UC DAVIS - DEPARTMENT OF PHYSICS

# εMSR

Submitted in Completion of PHY250 Course

**Jaspinder Paul Singh**

**6/6/2007**

Recently, it has been shown that the correlation function (gotten from the power spectrum by way of the Wiener-Kninchin theorem) can be used to reconstruct epsilon-machines. We demonstrate the reconstruction process with four processes – the fair coin, the period one, the period two, and the golden mean – and from them obtain probability plots similar to ones obtained directly from a data sequence.

## Introduction

A discrete sequence of data is an ideal data set. The period one process, which will be explained later, is a sequence of 1's. At any point, one knows the future and the past. A period two process, an alternating sequence of 1's and 0's, shares the same characteristic. One could have a slightly more complicated sequence of data, the golden mean, for example, but that too can be reduced into repeating patterns. Even seemingly random processes can be represented as finite state diagrams.

Can the same thing be said about the spectrum of a signal? Logically, if a signal contains useful information, so should transformations and/or manipulations of that signal. Unless the information is lost, or buried under noise, representing a signal as one would represent a discrete data sequence should be possible. We explore that notion using the ε-Machine Spectral Representation (εMSR) procedure (1). By examining the power spectrums with varying degrees of complexity, we ultimately and successfully reconstruct the probability information of the underlying process.

## Background

To understand εMSR, a concise understanding of ε-Machines will be given:

Given a stream of data, it is desirable to make future predictions. Assuming that the stream is not wholly random, this could be done by utilizing useful information from the past, hopefully with as little as needed. The intention is to find "causal states" – histories which share similar futures (2). An ε-Machine is essentially a probability model of a particular process. They are unique, deterministic, and minimal representations. Furthermore, they are Markovian – the conditional probability of being in a state from its entire history is equal to the conditional probability of being in that state from its immediate past state. The process of ε-Machine reconstruction will not be covered; all that is necessary to know is that it reproduces the word distribution of the process.

The *Wiener-Kninchin* theorem shows that the autocorrelation function of a sequence is related to its power spectrum through a Fourier transformation.

## Processes

The following stochastic processes were chosen for examination: period one, period two, unbiased ("fair") coin, and the golden mean process. For each process (except for the fair coin), an example sequence will be offered. The alphabet for each process is defined as $A = (0,1)$ .

### Period One

The period one process contains no zeroes anywhere: $\{..1111111..\}$

### Period Two

The period two process alternates ones and zeroes: $\{..1010101..\}$

### Golden Mean

The golden mean process contains no consecutive zeroes: $\{..101101011101..\}$

### Fair Coin

The fair coin is a uniform process where each equal-length word shares the same probability.

## Methods

For each process listed in the prior section, the correlation function must be obtained. The Wiesner-Knichin theorem states that the correlation function for a sequence of data is the Fourier transformation of the power spectrum. The power spectrum is defined as:

$$P(f) = |S(f)|^2$$

Where $S(f)$ is the Discrete Fourier Transformation of the sequence $S_N = s_0, s_1, \dots, s_n, \dots, s_{N-1}$:

$$S(f) = \frac{1}{\sqrt{N}} \sum_{m=0}^{N-1} e^{-2\pi i m f}$$

Without any loss of generality, any 0 in a data set has been replaced with a -1. Doing Fourier analysis of the power spectrum results in the correlation function $C(n)$:

$$C(n) = \int_0^1 P(f) \cos(2\pi n f) df$$

The correlation function is the probability that two members of a sequence of distance $n$ are identical. Therefore, for short-range correlations, the correlation function can be related to word probabilities – and likewise causal states. For reasons of convenience, we will make the following transformation to the correlation function:

$$q(n) = \frac{1}{2}[C(n) + 1]$$

It was shown that sequence probabilities return several useful constraints:

$$Pr(u) = Pr(0u) + Pr(1u) = Pr(u0) + Pr(u1)$$

$$\sum_{\omega \in \mathcal{A}^{r+1}} Pr(\omega) = 1$$

And that the sequence probabilities are related to the correlation function through

$$q(n) = \sum_{s=0,1} \sum_{\omega^r} \Pr(s\omega^r s)$$

For long-range $n$, the asymptotical value of the correlation function can be related to $Pr(1)$ and $Pr(0)$ as:

