then to extend this framework to data on lattices of higher dimensions, where, for topological reasons, we can observe more interesting structures. Unfortunately, none of the tools in the toolkit of classical Information Theory is equipped to tackle the problem even in 2 dimensions. An obvious way to circumvent this problem, is thus to apply these well known measures to 1-dimensional subsets of the data we are studying. However, several difficulties arise in this approach as well. The main one is that there is no canonical way to decide how to enumerate the steps of a path in 2-dimensions. For instance, we may consider the approach of Lempel and Ziv in *Compression of two dimensional data* of scanning two dimensional data by a particular space filling curve. In this process the symbols of some neighboring cells will inevitably have to appear far apart once the data is translated into a one dimensional string and much of the geometric structure will be ignored . A more logistical problem is the sheer amount of paths we can choose. For the sake of impartiality, we wish to sample as many as possible paths. For starters, we can impose the condition that they are self avoiding, as otherwise we will obtain misleading correlations. We could also discard walks (sequences) outside the typical set. At any rate, however, the number of self avoiding walks (c_n) on the lattice grows with length n as:

$$c_n \approx \mu^n n^{\gamma-1}$$

The constant μ is called *connective constant* and depends on the particular lattice; γ, on the other hand, is universal, meaning it only depending on the dimension.

In short, it is abundantly clear that Information Theory in 2-D is highly nontrivial. The many efforts to extend classical Information Theory to 2 dimensions have focused on generalizing certain specific notions, suggesting that the definitions in 2-D may never be as natural as in the 1-D setting. Likewise, in this project we will restrict our attention to considerations of periodicity and scale of measurements. Regardless, of how mathematically elegant the theory may turn out at the end, it is certainly true that many real world applications would greatly benefit from its further development.

Background

For our study of 1 dimensional data, the framework we will be working on is that of a discrete time stochastic process $\{X_n\}_{n\in\mathbb{Z}}$ emitting symbols from a finite alphabet $\mathcal{A}$. We will restrict our attention to the binary alphabet $\{1,-1\}$, from which we will generate binary words of length L (w^L) for each 1 dimensional path. The information measures we will be applying will be the following:

1. **Block Entropy H(L)**: Its formal definition is:

$$H(L) = -\sum_{s^L\in\mathcal{A}^L} \mathbb{P}(s^L)\log_2 \mathbb{P}(S^L)$$

   For this project,the sum in this formula is understood to range over all the binary words of length L. We generate these words via a path on $\mathbb{Z}^2$ with binary symbols at each site. Generally speaking, longer words will have higher block entropies so it is also of interest to look at the value $H(L)/L$. Recall that as $L\to\infty$, if the limit exists, this quantity will correspond to entropy rate, which we can interpret intuitively as the time density of information generated by a stochastic process[1].

2. **Excess Entropy** $E = I(X_{\leftarrow}; X_{\rightarrow})$: Formally it is defined as:

$$E = \lim_{L\to\infty} I[S_0, S_1..S_{L-1}; S_L, S_{L+1}...S_{2L-1}]$$

   We can think of this as the mutual information between the past and the future. For our purposes, it will be the mutual information between the two halves of the path we are studying. This measure is of particular interest because, another way it can be interpreted is as intrinsic redundancy [1]. More specifically, what this tells us is that for small L the entropy rate h_μ is overestimated by the approximation $H_\mu(L) = H(L+1)-H(L)$ precisely by the same amount of bits as E.

