Metric Entropy and the Optimal Prediction of Chaotic Signals

Divakar Viswanath, Xuan Liang, and Kirill Serkh

January 12, 2013

Abstract

Suppose we are given a time series or a signal $x(t)$ for $0 \leq t \leq T$. We consider the problem of predicting the signal in the interval $T < t \leq T + t_f$ from a knowledge of its history and nothing more. We ask the following question: what is the largest value of $t_f$ for which a prediction can be made? We show that the answer to this question is contained in a fundamental result of information theory due to Wyner, Ziv, Ornstein, and Weiss. In particular, for the class of chaotic signals, the upper bound is $t_f \leq \log_2 T/H$ in the limit $T \to \infty$, with $H$ being entropy in a sense that is explained in the text.

If $|x(T - s) - x(t^* - s)|$ is small for $0 \leq s \leq \tau$, where $\tau$ is of the order of a characteristic time scale, the pattern of events leading up to $t = T$ is similar to the pattern of events leading up to $t = t^*$. It is reasonable to expect $x(t^* + t_f)$ to be a good predictor of $x(T + t_f)$. All existing methods for prediction use this idea in some way or the other. Unfortunately, this intuitively reasonable idea is fundamentally deficient and all existing methods fall well short of the Wyner-Ziv entropy bound on $t_f$. An optimal predictor should decompose the distance between the pattern of events leading up to $t = T$ and the pattern leading up to $t = t^*$ into stable and unstable components. A good match should have suitably small unstable components but will in general allow stable components which are as large as the tolerance for correct prediction.

An optimal predictor for chaotic signals should have three properties. First, it should achieve the Wyner-Ziv entropy bound. Second, it should look something like the classical Wiener-Kolmogorov predictor when the signal has zero entropy and no positive Lyapunov exponents. Third, it must be numerically stable. For the special case of toral automorphisms, we use Padé approximants and derive a predictor which has these properties and which seems to point the way to the derivation of a more general optimal predictor.

*Email: divakar@umich.edu, x5liang@ucsd.edu, kserkh@umich.edu. Address: (DV and KS) Department of Mathematics, University of Michigan, Ann Arbor; XL Department of Economics, University of California at San Diego. Supported by NSF grant DMS-0715510.
1 Introduction

Prediction of signals is a basic problem in science. In canonical problems of science such as planetary motion or the spectral theory of matter, the prediction problem is tackled by understanding the phenomenon that produces the signal and formulating a physical model for it. However, the need to predict signals remains in many instances where a physical model is unknown or is too complicated to be of practical use.

The problem of predicting signals without deriving physical models was first formulated independently by Kolmogorov and Wiener [23]. The Wiener-Kolmogorov predictors are linear and assume that the signal is characterized solely by its auto-correlation function. Under these assumptions, the predictors minimize the root mean square error and are optimal in that sense. A key assumption in the theory is that the signals are stationary.

Chaotic signals, which are signals obtained from dynamical systems that exhibit sensitive dependence on initial conditions, are stationary as well. However, the characterization of these signals by their auto-correlation functions is unsatisfactory, and the Wiener-Kolmogorov predictor should not be expected to and does not perform well at all. Available methods for the prediction of chaotic signals do not use spectral decomposition. These methods are based on recurrence and embed the signal in phase space in one way or another [1, 5, 9, 20].

The contribution of this paper is three-fold. Firstly, we show that all current methods for predicting chaotic signals are necessarily suboptimal. Current predictors are lacking in an important respect, as we will show. Secondly, we make a firm and significant connection between the problem of predicting chaotic signals and a major result in modern information theory. Thirdly, we investigate ideas that point the way to the derivation of an optimal general purpose predictor for chaotic signals.

The auto-correlation of a real valued signal \( x(t) \) is defined as

\[
\varphi(\tau) = \lim_{M \to \infty} \frac{1}{2M} \int_{-M}^{M} x(t)x(t+\tau) \, dt.
\] (1.1)

The auto-correlation function is of little use in predicting a chaotic signal. However, some of the mechanics of the Wiener predictor, especially with regard to its use of the auto-correlation function, reappear in the Padé predictor that we will derive later. The discussion of the Wiener predictor and its limitations in the simplified treatment of Bode and Shannon [3] is illuminating and merits careful study. Bode and Shannon ask themselves how much is lost by the restriction to linear predictors and answer as follows:

The fact that nonlinear effects may be important in a prediction can be illustrated by returning to the problem of forecasting tomorrow’s weather. We are all familiar with the fact that the pattern of events over a period of time may be more important than the happenings taken individually in determining what will come. For example, the sequence of events in the passage of a cold or warm front is characteristic. Moreover, the significance
of a given happening may depend largely upon the intensity with which it occurs. Thus, a sharp dip in the barometer may mean that moderately unpleasant weather is coming. Twice as great a drop in the same time, on the other hand, may not indicate that the weather will be merely twice as unpleasant; it may indicate a hurricane.

One of the authors of this passage, who grew up near the 45th parallel where the annual snowfall is more than 10 feet, may have watched the barometer during the long and harsh winters. For our purpose, the significant point here is the emphasis on “the pattern of events.” When the authors speak of the sequence of events in the passage of a cold or warm front being characteristic, it appears as if they had part of the intuition behind phase space reconstruction using delay coordinates.

Many methods have been proposed for predicting chaotic signals. All the current predictors known to us use delay coordinates. Suppose the real valued signal \( x(t) \) is obtained as \( x(t) = b^TX(t) \), where \( X(t) \) takes values in \( \mathbb{R}^d \) and \( b \in \mathbb{R}^d \) is constant. Suppose \( X(t) \) satisfies the dynamical system \( \dot{X} = f(X) \). Delay coordinates are an attempt to reconstruct the dynamics of \( X \) in \( \mathbb{R}^d \) using the scalar signal \( x(t) \). Even though an individual signal value such as \( x(t_0) \) may give little idea of \( X(t_0) \), the pattern of events \( x(t_0), x(t_0 - \tau), \ldots, x(t_0 - (k-1)\tau) \) can be used to stand as a substitute for \( X(t_0) \) and to reconstruct dynamics in phase space for suitable values of the delay \( \tau \) and the embedding dimension \( k \). The idea of using coordinates in this manner can be traced back to some embedding theorems in topology, and we will have more to say about this connection later.

For the most part, our discussion of existing predictors is restricted to one of the most basic ones [9]. The prediction method in its most basic form is as good as any of its many variants as far as optimal prediction is concerned, as we argue in Section 5. We refer to this basic form of the predictor, which we describe in Section 5, as the embedding predictor for predicting chaotic signals. The embedding predictor matches the pattern of events at the present with some time in the past, and uses the best match to predict the future. In that respect, it stays close to the logic of Bode and Shannon, while using delay coordinates to obtain a concrete realization of their notion of the pattern of events.

Perhaps the central point of this paper is that matching “the pattern of events” is not the best way to predict chaotic signals in spite of its indubitable intuitive appeal. This point is illustrated in Figure 1. Before explaining that figure, we set down some notation that will be used throughout this paper. If the signal is \( x(t) \), the current time is always denoted by \( T \). It is assumed that the signal is recorded from \( t = 0 \) and that the stretch of signal that is available is \( x(t) \) for \( 0 \leq t \leq T \). The task is to use the available history, which is \( x(t) \) for \( 0 \leq t \leq T \), to predict \( x(T + t) \) for \( 0 < t \leq t_f \) for as large a value of \( t_f \) as possible.

In the two plots of Figure 1, the thin black lines show a chaotic signal obtained from the Lorenz system. The plots show only a part of the signal and \( T \) is given as \( 2^{20} \) symbols. Each symbol is equal to \( t_{\text{return}} = 0.7511 \) units of time, where \( t_{\text{return}} \) is the average time from one “turning point” to another. A turning point is defined as a peak or a trough of
Figure 1: In both plots, the current time $T$ is $2^{20}$ symbols. The Lorenz signal, which is shown as a thin black line, is the same in the two plots. The thick red line is: (a) best fit from the past; (b) suboptimal prediction using the embedding predictor.
the graph of $x(t)$ with only peaks or troughs with $|x(t)| > 6\sqrt{2}$ being counted. Since the two fixed points that are located in the holes in the wings of the Lorenz attractor have coordinates $(\pm 6\sqrt{2}, \pm 6\sqrt{2}, 27)$, turning points defined in this way are in correspondence with intersections of the signal with Poincaré sections of the Lorenz attractor [22]. In the plots, $T$ is given as $2^{20}$ symbols, which means $T = 2^{20} \times 0.7511$ and that the number of turning points of $x(t)$ in $[0, T]$ is approximately $2^{20}$. In the plots, the thin black lines go beyond $T$ to show how the signal develops so that we can visually assess the quality of the predictions.

The thick red lines in the two plots are obtained differently. In the top plot, we fix a tolerance $tol$ (the precise value of $tol$ is unimportant for the discussion here) and look for $t^* \in [0, T - t_{\text{return}}]$ such that the length of fit, which is

$$\text{length of fit at } t^* = \text{largest } t_f \text{ such that } |x(T + s) - x(t^* + s)| \leq tol \text{ for } s \in [0, t_f],$$

(1.2)

is maximized. The maximum value of the length of fit is denoted by $t_{\text{best}}$. Here we are looking into the future of the signal and trying to find the moment $t^*$ in the past which agrees with the signal’s future for the maximum period $t_f$ (within a specified tolerance). This method of determining $t^*$ and the maximum length of fit $t_{\text{best}}$ will be called the best fit from the past. Since it looks at $x(T + s)$ for $s > 0$, the best fit from the past is not a predictor.

We see that the best fit from the past in Figure 1a follows the signal for $t > T$ for more than 20 symbols. It is not difficult to see why predictor can follow the signal for longer. If we fit the signal starting at $x(t)$, $0 \leq t \leq T - t_{\text{return}}$, to the signal starting at $x(T)$ the fit will extend from $T$ to $T + t_f$ for some $t_f$ and then start diverging. The rate of divergence beyond $T + t_f$ will be exponential as the signal is from a chaotic source. By definition of $t_{\text{best}}$, we have $t_f \leq t_{\text{best}}$. Thus the past has no information about what happens to the signal beyond $T + t_{\text{best}}$ and no amount of algorithmic legerdemain can synthesize that information.