$$q_\infty = \big(Pr(0)\big)^2 + \big(Pr(1)\big)^2$$

By utilizing constraints amongst the probabilities, spectral equations for word various word lengths can be defined. For $r = 1$, there are 4 equations with 4 unknowns:

$$Pr(0) = Pr(00) + Pr(10) = Pr(01) + Pr(00)$$

$$Pr(11) + Pr(10) + Pr(01) + Pr(00) = 1$$

$$q(1) = Pr(11) + Pr(00)$$

$$q_\infty = \big(Pr(00)\big)^2 + \big(Pr(01)\big)^2 + \big(Pr(10)\big)^2 + \big(Pr(11)\big)^2$$

For $r = 2$, there are 7 equations with 8 unknowns:

$$Pr(001) - Pr(100) = 0$$

$$Pr(011) - Pr(110) = 0$$

$$Pr(001) + Pr(101) - Pr(011) - Pr(010) = 0$$

$$Pr(111) + Pr(101) + Pr(011) + Pr(001) + Pr(110) + Pr(100) + Pr(010) + Pr(000) = 1$$

$$q(1) = Pr(111) + Pr(110) + Pr(000) + Pr(001)$$

$$q(2) = Pr(111) + Pr(101) + Pr(000) + Pr(010)$$

$$q_\infty = \big(Pr(000) + Pr(001) + Pr(010) + Pr(011)\big)^2 + \big(Pr(100) + Pr(101) + Pr(110) + Pr(111)\big)^2$$

This cannot be solved without an 8th equation, which can be inferred by the following relation between 4th:

$$Pr(s_0s_1s_2s_3) = Pr(s_0s_1s_2)Pr(s_3|s_0s_1s_2) \approx Pr(s_0s_1s_2)Pr(s_3|s_1s_2) = \frac{Pr(s_0s_1s_2)Pr(s_1s_2s_3)}{Pr(s_1s_20) + Pr(s_1s_21)}$$

Combining this relation with the correlation function, the 8th equation is obtained:

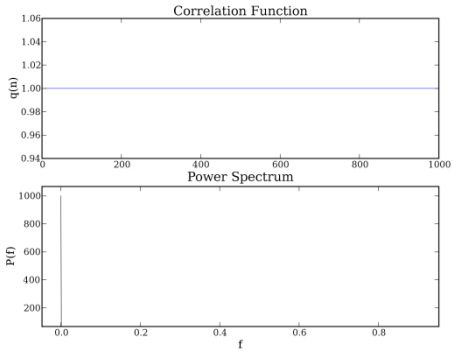
$$q(3) = Pr(1111) + Pr(1101) + Pr(1011) + Pr(1001) + Pr(0110) + Pr(0010) + Pr(0100) + Pr(0000)$$

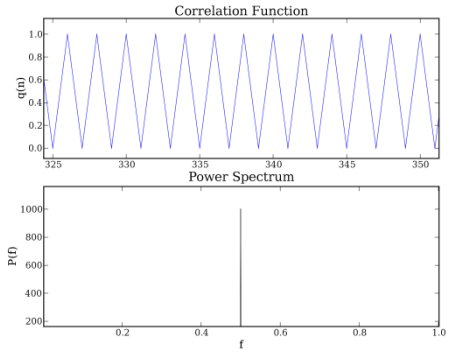
$$q(3) = \frac{Pr(111)^2}{Pr(110) + Pr(111)} + \frac{Pr(110)Pr(101)}{Pr(100) + Pr(101)} + \frac{Pr(101)Pr(011)}{Pr(010) + Pr(011)} +$$

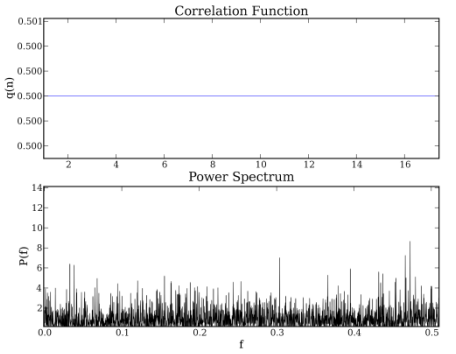
$$\frac{Pr(100)Pr(001)}{Pr(000) + Pr(001)} + \frac{Pr(000)^2}{Pr(000) + Pr(011)} + \frac{Pr(001)Pr(010)}{Pr(010) + Pr(011)} +$$