In order to generate one-dimensional paths we will make use of Hidden Markov Models (henceforth HMM). An HMM is a Markov Chain together with an alphabet $\mathcal{A}$. The states of Markov Chain are called *internal states* and may not be known to the observer. In fact, all the observer can see is a sequence of symbols drawn from $\mathcal{A}$. An HMM can emit a symbol either whenever a transition occurs or whenever we are at a given internal state. These are called *edge emitting* and *state emitting* HMMs respectively. In this project, we will be working with state emitting ones. One useful feature of Hidden Markov Models is that we can also obtain information measures about them. It must be noted we are using HMMs is in the context of generating of *self avoiding walks* (henceforth SAWs). In general, transition probabilities with lower "entropies" (in the sense of the entropy rate in transition emitting HMMs: $h_\mu = -\sum_{v\in V}\pi(v)\sum_{s\in A}\sum v' \in V T^{(s)}_{vv'}\log T^{(s)}_{vv'}$) generate walks that drift away from the origin at a faster rate and are thus less likely to self intersect. Considerations on the length of the SAWs will be relevant in the results section.

Finally, it is worth highlighting some of the many efforts that have been made to establish a clearer picture of 2-D information. In [2], the notion of *assymptotic compressibility* is introduced, which in some sense can be interpreted as a generalization of entropy rate. However, the goal is mainly compression and reproduction of data, which is at odds with faithfully capturing geometric structure. 1-D Information theory is performed along a Peano-Hilbert space filling curve, and consequently information in neighboring lattice sites will appear as unnatural correlations once translated into a string. Crutchfield and Feldman in [3] define a 2-D analogue of excess entropy and show that it is optimal under certain conditions. A noteworthy feature of it is that it distinguishes fundamentally distinct periodic structures, such as checkerboard patterns in different arrangements.

System and Methods

The system throughout the simulations performed for this project is a finite square lattice, of size ranging between 30x30 and 50x50. Each lattice site contains a binary symbol $\{\pm 1\}$. We will test different cases of interest such as random, periodic or symmetric configurations. A 1-dimensional path in the lattice thus induces a stochastic process from which we will obtain binary words. Measures of 1-D Information Theory will then be applied to such words.

Hidden Markov Models as Generators for Self Avoiding Walks

As mentioned in the beginning, in order to make as few (sampling) assumptions as possible we want to consider a sizeable enough ensemble of SAWs as opposed to a fixed path. The choice of employing HMM Markov Models for this task is a pragmatic one (despite the central role in the course). The simulations in this project were performed using MATLAB. One of its nice features is a built in HMM sequence generator. For our purposes the HMM generates sequences of four symbols: 1/2/3/4. Due to computational constraints, we cap their length at say 25 and then translate them to strings of the form $'uSRLL'$ where the first symbol is one of u, d, r, l (for each direction) and the subsequent ones are S, L, R (for "straight", "left" and "right" respectively). The translated sequence ends at the first point of intersection or at the (in this case) 25th symbol, whichever comes first.

```
for t=1:500

TRANS = [.9 .1; .5 .5];

EMIS = [1/4, 1/4, 1/4, 1/4;
1/4, 1/4, 1/4, 1/4];

[seq,~] = hmmgenerate(25,TRANS,EMIS);
for n=1:25
    A=seq(1:n);
    if sum(A(:) == 1)+sum(A(:) == 2) == sum(A(:) == 3)+sum(A(:) == 4)
      break

    else
```

Figure 1: Generating Self Avoiding Walks

In order to perform Information Theory on these words, we need probability distributions. This important task was developed in [4]. The idea is to keep track of all the distinct words (paths) that lead to a given lattice site and the number of times they occur. After appropriate normalizations, a probability distribution is obtained. Finally the information measures we are interested in are computed using the formulas in the background section. In particular, note that we are using the mutual information definition of excess entropy. This definition involves taking a limit. For obvious reasons, we will consider only the mutual information between the two halves of the word.

Spectrum of Mutual Information

Following the procedure in [4], after we have generated a significant enough sample of SAWs, we can construct a *spectrum of mutual information*. We simply plot the measures corresponding to each SAW in increasing order. The hope here is to determine whether qualitative aspects of the underlying pattern we are studying can be deduced by inspecting the spectrum of mutual information. For instance, Figure 2 contains the spectra for two random Ising configurations with different neighborhood definitions. It was hypothesized in [4], that the observable differences in their respective spectra could very well be an inherent feature of these two Ising models. An example of features we would hope to be able to read off from these spectra is, say, the intricateness of the boundary between the +1 and -1 subsets of this configurations. Does it, for example, explain the steadier slope in the case of the von Neumann neighborhood?