Thus it should be clear that the basic task of a predictor is to find $t^*$ which maximizes (1.2) or another $t^*$ which nearly maximizes it without looking into the future. Let us see how the embedding predictor goes about predicting the signal. For the discussion here, a brief account of the embedding predictor suffices. A more detailed discussion including extensions and modifications of the basic predictor will be given later. The embedding predictor works by finding $t^* \in [k\tau, T - t_{\text{return}}]$ such that

$$\sum_{p=0}^{k-1} (x(T - p\tau) - x(t^* - p\tau))^2$$

(1.3)

is minimized. There is much literature about the choice of the delay parameter $\tau$ and the embedding dimension $k$ (see [1] for instance). We will assume that $\tau$ and $k$ are suitably chosen (with $\tau k$ about a fifth of a symbol). The prediction of $x(T + s)$ is taken to be $x(t^* + s)$.

How well does $t^*$ which minimizes (1.3) work in terms of maximizing (1.2)? Before answering that question, let us ask ourselves why there should be a connection at all
between finding \( t^* \) to minimize (1.3) and finding it to maximize the length of fit defined by (1.2). When we minimize (1.3), we are looking for a \( t^* \) such that if we walk back from \( t = t^* \) the portion of the signal we see looks much like what we see when we walk back from \( t = T \). In other words, the pattern of events leading up to \( t = t^* \) should look like the pattern of events leading up to \( t = T \). The hope is that if the events immediately preceding \( t = t^* \) look like the events immediately preceding \( t = T \), the signal value \( x(t^* + s) \) will be a good predictor of \( x(T + s) \).

Unfortunately, this intuitively reasonable idea is fundamentally deficient. To see why, we go back to Figure 1 on page 4. The thick red line of part (a) of that figure is obtained by shifting \( t^* \), which corresponds to the best fit from the past, to coincide with \( T \) to permit comparison between the two patterns. The thick red line of part (b) is obtained by shifting \( t^* \) found using the embedding predictor to \( T \). In part (a), we see that the sequence of events leading up to \( t = T \) and \( t = t^* \) are not close at all. Yet the two portions of the signal slam into each other at \( T \) and follow each other for more than twenty symbols. In part (b), on the other hand, the sequence of events leading up \( t = T \) and \( t = t^* \) are actually quite close. If we walk backwards, the pattern of events matches for three symbols. Yet the fit into the future is not half as good as in part (a).

As we explain in Section 6 and later, the distance between the pattern of events leading up to \( t = T \) and the pattern of events leading up \( t = t^* \) must be decomposed into stable and unstable components. For optimal prediction or for the best fit from the past, the unstable components must be made suitably small by weighting with the corresponding Lyapunov exponents, as will be explained. However, the stable components will in general be as large as the tolerance for an acceptable fit or prediction allows.

If we split the distance between the pattern of events leading up to \( t = T \) (black line in Figure 1a) and the pattern of events leading up to \( t = t^* \) for the best fit (thick red line with \( t^* \) shifted to \( T \) in Figure 1b), the distance between the two patterns has a noticeably substantial stable component but a small unstable component. However, the stable component decreases exponentially fast beyond \( t = T \) which means the two signals slam into each other. The small unstable component allows the fit between the two signals to persist for the longest interval of time.

In Figure 1b on the other hand, both the stable and unstable components of the distance between the two patterns is small, which means that the pattern of events leading up to \( t = T \) and \( t = t^* \) are noticeably much closer. However, the unstable component of the distance between the two patterns is not as small as in Figure 1a. Therefore, when we advance beyond \( t = T \) the red line diverges from the black line much earlier.

The situation shown in Figure 1 is typical. Because of the nature of chaotic signals, best fits tend to slam into the signal at \( t = T \) and diverge rapidly beyond \( t = T + t_{\text{best}} \). This introduces a fundamental asymmetry between the immediate past and the immediate future. Good agreement in the immediate past does not imply that the two portions of the signal will agree closely in the future.

Current predictors for predicting chaotic signals try to find a \( t^* \) such that the pattern of events leading up to \( t = t^* \) closely resembles the pattern of events leading up to \( t = T \).
If the goal is to predict the signal as far into the future as possible, that is not the right idea. The right idea for an optimal predictor is to evaluate if the pattern of events leading up to \( t = t^* \) and \( t = T \) are such that the two patterns will come close to each other in the future and to calculate for approximately how long they will remain close. Such a calculation requires us to decompose the distance between the two patterns into stable and unstable components.

Our notion of optimality is the length of the fit into the future, which is defined by (1.2). Another notion of optimality may be to predict \( x(T + t_0) \) as accurately as possible for a fixed \( t_0 \). The notion of optimality that we adopt exposes the basic aspects of the prediction problem much better. Indeed, an optimal predictor according to our notion of optimality, such as the one we derive for toral automorphisms, can be easily modified to be optimal in the other sense. The basic point we make, which is the need to decompose the distance between two patterns into stable and unstable components, is valid no matter how the prediction problem is posed.

In our discussion here, we have stated that \( t^* \) is found in the interval \([0, T - t_{\text{return}}] \). Since \( t_{\text{return}} \) is equal to a symbol and the fits can extent for more than 20 symbols as evident from Figure 1 it may seem that the interval needs to be bounded away from \( T \) by a greater amount. That is a minor point, because \( t^* \) is nearly always smaller than \( T \) by a large multiple of \( t_{\text{return}} \). If we want to be certain, we can take the interval for \( t^* \) to be \([0, T - 100t_{\text{return}}] \).

Section 2 states a theorem of Wyner-Ziv [24] and Ornstein-Weiss [19] and Sections 3 and 4 develop the implications of the entropy bound in that theorem to the prediction of chaotic signals. Heuristically, the theorem says that

\[
\lim_{T \to \infty} \frac{t_{\text{best}}}{\log_2 T} = \frac{1}{H}
\]

with probability 1. Here \( H \) is entropy in a sense that will be described. A predictor is optimal if it predicts the signal in the interval \([T, T + t_f] \) and

\[
\liminf_{T \to \infty} \frac{t_f}{\log_2 T} \geq \frac{1 - \epsilon}{H}
\]

with probability 1 and for any \( \epsilon > 0 \). In Section 5, we discuss current predictors and point out why they are necessarily suboptimal. From Section 6 onwards, we develop a few ideas that take us closer to a general purpose optimal predictor for chaotic signals.

More introductory remarks can be made about the contents of the later sections. To bring this introduction to a close, we defer those remarks to later sections and point out a few aspects of the prediction problem that we do not consider.

The signals considered here are assumed not to be noisy, although it is true that the signals are noisy in experimental situations. It is difficult to imagine how optimal predictors for noisy chaotic signals can be derived when they are unavailable for noiseless chaotic signals. If the underlying properties of chaotic signals which enable and limit prediction are not understood, it does not seem possible to put prediction theory of
noisy chaotic signals on a secure foundation. A penetrating investigation of the effect of
noise on chaotic signals is due to Lalley [17]. Lalley’s algorithm D uses delay coordinates
much like many other investigations [1] [8] [12]. While proving the validity of algorithm
D as a method for removing the effect of noise, Lalley recognizes the need to require the
width of the window used by the delay representation to increase at a sub-logarithmic
rate.

We limit ourselves to stationary signals and assume that no physical model of the
source of the signal is available. For the possibilities created when something is known
about the physical model of the source of the signals, one may consult the use of Bayesian
models in the striking discoveries of Brown and others [4].

We limit ourselves to an single signal. The early work of Wiener [23] already considered
ways to improve prediction when several signals are recorded perhaps at different points
in space. As Weiner explained, given the tendency of weather to move east, Chicago
weather may be more important for predicting Boston weather than Boston weather
itself. We mention the intriguing work of Chernyshenko and Bondarenko [7] where it
is shown that an entire turbulent velocity field can be recovered using only 2% of the
modes. Remarks pertaining to spatially extended signals are found in the concluding
section. For a study of the ergodic nature of turbulent events, see [18].

2 Theorem of Wyner-Ziv and Ornstein-Weiss

In this section, we describe three results that apply to stationary and ergodic sequences:
the Poincaré recurrence theorem, a theorem of Kac, and the entropy theorem of Wyner-
Ziv and Ornstein-Weiss. Each of these results is pertinent to source coding and, as we
will show, to the prediction of chaotic signals.

The notion of stationarity can be defined for a sequence of random variables or for a
dynamical system. Since our interest is in the prediction of signals, we begin with the
definition for a sequence of random variables. A sequence of random variables

\[ X_0, X_1, X_2, \ldots \]

is stationary if

\[ \mathbb{P}\left( (X_n, X_{n+1}, X_{n+2}, \ldots) \in B \right) = \mathbb{P}\left( (X_{n+1}, X_{n+2}, X_{n+3}, \ldots) \in B \right) \]

for any Borel measurable subset \( B \) of \( \mathbb{R}^\infty \). The definition captures the idea that the
mechanism underlying the stochastic process does not change with time.

A stationary sequence is ergodic if every invariant event has probability 0 or 1. Events
phrased using means and correlations of the sequence are examples of invariant events.

For an alternative definition, let \( T : \Omega \rightarrow \Omega \) be a measurable transformation that
preserves the probability measure \( \mu \) on \( \Omega \). The set \( A \subset \Omega \) is invariant if \( T^{-1}A = A \).
The transformation \( T \) is ergodic if \( \mu(A) = 0 \) or \( \mu(A) = 1 \) for every invariant set \( A \). The
ergodicity condition precludes the dynamics from getting stuck in a part of phase space.

The Poincaré recurrence theorem does not assume ergodicity.
Theorem 2.1 (Poincaré recurrence [16]). Assume \( X_0 \) to be \( \mu \)-distributed and define the stationary sequence \( X_0, X_1, \ldots \) with \( X_n = T^n(X_0) \) for \( n = 1, 2, \ldots \). For a measurable subset \( B \) of \( \Omega \) with \( \mu(B) > 0 \), \( X_0 \in B \) implies \( X_n \in B \) infinitely often with probability 1.

Suppose a long stream of text is modeled as a stationary sequence of characters and suppose that the set \( B \) is chosen to prescribe the first ten characters of the text. The theorem then asserts that the sequence formed by the first ten characters will repeat again and again. The origin of the sequence \( X_0 \) can be taken anywhere in the text.

If \( \Omega \) is the phase space of a dynamical system, the theorem asserts that the dynamical system will revisit the same region \( B \) in phase space infinitely often. These revisitations are the basis for predicting chaotic signals.

The Poincaré recurrence is not quantitative. It does not tell us by what factor a long stream of text can be compressed if the repetitions are exploited or how well a chaotic signal can be predicted by tracking the recurrences. The first step to a quantitative version of the Poincaré recurrence theorem is a lovely theorem of Kac. This theorem assumes the sequence to be ergodic.