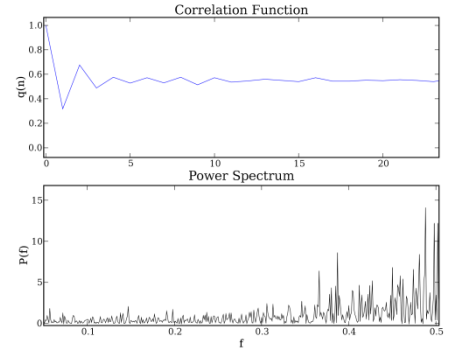
$$\frac{Pr(010)Pr(100)}{Pr(100) + Pr(101)} + \frac{Pr(010)Pr(100)}{Pr(110) + Pr(111)}$$

This is referred to as the *memory length reduction* approximation (2). With these spectral equations, we can determine the probabilities of different word lengths from the correlation function.




## Results – Correlation Functions and Power Spectrums

### Period One

Each term in a period one process is identical to every other term, hence the constant correlation (figure 1a). The εMSR probability distribution plots (figure 1b) confirm the existence of forbidden words; any word containing a zero.

### Period Two

The correlation function of the period two process (figure 1b) reveals a saw tooth structure, where $s_n$ is identical to $s_{n+2}$ with probability 1. The asymptotic value of $q(n)$ is taken to be 0.5.

### Unbiased Coin

εMSR reveals in the correlation function of the unbiased coin (figure 1c) that the probability between any two "flips" is 0.5.

### Golden Mean

For a more complicated process, the correlation function shows stronger fluctuations at short range distances with the spectral density increasing with frequency.

## Results – Probability Distributions

Taking the spectral equations defined in the previous section, the probability distributions are shown in Table 1. As an effective measure of accuracy for the εMSR process, they exhibit known characteristics of each function. The period one (figures 2a through 2c) forbid any words containing a zero. The period two process (figures 2d through 2f) forbids all words at any length which does not contain alternating 1's and 0's. For the

golden mean process, the L=1 plot (figure 2j) shows more probability for 1's and 0's, which is expected as the process limits the number of 0's. The L=2 plot (figure 2k) displays the first known forbidden word of the golden mean process, $\omega = 00$. The L=3 plot (figure 2l) forbids any work with sequential 0's, as well as shows more probability $\omega = 101$. The fair coin plots (figures 2g through 2i) are, essentially, a straight line, decreasing in likelihood as L increases.