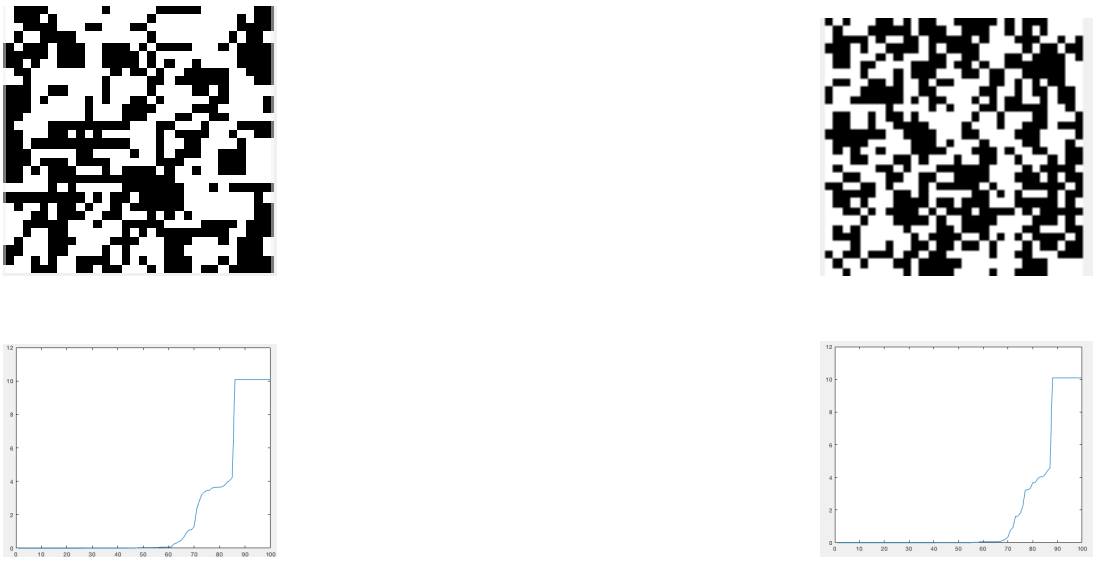


Figure 2: An example of Ising configuraration with Moore neighborhood (top left) and Von Neumann neighborhoods (top right) with their respective mutual information spectra with paths generated by the same HMM

Sliding Windows and Averaging

Some of the early simulations in this project motivated delving further into considerations of *scale*. Suppose we are trying to determine whether a given pattern is random or has some periodicity. Due to the limitations of the algorithm or computing power, our observations might turn out to be somewhat misleading. For instance, for computational limitations, we might have an ensemble of SAWs with length capped at 25 steps; effectively blinding us to periodicity at bigger scales. It is thus reasonable to wonder what happens when we discard detail at smaller scales and "zoom out" to observe more global behaviour. Fortunately, available to us is a technique from mathematical analysis, *convolution*, that generalizes quite nicely to matrices and allows us to essentially perform this operation of averaging out locally.

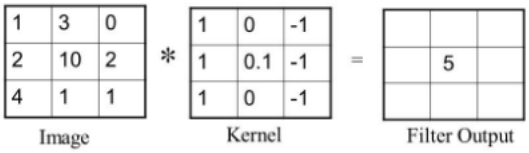


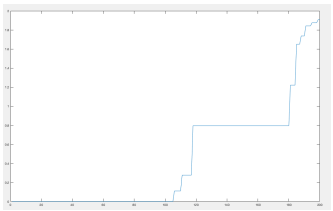
Figure 3: Convolution in Matrices

The exact mechanics of the procedure are depicted above. A kernel is a matrix of size $k \times k$. At any given site, the operation of convolution reassigns its value by getting the weighted average of its neighbors (within a $k \times k$ neighborhood) as specified by the kernel. We perform this for *every* lattice site, making the appropriate modifications when the $k \times k$ neighborhood overshoots the boundary. A similar convolution method involves covering the grid with nonoverlapping submatrices and then replacing *every* site inside them by their respective average. This method is more arbitrary and impractical when the side length of the grid has few divisors; it is thus discarded. The intuition behind this method is that it is in some sense analogous to the technique used in 1-D information of gathering statistics of substrings of length n. For the sake of simplicity, in our simulations we will fix a path and observe the block entropies of the words it generates as we increase k.