Theorem 2.2 (Kac’s theorem [15]). Suppose that the sequence \( X_0, X_1, \ldots \) is stationary and ergodic. Let \( B \subset \mathbb{R} \) with \( P(B) = P(X_0 \in B) > 0 \). Let \( n \geq 1 \) be the smallest integer such that \( X_n \in B \). Then \( \mathbb{E}(n|X_0 \in B) = 1/P(B) \).

Kac’s theorem says that the expected time to return to the set \( B \) is exactly equal to the inverse of the probability of \( B \). One would expect the recurrence time to sets of smaller probability to be greater. While the elegance of Kac’s theorem may lead one to suspect that the theorem should be obvious or easy to demonstrate, a perusal of Kac’s ingenious proof will dispel such a misperception.

The entropy theorem stated below characterizes recurrences more sharply than Kac’s theorem. It applies to sequences which are stationary, ergodic, and take values in a finite alphabet. The restriction to finite alphabets does not cause such a great loss of generality because information is fundamentally discrete in nature. Chaotic signals are real valued and often continuous in time. Yet we may obtain a notion of optimality of prediction of chaotic signals using the entropy theorem, as we will show in the following sections.

Theorem 2.3 (Ornstein and Weiss [19]). Let \( X_0, X_1, \ldots \) be a stationary and ergodic sequence, in which each \( X_n \) takes values in a finite alphabet \( A \). Let \( t_{\text{best}} \) be the greatest integer such that \( X_{T+1}, \ldots, X_{T+t_{\text{best}}} \) occurs as a contiguous subsequence of \( X_0, \ldots, X_T \). Then

\[
\lim_{T \to \infty} \frac{t_{\text{best}}}{\log_2 T} = \frac{1}{H}
\]

with probability 1. Here \( H \) is the entropy of the stationary, ergodic process \( X_0, X_1, \ldots \).

Theorem 2.2 tracks the re-occurrence of an event associated with \( X_0 \) for some \( X_n \) with \( n > 0 \). Theorem 2.3 checks if an event that follows the current symbol \( X_T \) repeats a past event. We will refer to either scenario as a recurrence.
Theorem 2.3 is a remarkable sharpening of the Poincaré recurrence theorem. If we regard $T$ as current time and that observations begin at 0, as we do throughout this paper, it gives a perfect characterization of the extent to which the pattern that will follow $T$ will repeat some pattern of events we have seen in the past. The $1/H$ bound was first stated by Wyner and Ziv [24], who were able to prove the convergence of $t_{\text{test}}/\log_2 T$ to $1/H$ in probability. Almost sure convergence of the type asserted by Theorem 2.3 was proved by Ornstein and Weiss [19].

The distinction between convergence in probability and almost sure convergence is pertinent to the prediction of chaotic signals. If predictions of weather or of hurricane tracks or of cardiac signals are to be really useful, the prediction method should apply to almost every signal and not only to a fraction of the signals that occur in practice. The distinction between almost sure predictions of individual signals and statistical predictability has not been made in extant work on the subject. Existing predictors of chaotic signals have been validated generally with statistical notions of accuracy such as mean square error and correlation plots [9, 20]. Our discussion of predictability of chaotic signals will be framed in terms of almost sure predictability.

Since $X_n$ takes values in a finite alphabet $A$ for $n \geq 0$, we refer to each value as a symbol. The entropy $H$ is defined as follows. Suppose we consider the following block of symbols of length $m$: $X_0, \ldots, X_{m-1}$. This block can take $|A|^m$ different values. Suppose the probabilities of the different possibilities are $p_1, p_2, \ldots, p_M$, where $M = |A|^m$. Then

$$H = \lim_{m \to \infty} \frac{1}{M} \sum_{i=1}^{M} -p_i \log_2 p_i.$$

In dynamics, natural logarithms are used instead of logarithms to base 2. We will follow the information theory convention and use logarithms to base 2.

The definition of entropy comes up in a natural way when we try to count states. Suppose we look at all $|A|^m$ possible values of the sequence $X_0, \ldots, X_{m-1}$. Some possible sequences are more probable and some are less probable. How may possible sequences have a probability that is approximately that of the average? The answer is $2^{mH}$. The entropy theorem of Shannon and others asserts that a sufficiently long segment of $X_0, X_1, \ldots$ looks like an average sequence with probability 1. Therefore to transmit $m$ symbols from such a stationary and ergodic source, we may be able to get by using $mH$ bits but no less. An optimal compression of the source will use $mH$ bits to encode $m$ symbols asymptotically.

Entropy comes up in statistical mechanics while counting the number of states of a system. The interpretation of entropy in terms of information originated with Shannon’s source coding theorem. However, the coding scheme implicit in Shannon’s theorem, which is to use long block codes, is useless in practice. The widely used source coding scheme of Lempel and Ziv relies on an entirely different interpretation of entropy, which is the interpretation given by Theorem 2.3.

Theorem 2.3 interprets entropy in terms of the maximum segment following $X_T$ that occurs as a subsequence of the segment preceding it. The entire segment following
3 APPlicability of the entropy theorem to Chaotic systems

$X_T$ can be encoded using a pointer to some place in the past and the length of the recurrence. Various source coding schemes based on that idea have been derived by Lempel, Ziv and others and have been widely used for decades. The distinction between almost sure convergence and convergence in probability is important for the practical success of these coding schemes.

In normal use, entropy theorem refers to the entropy theorem of Shannon. In this paper, entropy theorem and entropy bound will refer to Theorem 2.3. This convention saves us the trouble of using four names everytime we need to refer to the theorem and the bound contained in it.

If we look at the entropy theorem in the light of prediction, it appears as if $\log_2 T/H$ symbols can be predicted using a history of length $T$. The fallacy behind that surmise becomes evident if we consider an i.i.d. sequence made up of $\pm 1$, where each sign is equally probable. The entropy of such a sequence is 1. As the entropy theorem asserts, we may expect $\log_2 T$ symbols that follow a history of length $T$ to form a segment that repeats a segment that has already occurred. That type of repetition is useless for prediction. Given a knowledge of the history of the signal up to $X_T$, all that we know about $X_{T+1}$ is that it is equally likely to be $+1$ or $-1$. Knowledge of history is useless in the prediction of i.i.d. sequences.

Thus we need to precisely delineate the nature of chaotic signals which makes the entropy theorem relevant to their prediction. In Section 3, we describe the notion of entropy for chaotic signals, and in Section 4, we explain why the entropy theorem defines the limit of predictability of chaotic signals. At the end of Section 6, we describe what form on optimal predictor should take. While currently available predictors do not take that form, in the rest of the paper, we describe a few ideas that suggest that optimal predictors can in fact be derived.

3 Applicability of the entropy theorem to chaotic systems

Stationary and ergodic sequences can be generated in many ways. An i.i.d. sequence $X_0 = \pm 1, X_1 = \pm 1, \ldots$ with $p(+1) = p(-1) = 1/2$ is stationary and ergodic. Suppose we form another sequence $Y_n$ with $Y_n = 1$ or $Y_n = -1$ according as $+1$ or $-1$ is the majority among the seven entries $X_n, \ldots, X_{n+6}$. Then the $Y_n$ sequence is also stationary and ergodic. Regardless of the length of history neither the $X_n$ sequence not the $Y_n$ sequence is predictable in the manner we consider. For notions of prediction pertinent to such signals, see [10].

Suppose $X_{n+1} = f(X_n)$ is a dynamical system. The phase space of the the dynamical system can be any Riemannian manifold but for convenience we will assume it to be a subset of $\mathbb{R}^d$. Let $\mu$ be a probability measure that is invariant with respect to the dynamical system. If $X_0$ has $\mu$ as its distribution and $X_{n+1} = f(X_n)$ for $n = 0, 1, \ldots$, the sequence $X_0, X_1, \ldots$ is stationary. If $\mu$ is indecomposable (an assumption we will always make), the sequence is ergodic as well.

It is evident that a stationary and ergodic sequence $X_0, X_1, \ldots$ generated in this manner is quite different from an i.i.d. sequence of the type $\pm 1, \pm 1, \ldots$. While the i.i.d. sequence generates a random number for every new entry, in a stationary and ergodic
sequence derived from a dynamical system, every new entry is generated deterministically.

We do not assume the entire state vector $X_n$ to be observable. The observed sequence is $x_0, x_1, \ldots$ where $x_n$ is a real-valued function of $X_n$. For example, $x_n$ can be some component of $X_n$. This framework should be sufficiently general to allow for seismic signals, ECG signals and so on. Nearly all the theoretical discussion will be restricted to maps to avoid some of the technicalities that arise for flows. For both maps and flows, the dynamical system that generates the signal is assumed to be unknown.

One of the examples we consider is a signal obtained from the Lorenz flow:
\[
\begin{align*}
\frac{dx}{dt} &= 10(y - x) \\
\frac{dy}{dt} &= 28x - y - xz \\
\frac{dz}{dt} &= -8z/3 + xy.
\end{align*}
\]

The Lorenz system has fixed points at $(0, 0, 0)$ and $(\pm 6\sqrt{2}, \pm 6\sqrt{2}, 27)$. The two nonzero fixed points sit in the middle of two holes in the wings of the butterfly-shaped attractor’s two wings. The signal is generated by accurately integrating a random point $(x', y', z')$ for some time to generate the initial point $(x(0), y(0), z(0))$. The initial point generated in this may be assumed to be $\mu$ distributed, with $\mu$ assumed to be the physical measure of the Lorenz attractor. The signal $x(t)$ is generated for $t \geq 0$ by integrating this initial point. For the purpose of prediction, it is assumed that the model which generates the signal is unknown.

To apply the entropy theorem to the Lorenz signal $x(t)$, we need to specify the entropy of the Lorenz signal. We recall a few of the theoretical definitions related to the entropy of a dynamical system. For complete details, see [16] or [25]. Let $f : \mathbb{R}^d \to \mathbb{R}^d$ be a smooth dynamical system and let $A$ be an invariant set. Let $\mu$ be a probability measure on $A$ that is invariant with respect to $f$. Assume that $f$ is an ergodic transformation of $A$ with respect to the measure $\mu$. In this setting, the definition of metric or Kolmogorov-Sinai entropy is quite simple. Let $P$ be a finite partition of the set $A$. We can generate a finite-valued stationary ergodic process as follows. Pick $X_0$ according to $\mu$ and take $X_{n+1} = f(X_n)$ for $n = 0, 1, \ldots$ Let $Y_n$ be the partition in $P$ that $X_n$ belongs to. Then the finite valued process $Y_n$ is stationary and ergodic, and as such has a Shannon entropy which we denote by $h_\mu(f, P)$. In general, $h_\mu(f, P)$ can depend upon the partition $P$. The metric entropy $h_\mu(f)$ is defined as the maximum over all finite partitions $P$.