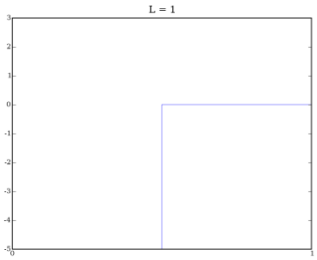



Figure 2a: Period 1, L=1

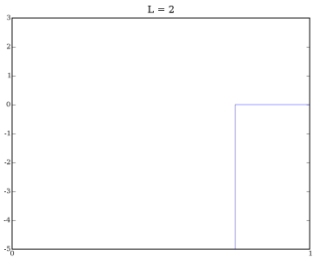



Figure 2b: Period 1, L=2

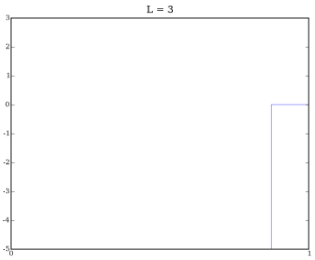



Figure 1c: Period 1, L=3

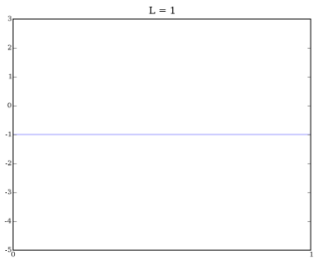



Figure 2d: Period 2, L=1

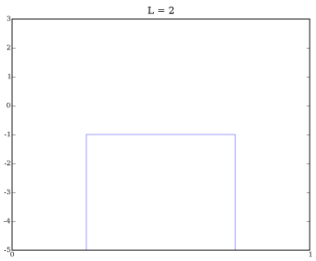



Figure 2e: Period 2, L=2

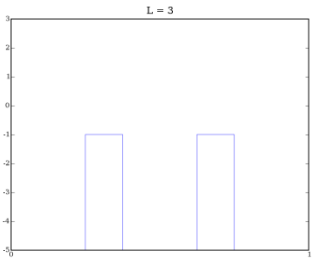



Figure 2f: Period 2, L=3

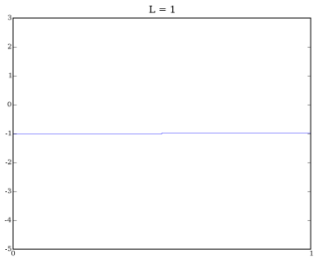



Figure 2g: Fair Coin, L=1

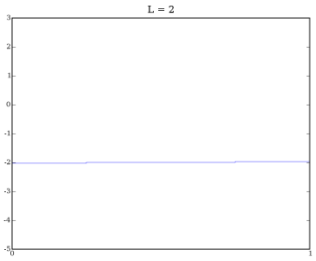



Figure 2h: Fair Coin, L=2

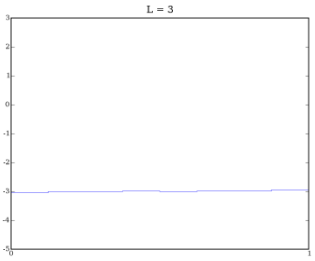



Figure 2i: Fair Coin, L=3

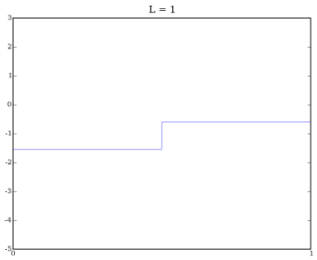



Figure 2j: Fair Coin, L=1

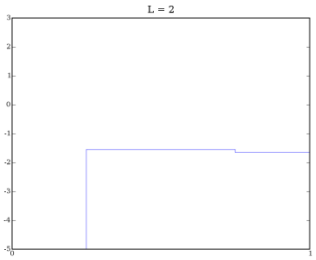



Figure 2k: Fair Coin, L=2

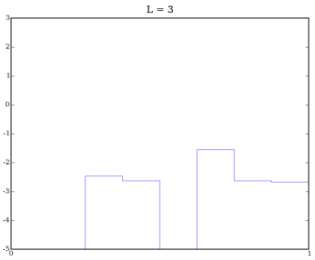



Figure 2l: Fair Coin, L=3

**Table 1: Probability distributions of the period one, period two, fair coin, and golden mean process (from top row to bottom row, respectively) at word lengths 1 through 3 (from left column to right column). The x-axis is word ω ,the y-axis is $log(Pr(\omega))$.**

## Conclusion

The εMSR procedure has been demonstrated to successfully result in probability distributions for stochastic processes up through $L = 3$. The distributions, based upon the power spectrum of the process, show accurate statistics indicate the presence of forbidden words. The correlation function still maintains the causal state information of the ε-Machine, and εMSR can be used to produce them. For more complicated processes, it is possible that higher-order correlations may be needed. The number of equations increases to 32 for length 4 words, and 64 for length 5 words, showing that solutions to higher-order spectral equations must be determined by computers. Furthermore, this procedure has shown promising results with zinc sulphide (3) and in the field of close-packed structures (4).


## Special Thanks

This project would not have been successfully completed without the help of Dowman Varn, Dr. Jim Crutchfield, and Christopher Ellison. Their aid is greatly appreciated.


## Bibliography


1. *"Inferring pattern and disorder in close-packed structures from x-ray diffraction studies, Part I: Type the background of the document here. e-Machine spectral reconstruction theory."* **Varn, D. P., Canright, G. S. and Crutchfield, J. P.** 2003, pp. [cond-mat/0302585].

2. *Natural Computation and Self-Organization Lecture Notes.* **Crutchfield, J. P.** 15-18a,

3. *"Inferring pattern and disorder in close-packed structures from x-ray diffraction studies, Part II: Application to zinc sulphide."* **Varn, D. P., Canright, G. S. and Crutchfield, J. P.** 2003, pp. [cond-mat/0302587].

4. *"Discovering planar disorder in close-packed structures from x-ray diffraction: Beyond the fault model"* **Varn, D. P., Canright, G S and Crutchfield, J. P.** 2002, Physical Review B, p. 66: 174110.