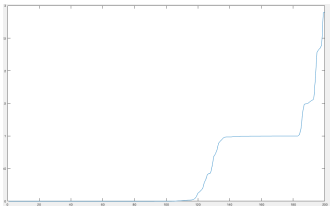
Results

Spectrum of Mutual Information

The starting point of our investigations will be to compare information spectra of markedly different configuration. One should expect that the overall shape of the spectrum to be a feature of the HMM we used to generate the paths; this in fact appears to be the case. However, information spectra yield some hints that we can confidently say are inherent to the configuration we are studying. An initial inspection of the mutual information spectrum for a random and periodic configuration suggests that a feature of particular interest is the range of achieved values.



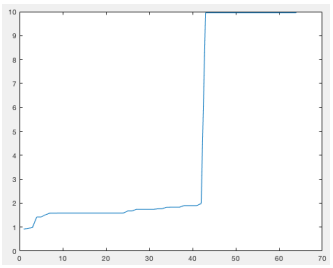
(a) 3 x 3 checkerboard



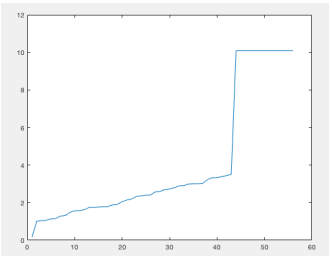
(b) Random Configuration

Figure 4: Two types of Neighborhoods

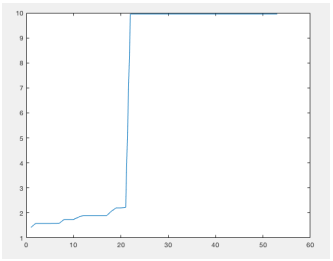
Periodic configurations tend to yield step-function-like information spectra. This should not be surprising as we will only observe a limited range of binary strings in the periodic case. In figure 4 above, the mutual information values for 200 SAWs of up to length 25 have been plotted with values ranging from 0 to almost 3 bits (in the random configuration).



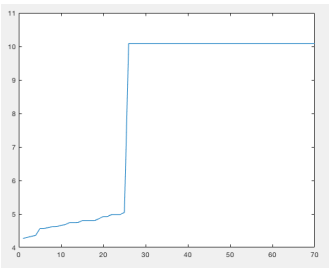
(a) 3x3 checkerboard (10)



(b) Random (10)



(c) 3x3 checkerboard (20)



(d) Random (20)