At first sight, it might seem as if the metric entropy can depend upon $P$. However, this dependence is not as severe as one might think. For example, one may modify $P$ to the finer partition $P \vee P$, where the finer partition keeps track of the partitions in $P$ that $x$ and its iterate $f(x)$ belong to. Even though $P \vee P$ is a finer partition, $h_\mu(f, P \vee P) = h_\mu(f, P)$ because it is readily evident that combining the $n$-th and the $(n + 1)$-st symbols into a single symbol in the $n$-th position will neither increase nor decrease the information per symbol. In fact, $h_\mu(f) = h_\mu(f, P)$ if the partition $P$ is
3 APPLICABILITY OF THE ENTROPY THEOREM TO CHAOTIC SYSTEMS

| log₂ T | t_{best} | Matching sequence | log₂ T | t_{best} | Matching sequence |
|--------|----------|-------------------|--------|----------|-------------------|
| 2      | 2        | BA                | 12     | 13       | AAAABAAABBBAAA   |
| 3      | 3        | BBB               | 13     | 12       | BBBAABAAABAA     |
| 4      | 7        | ABBBAAA           | 14     | 16       | BBBABBBAAAAAABBA|
| 5      | 7        | BBBAABAA          | 15     | 17       | BBBAABBBABBBABBB|
| 6      | 9        | BBBAABBBAAA       | 16     | 16       | BBABBBAAAABBB    |
| 7      | 10       | AAAAAAAAAAAAAAAAA | 17     | 20       | AAABBBAAABBBAAA  |
| 8      | 14       | AABAAABAAAAAAAAAA| 18     | 24       | BABBABBBABBBBABBB|
| 9      | 9        | BBBAABAAA         | 19     | 20       | BBBAABBBAAABAAA  |
| 10     | 10       | BBABBBBABB        | 20     | 16       | ABBAABBBAAABBA   |
| 11     | 9        | ABBBAAAA          | 21     | 30       | BAABBBAAABBBAAAA|

Table 1: Recurrences of a Lorenz signal.

generating. Generating partitions are defined using conditional entropy [10]. If the partitions in \(P \lor \ldots \lor P\) become fine enough to closely approximate any given partition \(Q\) of \(A\), the partition \(P\) is generating.

Later the theoretical discussion will focus on hyperbolic attractors \(A\). For such invariant sets, Markov partitions are generating. But now we will explain how the concept of metric entropy allows us to apply the entropy theorem to Lorenz signals.

Table 1 shows a calculation of \(t_{best}\), in accord with its definition in the entropy theorem (Theorem 2.3), using a Lorenz signal. The symbols \(A\) and \(B\) have the following meaning. Every intersection of the Lorenz signal with the “quarter” plane \(x < -6\sqrt{2}, y < -6\sqrt{2}, z = 27\) is recorded as the symbol \(A\) and every intersection with \(x > 6\sqrt{2}, y > 6\sqrt{2}, z = 27\) is recorded as the symbol \(B\). In this manner the Lorenz signal is turned into a stationary and ergodic sequence of \(A\)s and \(B\)s. For evidence that the partition into \(A\) and \(B\) is generating, see [22].

A convenient way to estimate the entropy of the sequence of \(A\)s and \(B\)s is using Lyapunov exponents. Lyapunov exponents are the exponential rates with which infinitesimal perturbations to a point on \(A\) grow or decay. For a definition, see [16]. The standard definition uses natural logarithms and not logarithms to base 2 as in the case of entropy. If the system is of dimension \(d\), there are exactly \(d\) Lyapunov exponents with repetitions. With probability 1 with respect to the measure \(\mu\), these are the only possible rates of growth or decay.

If the Lyapunov exponents are \(\lambda_1, \ldots, \lambda_d\), the metric entropy satisfies

\[
h_\mu \leq \sum_{\lambda_i > 0} \lambda_i.
\]  \(3.1\)

This is Ruelle’s inequality [25] (the same logarithm must be used in defining \(h_\mu\) and the Lyapunov exponents \(\lambda_i\)). In some cases, equality holds in \(3.1\).

For the Lorenz system, the continuous time Lyapunov exponent is approximately 0.905 (using natural logarithms). The average time from an intersection with one of the
4 Recurrence and predictability

Table 2: Recurrences of flips of a fair coin.

| $\log_2 T$ | $t_{\text{best}}$ | Matching sequence | $\log_2 T$ | $t_{\text{best}}$ | Matching sequence |
|------------|--------------------|-------------------|------------|--------------------|-------------------|
| 2          | 6                  | AABAAA            | 12         | 12                 | AAABAAABABBB      |
| 3          | 4                  | AAAB              | 13         | 12                 | BABAAAAABAAB      |
| 4          | 8                  | AAAAAABB          | 14         | 13                 | BAAABABABABBB     |
| 5          | 7                  | BBABBBAB          | 15         | 14                 | BBAAAAAAAAAAAB    |
| 6          | 6                  | BBAAAA            | 16         | 16                 | BAAABABABBAABBB   |
| 7          | 11                 | ABBABBBB          | 17         | 20                 | AAAABABAABABABABAB |
| 8          | 5                  | BBBBB             | 18         | 17                 | BAAABAAABABABBB   |
| 9          | 11                 | BBBBABBABBABBA    | 19         | 22                 | BBBBBAAAABBBABBBBAAB |
| 10         | 12                 | BBABBAABBAABA     | 20         | 19                 | BBBBBAAAABBBBBBABB |
| 11         | 10                 | BBABABBBABBABBA   | 21         | 20                 | ABBBBBBBBBBBBBBBB |

Very probably the entropy is quite close to $0.98$. Table 1 appears to be in agreement with this estimate of the entropy.

Table 2 tabulates $t_{\text{best}}$ (defined as in Theorem 2.3) for tosses of a fair coin (with $A$ for heads and $B$ for tails). The entropy of the coin toss process is 1 and very close to the entropy of the Lorenz signal. Yet Table 2 looks quite different from Table 1. The fluctuations of $t_{\text{best}}$ are more pronounced for the Lorenz signal. For the special case of i.i.d. sequences such as coin tosses, Theorem 2.3 was proved by Erdos and Renyi.

The intersection with the quarter-planes $A$ or $B$ to another is $t_{\text{return}} = 0.7511$. By Ruelle’s inequality (3.1), the entropy of the sequences of As and Bs is bounded above by $0.905 \times 0.7511 / \log 2 = 0.98$. Very probably the entropy is quite close to 0.98. Table 1 appears to be in agreement with this estimate of the entropy.

When we think of the Lorenz signal as a sequence made up of the symbols $A$ and $B$, it is natural to define $t_{\text{best}}$ as in the entropy theorem (Theorem 2.3). However, for continuous time signals the definition of $t_{\text{best}}$ which follows (1.2) is more natural. We take

$$tol = 5$$

(3.2)

to be the tolerance for Lorenz signals throughout this paper. Table 3 reports $t_{\text{best}}$ with $tol = 5$. The $t_{\text{best}}$ numbers with $tol = 5$ are somewhat smaller than the $t_{\text{best}}$ numbers in Table 1. This is because $tol = 5$ is a stiffer requirement than simply requiring the symbol sequences to match. When other methods are compared to the best fits in Table 3 later, the length of match is reported in symbols but not as a real number.

4 Recurrence and predictability

Suppose we are trying to predict a signal $x_0, \ldots, x_T$. The entropy theorem says that $t_{\text{best}} \approx \log_2 T/H$ for large $T$. Thus it appears the past of the signal does not have sufficient information to predict $x_{T+t}$ for $t > \log_2 T/H$. We expect that no algorithm
Figure 2: Best fits from the past (in thick red) to a Lorenz signal (in thin black). (a) $T = 2^{14}$ symbols. (b) $T = 2^{21}$ symbols.

| $\log_2 T$ (in symbols) | $t_{\text{best}}$ (in symbols) | $t_{\text{best}}$ (as a real) | $\log_2 T$ (in symbols) | $t_{\text{best}}$ (in symbols) | $t_{\text{best}}$ (as a real) |
|--------------------------|---------------------------------|-------------------------------|--------------------------|---------------------------------|-------------------------------|
| 2                        | 1                               | 0.58                          | 12                       | 9                               | 6.54                          |
| 3                        | 3                               | 2.20                          | 13                       | 13                              | 9.84                          |
| 4                        | 8                               | 6.04                          | 14                       | 15                              | 11.02                         |
| 5                        | 6                               | 4.81                          | 15                       | 13                              | 9.60                          |
| 6                        | 6                               | 4.78                          | 16                       | 17                              | 12.41                         |
| 7                        | 8                               | 5.97                          | 17                       | 18                              | 13.59                         |
| 8                        | 4                               | 3.04                          | 18                       | 14                              | 10.54                         |
| 9                        | 8                               | 5.85                          | 19                       | 18                              | 13.22                         |
| 10                       | 9                               | 6.69                          | 20                       | 22                              | 16.47                         |
| 11                       | 8                               | 5.92                          | 21                       | 25                              | 18.91                         |

Table 3: Best fits to a Lorenz signal, where $t_{\text{best}}$ in symbols equals $t_{\text{best}}$ as a real divided by $t_{\text{return}} = 0.7511$. 
can predict $x_{T+t}$ for $t > (1 + \epsilon) \log_2 T/H$ for $\epsilon > 0$. In this section, we formalize this claim to some extent to bring out in outline what form the proof of such a claim might take.

As in Section 3, the sequence $x_0, \ldots, x_T$ is assumed to be generated from the state vectors of a dynamical system $X_{n+1} = f(X_n)$. Since our aim is to upper bound the extent of predictability of the sequence, we may, without loss of generality, assume the entire state vector to be observable. We assume that the map $f$ possesses a hyperbolic attractor $A$. We assume that $f$ is transitive on $A$. Within a hyperbolic attractor, periodic points are dense and therefore a hyperbolic attractor satisfies the Axiom-A conditions.

We define a predictor as a measurable function and write it as $P(X_0, \ldots, X_T) = (\tilde{X}_{T+1}, \tilde{X}_{T+2}, \ldots)$. The measurable function $P$ captures our notion of an algorithm which will take the $T$ successive state vectors $X_0, \ldots, X_T$ and will generate approximations $\tilde{X}_{T+s}$ to $X_{T+s}$ for $s = 1, 2, \ldots$. The algorithm is not required to output an approximation for every $s > 0$. We will assume that it outputs approximations for $s = 1, 2, \ldots, t_f$.