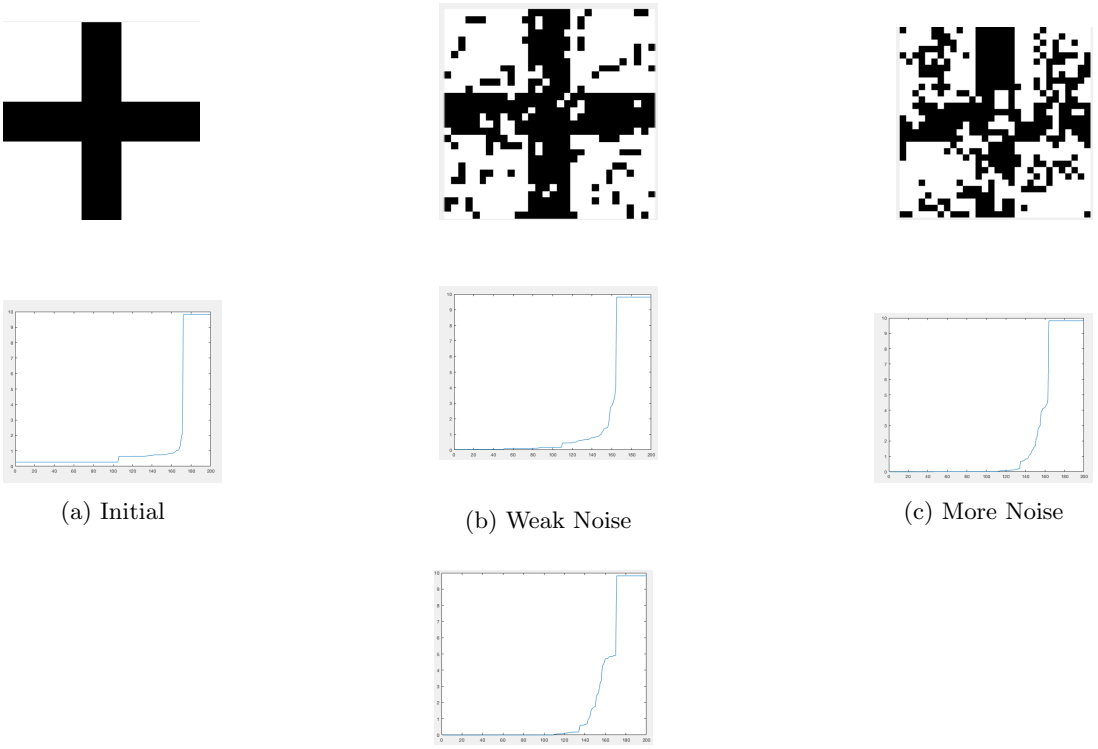
Figure 5: Spectra of fixed length (in parenthesis)

There are two inmediately apparent weaknesses in our initial analysis of figure 4. First of all we are ignoring any possible effect that length may have on mutual information. Secondly, and more importantly, we must recall that mutual information is being computed by looking at the two halves of the word; so for short ones in particular the data we obtain may not be

very illuminating. The next step in the simulations is therefore to consider spectra where all the SAWs have a fixed length and observe how they evolve as we increase it. Once again, we can observe that the random configuration appears to attain more values, even when imposing restrictions on the length of the SAWs (Figure 5). This confirms our suspicion that the range of values achieved is a relatively good proxy for randomness. It must be noted however, that as we consider longer walks an increasing proportion of them converges to a maximum value. Although not included in the figure above, by the time we consider walks of length 35 and above (note this is a 33x33 grid) effectively all walks attain this value. It is unlikely, that this particular value, is of any significance four our purposes since it coincides for different configurations. Consequently, as longer and longer SAWs start saturating the grid, the inspection of their information spectrum is decreasingly meaningful. A possible takeaway from this is that for a given grid size there is a theoretically optimal length for the SAWs we want to sample.

Limits of this Method

In this (sub)section we make a tentative proposal to quantify more precisely the quality of the mutual information approach and in turn to explain what is meant by optimal length in the previous paragraph. Due to time constraints we will only focus on the example in Figure 6 and leave a hypothetical formulation for the future. From previous observations, we can deduce that the utility of the mutual information spectrum relies crucially on how much it can differ depending on the underlying configurations. A more or less canonical reference point to measure this variation is that of a random configuration.



(a) Initial

(b) Weak Noise

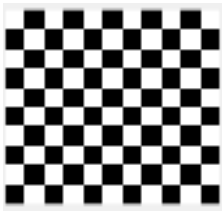
(c) More Noise

Figure 6: The evolution of MI spectra with increasing noise. For reference the last image(below) is the spectrum of a random configuration generated with the same HMM

Figure 6 displays an initial configuration and resulting configurations after noise has been added together with their information spectra. A perhaps trivial yet important feature of the mutual information spectrum is that as we add noise it shapes starts resembling that of a random configuration in a monotonic fashion. A couple of examples of gradual shifts from this to other nonrandom configurations also appear to exhibit the same monotonicity. Hypothetically, this behaviour could be exploited to devise a metric to measure distance between spectra (possibly one similar to known metrics of statistical distance such as Wasserstein, Kullblack-Leibler or the distance of total variation). For further investigations involving information spectra, such measures would give us a way to optimize our experiments. For instance, when considering ensembles of SAWs we would want to restrict our attention to lengths at which the distance between the spectra is maximized.

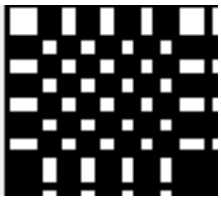
Block Entropy in the Sliding Windows

The initial simulations suggested that the lengths of SAWs was a crucial variable to account for. In fact, rescaling the checkerboard period and overall length of the SAWs accordingly appeared to have no impact in the profiles of MI spectra, but a more direct approach at observing periodicity at different scales was warranted. In particular, we wanted to remain agnostic about the scales at which some degree of structure may be observable. In this section we describe a few features about the method of sliding windows and in particular its suitability for detecting periodicity. In figure 7 we start with a 3x3 periodic pattern in 33x30 grid. The subsequent images correspond to the images obtained by convolving with sliding windows of the size k given below.



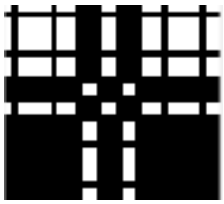
(a) Initial Configuration

(b) 6×6 windows



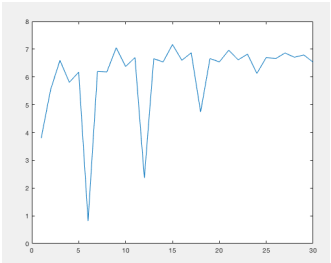
(c) 10×10 windows

(d) 12×12 windows



(e) 20×20 windows

(f) 30×30 windows

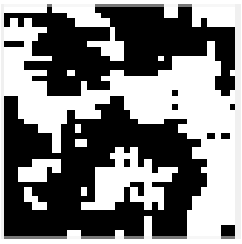


(g) Block Entropy vs k

Figure 7

In (g) we have also plotted the block entropies for the words generated by a fixed path that more or less travels along one of the grid's diagonals. A very salient phenomenon upon initial inspection are the "blackouts" whenever k is an even multiple of 3 (the periodicity). As it might be expected, these are accompanied by sudden drops in the block entropy plot (observe 6,12,18). This phenomenon is quite robust and can be replicated successfully in other checkerboard patterns. Simulations on other configurations and choices of path suggest some broad patterns in the behaviour of the block entropy plot and a couple of examples are included in the appendix. The block entropy value seems to stabilize as k gets closer to the size of the grid (note the smaller drop at 24). However, unlike in the case of the mutual information spectrum, different configurations do converge to different values although it is hard to give a satisfactory interpretation of that number. A second observation, one not that is perhaps not as strong is that, just how we could tell randomness by looking at the mutual information spectrum, randomness appears to leave a "signature" in the block entropy graph. In particular, in a random configuration the graph has a more or less logarithmic shape until it stabilizes and the fluctuations decay (Figure 8). Nonrandom configurations like the checkerboard either have a flat "trendline" or one that is more or less linear until they stabilize. Moreover, just as with the MI spectra, we have some degree of monotonicity as we add noise, meaning that the block entropy graph becomes more "log-like". One final observation is that the checkerboard pattern might be a singular one in the sense that not all structured (or even other periodic) configurations have the same flat profile. It is however the case that there might be other indications of periodicity. (Appendix)

(a) Initial Configuration

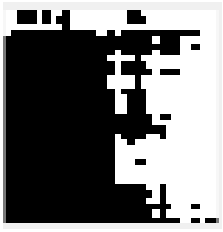


(b) 5x5 windows

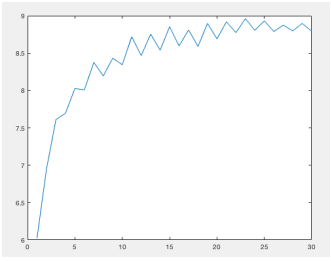
(c) 10x10 windows

(d) 20 x 20 windows

(e) 25 x 25 windows



(f) 30 x 30 windows



(g) Block Entropy vs k.