At this point, we have to decide when a prediction is termed as valid. One possibility seems to be to use Markov partitions. Markov partitions use the stable and unstable manifolds of the hyperbolic set $A$ to partition $A$ into finitely many pieces. Markov partitions of arbitrarily small diameters are guaranteed to exist. The key advantage of Markov partitions is that they facilitate symbolic encoding of the dynamics of $f$ on the hyperbolic attractor $A$. The way the encoding works is similar to the encoding of Lorenz trajectories using the symbols $A$ and $B$, which we explained in the previous section. However, defining the validity of predictions using Markov partitions seems problematic. Two points which are very close to each other can fall in different Markov partitions if they happen to lie on opposite sides of the boundary defining the partitions.

Thus we go back to using a tolerance to define a valid prediction. A prediction $\tilde{X}_{T+s}$ is deemed to be valid if $|X_{T+s} - \tilde{X}_{T+s}| \leq tol$ for some tolerance $tol$.

We require the prediction algorithm to output $\tilde{X}_{t+s}$ as a valid prediction for $s = 1, \ldots, t_f$. In other words, each prediction output by the prediction algorithm $P$ must be valid. Alternatively, we can allow the prediction algorithm to output anything it wants and define $t_f$ by counting only the valid predictions in the segment that immediately follows $t = T$. At this point, there seems to be little to choose between the two possibilities. So we adopt the more restrictive definition of a prediction algorithm.

Now our claim can be stated as follows: if $P$ is a valid prediction algorithm

$$\limsup_{T \to \infty} \frac{t_f}{\log_2 T} < \frac{1 + \epsilon}{H}$$

with probability 1 for any $\epsilon > 0$. The notion of entropy $H$ that we adopted in the previous section was metric entropy $h_\mu(f)$ relative to the physical measure $\mu$ on $A$. For a hyperbolic attractor, the physical measure is the SRB measure and it is guaranteed to exist. Thus we are assuming $X_0$ to be $\mu$-distributed, $X_1 = f(X_0)$, $X_2 = f(X_1)$, and so on. But it is the same prediction algorithm for any $f$ and any hyperbolic attractor $A$. It is not allowed to assume information about $f$. 

16
In order to explain why every prediction algorithm must satisfy the bound (4.1), we turn to another notion of entropy, namely topological entropy $h_{\text{top}}(f)$ \cite{[16]}. To begin with we have a metric $d$ on $\mathcal{A}$. We can define $d_n(x,y)$ to be the maximum of $d(f^i(x), f^i(y))$ over $i = 0, \ldots, n - 1$. If $N(\delta, n)$ is the number of $\delta$ balls required to cover $\mathcal{A}$ in the metric $d_n$, topological entropy is defined using the relation $N(\delta, n) \approx C 2^{nh_{\text{top}}}$ for small $\delta$. It is independent of the metric. In general, $h_\mu(f) \leq h_{\text{top}}(f)$ (see Theorem 4.5.3 of \cite{[16]}). With the assumptions we have made on $\mathcal{A}$ and $\mu$, $h_\mu = h_{\text{top}}$.

Suppose we are given the sequence $X_0, \ldots, X_T$. That is equivalent to assuming that we know the iterates $f, f^2, \ldots, f^n$ at $T - n + 1$ points on $\mathcal{A}$. For example, we know $f(X_1) = X_2$, $f^3(X_2) = X_5$, and so on. We are assuming $n$ to be of the order of $\log_2 T$. These $T - n + 1$ points on $\mathcal{A}$ at which $n$ iterates of $f$ are known may be assumed to be approximately $\mu$-distributed. To predict $n$ iterates of $X_T$ with tolerance $\text{tol} = \delta$ from that information, we require one of the points $X_0, \ldots, X_{T-n}$ to be within $\delta$ of $X_T$ in the $d_{n+1}$ metric. For such a thing to be possible, we require $T - n \geq C 2^{nh_{\text{top}}}$ or $n \leq \log_2 T/h_{\text{top}}$ asymptotically.

It may seem that one may extract some more information about $f^n(X_T)$ by clever interpolation of $f^n$ whose value is known at $X_0, \ldots, X_{T-n}$. It is true that clever interpolation can improve the accuracy dramatically if the function being interpolated is smooth. In this context, however, no such thing is possible even if $f$ is infinitely differentiable or real analytic. The key reason is that the exponential divergence of trajectories is enough to defeat any attempt at clever interpolation.

Perhaps this point will be clearer with an example. The map $x_{n+1} = f(x_n)$ with $f(x) = 4x(1-x)$ over the interval $[0,1]$ has topological entropy equal to 1. Suppose we want to predict $f^n(x_T)$. Given the shape of $f$, $f^n$ will have $2^{n-1}$ oscillations. By an oscillation we mean a monotonic increase in $f^n(x)$ from 0 to 1 and then a monotonic decrease to 0. If $T < C 2^{nh_{\text{top}}}/(1+\epsilon) = C 2^n/(1+\epsilon)$, it is clear that $T$ points will be too few to track all the oscillations of $f^n$. No interpolation scheme can make up for that kind of undersampling. The consideration here is somewhat analogous to that in the sampling theorem. According to the sampling theorem, to reconstruct a band-limited signal exactly, we must sample at least twice per wavelength.

As indicated earlier, the theoretical discussion in this section is restricted to maps. However, a new point comes up in relation to flows that is worth mentioning. Suppose we have a continuous signal $x(t)$ for $0 \leq t \leq T$ from a real analytic flow. Then $x(t)$ is analytic in a neighborhood of the real line. Thus in principle we may use the known stretch of the signal to predict it forever into the future using analytic condition. Analytic continuation is numerically unstable and often not feasible as an extrapolation strategy. Limitations to the applicability of analytic continuation become evident the moment we note that the continuous signal must be sampled at some finite rate and that it is incorrect to assume the entire signal to be available. A very similar point comes up in the context of the Wiener-Kolmogorov predictor. See Section 1.7 of \cite{[23]}.

A prediction algorithm $P$ is optimal if

$$\liminf_{T \to \infty} \frac{t_f}{\log_2 T} \geq \frac{1 - \epsilon}{H}$$

(4.2)
with probability 1 for any \( \epsilon > 0 \). Our view of optimality is tied to almost sure prediction and not to statistical predictability. The practical significance of almost sure convergence is accepted in information theory. See the discussion in [24] for an example.

5 The embedding predictor, related predictors, and their suboptimality

The Wiener-Kolmogorov predictors were derived for statistical time series which are well-characterized by their auto-correlation functions. They should not be used for the extrapolation of smooth curves. In this regard, Wiener wrote [23 0.71]: geometric facts must be predicted geometrically and analytical facts analytically, leaving only statistical facts to be predicted statistically.

There are two geometrical facts that are central to the prediction of chaotic signals. The first is recurrence and the second is the need to decompose close recurrences into stable and unstable components. Existing predictors have exploited recurrence but have not attempted to decompose close recurrences into stable and unstable components. As a result, they fall well short of being optimal in the sense of (4.2).

In this section, we discuss a few existing predictors of chaotic signals. Some of the ideas used by existing predictors, which we find to be deficient with respect to optimal prediction, may become more useful once a good method is found to decompose close recurrences into stable and unstable components. For example, polynomial interpolation has been suggested and used for limited improvement of the accuracy of predictions of chaotic time series. It is of little use in getting closer to optimality. However, if close recurrences are decomposed appropriately into stable and unstable components, polynomial interpolation may indeed be useful for improving the accuracy of the prediction of \( x(T + s) \), especially for \( s < \alpha \log_2 T/H \), where \( \alpha \) is a small fraction.

Phase space reconstruction using delay coordinates is used by all existing predictors. We term the most basic of these predictors as the embedding predictor \([5, 9] \). Given a signal \( x(t) \) for \( 0 \leq t \leq T \), the embedding predictor finds \( t^* \) to minimize (1.3), as we have already discussed. The key idea behind embedding predictors is to indirectly recover the location of the dynamical system in phase space at time \( t \) using the delay coordinates \((x(t), x(t - \tau), \ldots, x(t - (k - 1)\tau))\). Suppose the state vector of the dynamical system at time \( t = t_1 \) is \( X_1 \) and the state vector at time \( t = t_2 \) is \( X_2 \). It is quite possible that \( x(t_1) = x(t_2) \) even if \( X_1 \neq X_2 \) or that \( |x(t_1) - x(t_2)| \) is small even if \( X_1 \) is not close to \( X_2 \). However, the pattern of events preceding \( t = t_1 \) and \( t = t_2 \) as recorded using delay coordinates gives us better information to decide if \( X_1 \) and \( X_2 \) are close to each other or not.

Although the choice of the delay parameter \( \tau \) and the embedding dimension \( k \) have been discussed extensively, it is difficult to make definite statements about what the best choices are. One approach is to use mutual information—see [1]. In this approach it is assumed that \( \tau \) should not be too small because nearby values are well-correlated and not too large because distant points on the signal are very weakly correlated. Mutual information is used to find some kind of a compromise. Regarding the embedding
Figure 3: Suboptimal predictions (in thick red) using the embedding predictor to a Lorenz signal (in thin black). (a) $T = 2^{14}$ symbols. (b) $T = 2^{21}$ symbols.
5 THE EMBEDDING PREDICTOR, RELATED PREDICTORS, AND THEIR SSUBOPTIMALITY
dimension \( k \), it is stated that it should be at least as large as the dimension of the underlying chaotic set. We comment about time series embedding using delay coordinates in Section 10.

Figure 3 shows suboptimal predictions of Lorenz signals using the embedding predictor and \( \tau = 0.03, k = 5 \). Our choice of the delay parameter at \( \tau = 0.03 \) is much smaller than what the mutual information criterion would imply. The mutual information criterion would imply a \( \tau \) that is large enough to span a few oscillations of the signal. It is difficult to see what advantage using information from such distant points may have with regard to prediction, where the game is to exploit local information optimally. Indeed, use of a larger delay parameter gives no improvement at all. Some of the extent discussion about choosing the delay parameter appears to be based on a desire to obtain good plots and not good predictions.

For a study of the effect of the delay parameter on the quality of prediction, see Figure 22 of Casdagli et al. [6]. For the Ikeda map, the optimal delay for prediction is found to be the smallest delay possible. In Figure 22 of that paper, an attempt is made to predict only one iteration using a history that is equal to \( 10^4 \) iterates in length. The entropy theorem indicates that more than 20 iterates could be predictable using a history of that length. Here the advantage of defining optimal prediction as in (4.2), which we mentioned earlier in the introduction, becomes evident in a more concrete way. If we attempt to predict only one iterate, different prediction methods will differ in terms of accuracy, but the difference will be quite delicate. Even for predicting a single iterate optimally, it is important to resolve close recurrences into stable and unstable components. However, the gain in accuracy to be obtained by resolving close recurrences in that manner is not easily noticed. In contrast, the optimality criterion (4.2) which emphasizes the length of the fit into the future, exposes the central deficiency of existing predictors in a way that is quite easy to see.