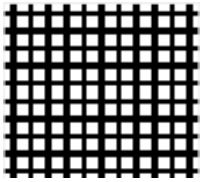
Figure 8

Conclusions

Presented with the conundrum of generalizing Information Theory to 2 dimensions, it appears to be the case that we ought to probe different approaches until a more clear picture is arises from the accumulation of partial results. The work in [4], for instance, explores informational signatures of symmetry. In this project, we have mainly focused on analyzing periodicity and, in particular, the role of *scale* in our measurements; namely length of SAWs and size of sliding windows respectively. As it was mentioned in the introduction, the main challenge in 2-D information theory is that it lacks the straightforward approach of classical IT to discrete time indexed stochastic processes. Consequently, any further exploration or refinement of the experiments presented in this project should focus on questioning some of the assumptions we made. For instance, what would constitute a good choice for the path to get the block entropies in the sliding window experiment? Should this choice be different if we were looking for nonperiodic structure? At any rate, as it is usually the case in science, the insights that Information Theory may provide will be enriched if complemented by other techniques from areas like computer science or signal processing.

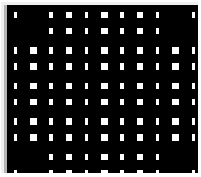
Appendix

In this appendix we give a couple of technical remarks on the sliding window method as well as two examples that suggested some of the more speculative claims made in the results section. First, it must be noted that in general any nontrivial convolution ($k > 1$) will yield noninteger values. We will simply consider positive values to be "+1" and negative ones "-1". Furthermore this also implies that initial configurations (such as Figure 9) where the "+1" and "-1" are not close to being evenly distributed, will quickly converge to a monochromatic configuration, thus limiting the usefulness of this method. Anyways, figure 9 and figure 10 show an example of a non-checkerboard periodic configuration and a noisier version of it. The block entropy graph in figure 9 shows consistently higher entropies than in the checkerboard case. However the fluctuations are noticeably less random (two monotonic subsequences can be identified) than in other examples; possibly being attributable to the periodicity. On the other hand, the block entropy graph in figure 10, shows faster growth for small values of k, more closely resembling the behaviour of a random initial configuration.



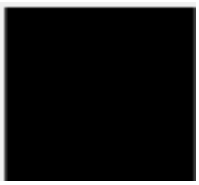
(a) Initial Configuration

(b) 3x3 windows

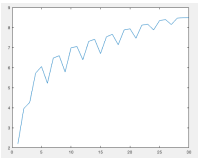


(c) 11 x 11 Windows

(d) 14x14 windows



(e) 19 x 19 windows



(f) Block Entropy vs k

Figure 9: Sliding windows on a periodic (3x3) cross pattern

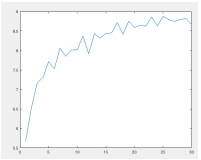
(a) Initial Configuration

(b) 5x5 windows

(c) 10x10 windows

(d) 15x 15 windows

(e) 20x20 windows



(f) Network 2

Figure 10: Half of the 3x3 cross blocks in the previous example are replaced by a random block

References and Acknowledgements

[1] J. P. Crutchfield and D. P. Feldman, "Regularities Unseen, Randomness Observed: Levels of Entropy Convergence", CHAOS 13:1 (2003) 25-54.

[2] Lempel, Abraham, and Jacob Ziv. "Compression of two dimensional data." Information Theory, IEEE Trans actions 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

[4] Jordan Snyder. "Pathwise Information Theory in Two Dimensions" NCAS0 2014 Final Project

This project was inspired by Jordan's work in [4]. I would like to thank him for generously lending his code and a slide for my presentation as well as suggestions for this project