If we compare Figure 3 with Figure 2, it is abundantly clear that the embedding predictor does not extract the information in the history of the signal in an optimal manner. The embedding predictor gives a closer fit in the immediate past of \( t = T \), but that is precisely why it does not do the best job of predicting the future. Still from Figure 3 we see that the fit into the future is much better than the fit into the past. Does the embedding method have a bias to the future after all? The answer is no. The embedding method treats the past and the future equally. There is nothing in it to say that it is attempting to predict the future rather than fit the past. The better fit into the future we see in the figure is a consequence of the Lyapunov exponents of the Lorenz attractor. The lone negative exponent of the Lorenz attractor is \(-14.5\) (using natural logarithms) and is much larger in magnitude than the lone positive exponent, which is \(0.905\). Therefore if we pick two points close to each other on the Lorenz attractor, the corresponding trajectories will typically diverge faster in the past.

The \( t_{\text{embed}} \) column of Table 4 is obtained as follows. The metric (1.3) is used to pick \( t^* \) so that the distance between the delay coordinates at \( t = t^* \) and \( t = T \) is the smallest. The length of the fit into the future is given by \( t_{\text{embed}} \): \(|x(t^* + s) - x(T + s)| \leq tol \) for \( 0 \leq s \leq t_{\text{embed}} \) but not for \( 0 \leq s \leq t \) with \( t > t_{\text{embed}} \). Comparison of \( t_{\text{embed}} \) in Table 4
with $t_{\text{best}}$ in Table 3 shows that the embedding predictor does not approach optimality.

In the rest of this section, we consider a number of ideas for improving the basic embedding predictor. All these ideas have merits. However, to be fully effective, they need to take into account an essential aspect of chaotic signals, which is their tendency to separate or come together depending upon the relative sizes of the stable and unstable components.

The first idea we mention is from the paper by Farmer and Sidorowich [9]. To predict $x(T + s)$ the basic embedding predictor picks a single $t^* \in [k\tau, T - s]$ using the metric (1.3). Instead, a predictor may pick $p$ different instants $t_{i}^*, ..., t_{p}^*$ where the delay coordinates are the $p$ closest to the delay coordinates at $t = T$. Assuming $p \geq k$, the prediction of $x(T + s)$ is generated as a linear combination of the delay coordinates at $t = t_{i}^*$ by fitting $x(t_{i}^* + s)$ as a linear combination of the delay coordinates at $t = t_{i}^*$, for $i = 1, ..., p$, using linear least squares.

Let us first understand the merit of this idea. Suppose we are looking at a Lorenz signal and we fix $s = 1$, which means we are trying to predict the signal at a point that is somewhat more than one return time ($t_{\text{return}} = 0.7511$) from $t = T$. For sufficiently large $T$, the signal will have delay coordinates at $t = t_{i}^*$ close to that at $t = T$ for each of the $p$ values of $i$. More importantly, they will be sufficiently close that none of the $p$ segments $x(t), t_{i}^* \leq t \leq t_{i}^* + s$, will diverge from each other for $i = 1, ..., p$. Therefore extrapolation using least squares will improve the order of accuracy (see Figure 2 of [9]).

The situation is quite different if we take $s = \alpha \log_{2} T/H$, with say $\alpha = 0.75$. In this case, we want to predict an instant that gets farther out in time as $T$ increases. In this situation the $p$ segments $x(t), t_{i}^* \leq t \leq t_{i}^* + s$, with $i = 1, ..., p$ will diverge from each other with high probability ruining any attempt to extrapolate using linear least squares. One may attempt to patch the situation by trying to classify the $p$ segments into clusters that stay close to each other and then picking one of the clusters to extrapolate

| $\log_{2} T$ (in symbols) | $t_{\text{embed}}$ (in symbols) | $\log_{2} T$ (in symbols) | $t_{\text{embed}}$ (in symbols) |
|--------------------------|---------------------------------|--------------------------|---------------------------------|
| 2                        | 0                               | 12                       | 8                               |
| 3                        | 0                               | 13                       | 10                              |
| 4                        | 1                               | 14                       | 7                               |
| 5                        | 5                               | 15                       | 9                               |
| 6                        | 6                               | 16                       | 7                               |
| 7                        | 4                               | 17                       | 9                               |
| 8                        | 4                               | 18                       | 10                              |
| 9                        | 6                               | 19                       | 9                               |
| 10                       | 2                               | 20                       | 8                               |
| 11                       | 3                               | 21                       | 9                               |

Table 4: Length of suboptimal predictions of a Lorenz signal ($t_{\text{embed}}$) using the embedding predictor.
from $t = T$ to $t = T + s$. But to do so would be to get back to our point that one has to decompose the distance between segments of the signal into stable and unstable components for optimal prediction.

Even with $s = 1$, in which case extrapolation using least squares improves the accuracy of the basic embedding predictor, there are advantages to decomposing the distance between segments of the signal into stable and unstable components. Such a decomposition will allow us to weight the different segments from the past and wring all the information out of the signal. Conversely, ideas such as extrapolating using linear least squares may be useful once the basic issue of resolving the distance between segments into stable and unstable components is addressed.

Other ideas for improving the basic embedding predictor are to use higher order polynomials for extrapolation [9], to trap the delay coordinates at $t = T$ within a simplex in reconstructed phase space [20], or to weight close recurrences using the closeness of the approach [11]. The merits and demerits of these ideas are as in the discussion above and nothing more needs to be said. Another idea is to extrapolate from $t = T$ to $t = T + 1$ using the embedding predictor possibly with enhancements and then iterate the extrapolation from $t = T$ to $t = T + 1$ a total of $s$ times to extrapolate from $t = T$ to $t = T + s$. The merit of this idea is to bring in new information from the signal to evaluate intermediate points such as $t = T + 1$ and $t = T + 2$. However, the embedding predictor continues to be suboptimal even with this enhancement. The problem is that a single step of extrapolation will throw away all the information about stable and unstable manifolds in the vicinity of $t = T$. The way the stable and unstable components of the distance between two segments of the signal must be taken into account depends upon how far into the future we want to extrapolate, as will become clear in the next section.

6 Character of an optimal predictor

In this section, we give a sense of how an optimal predictor might work. Although a general purpose optimal predictor has not yet been derived, it is possible to give a sense of what such a predictor should do.

Suppose $c$ is a fixed point of the map $f$. The iterates at $c$ will obviously look like

$$c, c, c, \ldots$$

Suppose we pick a point $X_0$ within a distance $\epsilon$ of $c$ and look at the sequence

$$X_0, f(X_0), f^2(X_0), \ldots$$

When is the latter sequence closest to the former sequence? The answer is they are closest when $X_0$ lies on the stable manifold of $c$. If it lies on the unstable manifold of $c$, on the other hand, the latter sequence will quickly diverge from the former. Here we already see the basic ingredient for optimal prediction. For a good match between the sequences, it is not enough to pick $X_0$ close to $c$ but we have to pick $X_0$ to be on or close to the stable manifold of $c$. An optimal predictor has to implement this idea using time series data and nothing more.
In general, it is impossible to pick a point that is exactly on the stable manifold. Therefore, we expand upon what it means to pick a point that is close to the stable manifold. Let $c$ be a point on the hyperbolic attractor. Let us suppose that $x$ is close enough to $c$ and that we may write $x$ as

$$x = c + \sum_{i=1}^{u} a_i v_i^+(c) + \sum_{i=1}^{s} b_i v_i^-(c).$$  

(6.1)

Here $v_i^+(x)$ are unit vectors in the tangent space at $x$ corresponding to positive Lyapunov exponents and the $v_i^-(x)$ are unit vectors corresponding to negative Lyapunov exponents. For simplicity, we assume the Lyapunov exponents to be distinct with $u$ positive exponents and $s$ negative exponents. Let $\lambda_i^+$ be the characteristic multiplier corresponding to $v_i^+$ and similarly let $\lambda_i^-$ correspond to $v_i^-$ (if $l$ is a Lyapunov exponent defined using natural logarithms, $\exp(l)$ is the corresponding characteristic multiplier).

We have

$$f^n(x) \approx c_n + \sum_{i=1}^{u} a_i (\lambda_i^+)^n v_i^+(c_n) + \sum_{i=1}^{s} b_i (\lambda_i^-)^n v_i^-(c_n) \quad \text{where} \quad c_n = f^n(c).$$  

(6.2)

Here we have assumed that the expansion along the directions $v_i^+$ and $v_i^-$ is by the same factor with each iteration. With that assumption, it is easier to bring out the essential aspects of the heuristic argument we are developing here. Note that $|\lambda_i^+| > 1$ and $|\lambda_i^-| < 1$.

To eliminate some linear algebra from the discussion, we will assume that $v_i^+(x)$, $1 \leq i \leq u$, and $v_i^-(x)$, $1 \leq i \leq s$, form an orthonormal basis for the tangent space at each point $x$ on the hyperbolic attractor. For the related concepts of adapted metric and adapted coordinates, see [16].

Suppose (as usual) that the points in the available trajectory are $x_0, \ldots, x_T$ with $x_T = c$. To predict the sequence $f(c), f^2(c), \ldots, f^k(c)$, with $k \approx \log_2 T/H$, we will look at points from the sequence $x_0, \ldots, x_{T-k}$ that are close enough to $x_T$ and $c$ can be represented in the form (6.1). Here we will examine what kind of points $x$ are available in the sequence and which ones will be useful predictors.

Let us try to find an $x$ of the form (6.1) in the available history with $a_i = A_i \delta$ for $1 \leq i \leq u$ and $b_i = B_i \delta$ for $1 \leq i \leq s$ with $A_i$ and $B_i$ fixed to determine the shape of the box around $c$ and with as small a $\delta$ as possible. Kac’s theorem (Theorem 2.2) suggests that we may find a point in the available history in a box around $c$ if the volume of the box is $1/T$ or more. Thus in a box of shape determined by $A_i$ and $B_i$, the smallest $\delta$ that leaves the box large enough to be likely to include a point from the available history is given by $A_1 \ldots A_u B_1 \ldots B_s \delta^{i+u} \approx 1/T$. In fact, we will allow the stable components $b_i$ to be as large as the tolerance allows. In that case, the box has dimensions $a_i = A_i \delta$ and $b_i = O(1)$. The smallest delta should then satisfy

$$A_1 \ldots A_u \delta^u \approx \frac{C}{T}$$  

(6.3)
for some constant $C$.

We may now try to choose the shape of the box to allow $f^n(x)$ to stay close to $f^n(c)$ for $n = 1, \ldots, k$. If we estimate the distance between $f^n(x)$ and $f^n(c)$ using \((6.2)\), the distance comes out as follows:

$$
\| f^n(x) - f^n(c) \| \approx \sqrt{\sum_{i=1}^{u} a_i^2 \left( \lambda_i^+ \right)^{2n}}.
$$

\((6.4)\)

Here we have neglected the $\mu_i^-$ components because $\mid \mu_i^- \mid < 1$ and these stable components diminish rapidly with $n$. As long as the stable components are less than a tolerance, we do not need to worry about them. Given the constraint on how small the box can get, the best shape is obtained by taking $A_i = 1/\left( \lambda_i^+ \right)^n$. The value of $\delta$ implied by \((6.4)\) is

$$
\delta^n \approx C \left( \prod \lambda_i^+ \right)^n T.
$$

\((6.5)\)

and the minimum possible value of $\| f^n(x) - f^n(c) \|$ is approximately $\delta \sqrt{u}$.

From this heuristic calculation, we learn two things. If we want to pick an $x$ from the available history to minimize $\| f^n(x) - f^n(c) \|$ it is not enough to simply pick an $x$ from the history that is as close to $c$ as possible. We have to balance the sizes of the unstable components $a_i$ carefully. The stable components $b_i$ can be as large as the tolerance of the problem allows, which means that the best $x$ for predicting $f^n(c)$ may not be particularly close to $c$.

For valid prediction of $f^n(c)$, we require $\| f^n(x) - f^n(c) \| \approx \delta \sqrt{u} \leq tol$. If we use expression \((6.5)\) for $\delta$, we get

$$
n \leq \frac{\log_2 T + u \log_2 tol - (u/2) \log_2 u - \log_2 C}{\sum \log_2 \lambda_i^+}.
$$

\((6.6)\)

For a hyperbolic attractor, metric entropy is equal to $\sum \log_2 \lambda_i^+$. From this calculation, we understand why the metric entropy shows up the way it does in the entropy theorem.

In the argument leading up to \((6.6)\), we assumed $x$ and $c$ to be points on the hyperbolic attractor. A predictor which predicts $x_{T+n}$ for $n$ that approaches the upper bound in \((6.6)\) or is optimal in the sense of \((4.2)\) has to calculate the $a_i$ in \((6.1)\) using time series data alone.

Given a Lorenz signal, suppose we want to assess if $t = t^*$ will give a long fit to the segment following $x(T)$, with the length of fit defined as in \((1.2)\). If we knew the points $X(t^*)$ and $X(T)$ in the three-dimensional phase space of the Lorenz flow, as well as the decomposition $X(t^*) - X(T) = s + f + u$—where $s$ is along the stable direction at $X(T)$, $f$ is along the flow at $X(T)$, and $u$ is along the unstable direction at the same point—the assessment would be easy to make. As long as the components $f$ and $s$ are below the tolerance, we want the minimum $\| u \|$ possible for the longest fit.

The embedding method attempts to estimate the distance between $X(t^*)$ and $X(T)$ using delay coordinates and the formula \((1.3)\). It does not even attempt to resolve the close recurrences into $s, f,$ and $u$ components as an optimal predictor should.
7 The Wiener-Kolmogorov predictor

Let \( x(t) \) be a signal that arises from a continuous dynamical system \( \dot{X} = f(X) \). From the discussion in the previous section, the following picture emerges with regard to the prediction of \( x(T + s) \) for \( 0 < s \leq t_f \) given the history of the signal, which is \( x(t) \) for \( 0 \leq t \leq T \). At some point \( t^* \) with \( 0 \leq t^* \leq T - t_f \), the underlying dynamical system will have the state vector \( X(t^*) \). For optimal prediction, we have to use the available signal to form an estimate of the state vectors \( X(T) \) and \( X(t^*) \), as well as the decomposition of \( X(t^*) - X(T) \) into stable and unstable components relative to the splitting of the tangent space at \( X(T) \). In Section 9, we give some idea of how one might go about doing such a thing. If two segments of the signal are similar to each other, decomposing the distance between the two segments into stable and unstable components is somewhat similar to spectral analysis. Indeed the Padé predictor described in Section 9 is similar to the Wiener-Kolmogorov predictor in a few respects.

In this section, we briefly describe the Wiener-Kolmogorov predictor to help bring out those points of similarity and to draw attention to the case of certain zero entropy signals. Consider the dynamical system

\[
\frac{d\theta_1}{dt} = \omega_1, \quad \frac{d\theta_2}{dt} = \omega_2, \ldots, \quad \frac{d\theta_d}{dt} = \omega_d
\] (7.1)

where the \( \theta_i \) are angular variables and the \( \omega_i \) are the corresponding frequencies. Suppose the observed signal is \( x(t) = \sum_{i=1}^{d} \cos(\theta_i) \). Predicting such a signal is chiefly a matter of spectral analysis and the Wiener-Kolmogorov filter can handle such signals with ease. If an optimal predictor for chaotic signals is to be well-behaved in the limit of vanishing entropy, it too should be able to handle such signals.

The entropy theorem (Theorem 2.3) states that \( t_{\text{best}}/\log_2 T \) diverges as \( T \to \infty \) for zero entropy signals. Therefore \( t_{\text{best}} \) should increase super-logarithmically. For a signal derived from (7.1), a calculation using Kac’s theorem suggests that \( t_{\text{best}} \) will be proportional to \( T^{1/d} \) (assuming no rational relationship between the \( \omega_i \)). This is because each side of a \( d \)-dimensional cube must be of length \( T^{-1/d} \) for the volume of the cube to be \( 1/T \).

This allows us to comment on an aspect of the entropy theorem, which is the independence of the bound on \( t_{\text{best}}/\log_2 T \) from the dimension. The entropy theorem leads us to think that the recurrences of a signal obtained from a hyperbolic attractor of large dimension and a hyperbolic attractor of small dimension will be of the same quality if the two attractors have the same metric entropy. Can such a thing really be true? It has to be true in the limit \( T \to \infty \) as asserted by the theorem, but the problems associated with large dimensionality will certainly be an issue for practical values of \( T \).

For zero entropy signals of the type considered above, having recurrences of length \( T^{1/d} \) may seem much better than recurrences of length \( \log_2 T \) as guaranteed by the entropy theorem for signals of entropy 1. However, \( T \) would have to be very large for \( T^{1/d} \) to be greater than \( \log_2 T \) for even \( d = 20 \). Since the Wiener-Kolmogorov predictor is based purely on spectral analysis and not on tracking recurrences, it can handle signals of the type \( x(t) = \sum_{i=1}^{d} \cos(\theta_i) \) with ease for \( d = 20 \).
In our brief account of the Wiener-Kolmogorov predictors, we try to bring out the role of Toeplitz operators and matrices because Toeplitz matrices will appear later in Section 9. Suppose \( f(t), -\infty < t < \infty \), is a real-valued function whose auto-correlation \( \phi(\tau) \) is given by (1.1). The Wiener-Kolmogorov predictor uses

\[
\int_0^\infty f(T-s) dK(s)
\]

to predict \( f(T+\alpha) \) using \( f(t) \) for \( -\infty < t < T \) and a suitable function \( K(s) \) of finite total variation. The function \( K(s) \) is found to minimize the least squares error which is

\[
\lim_{T \to \infty} \frac{1}{2T} \int_{-T}^T \left| f(t+\alpha) - \int_0^\infty f(t-s) dK(s) \right|^2 dt.
\]

The \( K(s) \) which minimizes the least square error should satisfy the Wiener-Hopf equation

\[
\phi(\alpha + t) = \int_0^\infty \phi(t-s) dK(s).
\]

By choosing appropriate function spaces, the right hand side can be interpreted as the application of a Toeplitz operator to \( K \).

To solve for \( K \), let \( \Phi(\omega) \) be the Fourier transform of \( \phi(t) \) and assume \( \int_{-\infty}^{\infty} |\log |\Phi(\omega)||/(1+\omega^2) d\omega \) to be finite. Then

\[
\Psi(u + iv) = \exp \left( \frac{1}{2\pi} \int_{-\infty}^{\infty} \log |\Phi(\omega)| \frac{d\omega}{-i(\omega - (u + iv))} \right)
\]

is free of singularities in the lower half of the complex plane and \( \Phi(u) = |\Psi(u)|^2 \). The Fourier transform \( k(\omega) \) of \( K(s) \) is given by

\[
k(\omega) = \frac{1}{2\pi \Psi(\omega)} \int_0^\infty \exp(-iwt) dt \int_{-\infty}^{\infty} \Psi(u) \exp(iu(t+\alpha)) du.
\]

For more details, see [23].

In practice, all signals must be sampled at a finite rate and no filter can have infinite memory. The theory can be modified to handle prediction of series \( a_n, n \in \mathbb{Z} \), using finite memory. In this case, the auto-correlation is

\[
R(k) = \lim_{N \to \infty} \frac{1}{2N+1} \sum_{l=-N}^{N} a_l a_{l-k}.
\]

With this definition, \( R(-k) = R(k) \). Let \( r_n = R(n)/R(0) \). The prediction error

\[
a_{k+s} = \sum_{n=0}^{M} A_n a_{k-n}
\]

26
is minimized in the least squares sense if
\[ \sum_{n=0}^{M} r_{k-n} A_n = r_{k+s} \]
for \( k = 0, \ldots, M \). The matrix system that must be solved to calculate \( A_n \) is symmetric and Toeplitz.

We have described the Wiener-Kolmogorov predictor for a single noiseless signal. For noisy signals, multiple noiseless signals, and multiple noisy signals, see [23].

8 Toral automorphisms

Let \( A \) be a \( d \times d \) matrix with integer entries and \( \det A = \pm 1 \). The map \( X_{n+1} = AX_n \mod 1 \) is a hyperbolic toral automorphism if no eigenvalue of \( A \) has unit modulus. Here \( X_n \) is a vector with \( d \) entries each of which is assumed to be in the interval \([0, 1)\). Each entry of the matrix vector product \( AX_n \) is taken modulo 1 in the interval \([0, 1)\) to get \( X_{n+1} \). The space \([0, 1)^d \) is used as the coordinate space of the torus \( \mathbb{T}^d \).

The class of hyperbolic toral automorphisms is a basic example in theoretical dynamics [16]. Such automorphisms are topologically transitive on the torus and possess Markov partitions of arbitrarily small diameter. The physical measure is the Lebesgue measure and the entropy is positive.

In the next section, we consider the prediction of the signal \( x_0, \ldots, x_T \), where \( x_n \) is the first entry of \( X_n \) for each \( n \), \( X_0 \) is uniformly distributed on \( \mathbb{T}^d \), and \( X_{n+1} = AX_n \mod 1 \) for \( n \geq 0 \). The first toral automorphism that is considered is
\[ X_{n+1} = AX_n \mod 1, \quad A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}. \] (8.1)

This matrix \( A \) has eigenvalues 2.61803 and 0.381966 and its entropy is \( \log_2 2.61803 = 1.3885 \). The second toral automorphism that is considered is
\[ X_{n+1} = AX_n \mod 1, \quad A = \begin{pmatrix} 0 & -1 & 0 \\ 1 & -2 & 1 \\ 2 & -3 & 3 \end{pmatrix}. \] (8.2)

This matrix \( A \) has eigenvalues 2.1479 and \(-0.57395 \pm i0.368989\). In both instances, \( \det A = 1 \).

The optimal prediction scheme given in the next section is specialized to hyperbolic toral automorphisms. The purpose of that prediction scheme is to show how the distance between close recurrences can be decomposed into stable and unstable components. Restricting ourselves to toral automorphisms has the advantage that the tangent space splits into stable and unstable directions in exactly the same way at every point on the torus (the vectors \( v_i^+ \) and \( v_i^- \) in (6.1) do not depend upon \( c \)). In the next section, we show how signals obtained from hyperbolic toral automorphisms can be predicted optimally in the sense defined by (4.2).
However, it is important to note that restricting ourselves to the class of hyperbolic toral automorphisms means that some oddities occur that would not occur with a general purpose optimal predictor. Hyperbolic toral automorphism of dimension $d$ are defined using finitely many parameters each of which is an integer (entries of the matrix $A$). One may exploit that fact and tweak the predictor in the next section to reconstruct the toral automorphism exactly. We do not overly specialize the prediction scheme in that way. The purpose of the prediction scheme is to show what kind of considerations may arise in the derivation of a general purpose predictor and the exact reconstruction of the toral automorphism from time series data is irrelevant in that regard.

Hyperbolic toral automorphisms are maps. Although the theory has fewer technicalities for maps, continuous signals may be better targets for general purpose predictors. Typically maps arise as suspensions of flows or as maps between Poincaré sections of flows. A lot of information contained in the path taken by the flow is lost when a map is derived from a flow.

9 Optimal prediction of toral automorphisms

We begin by considering the so-called exponential extrapolation problem. Suppose a sequence is defined by

$$ s_n = \sum_{k=1}^{d} c_k \lambda_k^n \quad n = 0, 1, \ldots $$

The problem is to find $s_{2d}$, $s_{2d+1}$, and so on given $s_0, \ldots, s_{2d-1}$. Since the sequence is defined by $d$ parameters $c_k$ and $d$ parameters $\lambda_k$, it is reasonable to expect that the first $2d$ numbers of the sequence may determine the rest of the sequence. The exponential extrapolation problem is to determine the rest of the sequence. It was solved by Prony late in the 18th century (see [13] for a discussion of Prony’s method). We present a solution based on Padé approximants. Our presentation could be new. Padé approximants generalize naturally to vector Padé approximants, which may turn out to be useful in deriving a general purpose predictor. For an introduction to Padé approximation, see [2].

Define $f(z) = \sum_{k=0}^{\infty} s_k z^k$. Using (9.1), we get

$$ f(z) = \sum_{k=1}^{d} \frac{c_k}{1 - \lambda_k z_k} = \frac{a_0 + a_1 z + \cdots + a_{d-1} z^{d-1}}{1 + b_1 z + \cdots + b_d z^d}. $$

The right hand side is the $(d-1, d)$ Padé approximant of $f(z)$. Determining the $b_i$ is the key to exponential extrapolation. We have

$$ a_0 + \cdots + a_{d-1} z^{d-1} = \left(1 + b_1 z + \cdots + b_d z^d\right) \sum_{k=0}^{\infty} s_k z^k $$

$$ = \sum_{k=0}^{\infty} z^k \left(s_k + \sum_{j=1}^{\min(d, k)} s_{k-j} b_j \right). $$
Equating coefficients of $s^k$ for $k = d, \ldots, 2d - 1$, we get the $d$ equations

$$\sum_{j=1}^{d} s_{k-j}b_j = -s_k. \quad (9.2)$$

This Toeplitz system must be solved to determine $b_j$. Its solvability is a necessary condition for exponential extrapolation. Once the $b_j$ are determined, $s_{2d}, s_{2d+1}$, and so on.

The analogy of this process to the Wiener-Kolmogorov predictor is unmistakable. In both cases, a Toeplitz system must be solved. Once the Toeplitz system is solved, new numbers in the sequence are obtained as fixed linear combinations of preceding numbers in the sequence. Indeed, it is quite possible that there may be a way to view the Wiener-Kolmogorov predictors as variations or extensions of Prony’s method as presented here.

Let $x_0, \ldots, x_T$ be a signal obtained from a hyperbolic toral automorphism as explained in the previous section. Suppose we want to compare the segment

$$x_{t^*-2d+1}, \ldots, x_{t^*-1}, x_{t^*}$$

with the segment

$$x_{T-2d+1}, \ldots, x_{T-1}, x_T.$$  

We first form the differences $\Delta x_i = x_{t^*-2d+1+i} - x_{T-2d+1+i}$ for $i = 0, \ldots, 2d - 1$. Our intention is to extrapolate the $\Delta x_i$ sequence to figure out how well $x_{t^*+s}$ will predict
$x_{T+s}$. Since the toral automorphisms are carried out modulo 1, we begin by making the following modification to the $\Delta s_i$ sequence. For each $i$ with $0 \leq i \leq 2d - 1$, if $\Delta x_i > 1/2$, we replace $\Delta x_i$ by $\Delta x_i - 1$. On the other hand, if $\Delta x_i \leq -1/2$, we replace $\Delta x_i$ by $\Delta x_i + 1$. After these operations, we will have $|\Delta x_i| \leq 1/2$ for $i = 0, \ldots, 2d - 1$.

If the point on the torus $\mathbb{T}^d$ that corresponds to $x_n$ is $X_n$, we have $X_{n+1} - X_{m+1} = A(X_n - X_m) \mod 1$. Therefore if $X_{T-2d-1} - X_{T-2d-1}$ is small enough, the sequence $\Delta x_i$, $i = 0, \ldots, 2d - 1$, can be written as a linear combination of exponentials like the $s_i$ sequence in $[0,1]$. The $\lambda_i$ will be the eigenvalues of $A$. We use a tolerance to check if the $\Delta x_i$ are small enough to permit sensible exponential extrapolation.

Using exponential extrapolation, we compute $\Delta x_{2d}$, $\Delta x_{2d+1}$, and so on, and find the maximum $n$ such that each of the numbers

$$|\Delta x_{2d}|, \ldots, |\Delta x_{2d+n-1}|$$

is less than $tol$. For the computations reported in this section, $tol = 0.1$. The $n$ found in this way is the expected length of fit. The $t^*$ which gives the maximum expected length of fit is chosen. The sequences $x_{t+1}, x_{t+2}, \ldots$ and $x_{T+1}, x_{T+2}, \ldots$ are compared to determine the actual length of fit, which is denoted by $t_{pade}$.

In Table 3, we list $t_{best}$ (the best fit from the past defined as in $[1,2]$), $t_{embed}$, and $t_{pade}$. For the embedding predictor, we took $2d$ to be the embedding dimension. By going down the table, we can easily detect that the entropy $H$ is greater than 1. It is evident that the embedding predictor falls well short of being optimal, while the Padé predictor which resolves the distance between segments of the signal into stable and unstable components, approaches optimality.

Both the embedding predictor and the Padé predictor implicitly assume the dimension of the torus to be known, which is not true if we are given the signal $x_0, \ldots, x_T$ explicitly and nothing more. The dimension can be calculated using one of the methods mentioned in the next section or it can be found by trial and error by checking the effectiveness of the Padé predictor. Such details are not terribly relevant here as our purpose is simply to show that the distance between segments of a signal can be decomposed into components in a restricted setting and that such a decomposition leads to optimal prediction.

From Figure 4, we see that the best fit from the past does not agree too well with the signal at $T - 1, T - 2$, and so on. However, it suddenly slams into the signal starting at $T$ and closely tracts the signal for more than 12 iterates. The embedding predictor on the other hand does too good a job of fitting the past, but tracks only 5 iterates from $T$ onwards. The Padé predictor produces a match that requires a few iterates in the past to be close enough for exponential extrapolation. Except for that, it reproduces the behavior of the best fit where the signal segment that is chosen from the history of the signal slams into the signal at $t = T$ and then tracks it for a number of iterates.

Table 6 and Figure 5 refer to the toral automorphism defined by $[8,2,1]$. By going down Table 6 and comparing it with Table 5, we notice that the automorphism of $\mathbb{T}^3$ has lower entropy than the automorphism of $\mathbb{T}^2$. The tendency of the embedding predictor to fit into the past is very pronounced in the middle plot of Figure 5.
Figure 4: In each of the three plots, the black dots are part of a signal obtained from the iterates of the automorphism (8.1) of the two dimensional torus $\mathbb{T}^2$. Here $T = 2^{21}$. The bigger red dots are: (a) the best fit from the past; (b) suboptimal prediction using the embedding method; (c) optimal prediction using the Padé method.

Table 6: Length of best fit from the past, suboptimal prediction using the method of embedding, and optimal Padé prediction of a signal obtained from the automorphism (8.2) of the three dimensional torus $\mathbb{T}^3$. 

| $\log_2 T$ | $t_{\text{best}}$ | $t_{\text{embed}}$ | $t_{\text{pade}}$ | $\log_2 T$ | $t_{\text{best}}$ | $t_{\text{embed}}$ | $t_{\text{pade}}$ |
|-----------|-------------------|-------------------|-------------------|-----------|-------------------|-------------------|-------------------|
| 5         | 3                 | 0                 | 0                 | 17        | 12                | 0                 | 8                 |
| 6         | 5                 | 0                 | 2                 | 18        | 14                | 0                 | 13                |
| 7         | 1                 | 1                 | 0                 | 19        | 15                | 0                 | 11                |
| 8         | 4                 | 1                 | 1                 | 20        | 15                | 7                 | 15                |
| 9         | 6                 | 0                 | 2                 | 21        | 15                | 0                 | 12                |
| 10        | 4                 | 0                 | 3                 | 22        | 18                | 0                 | 14                |
| 11        | 8                 | 2                 | 2                 | 23        | 17                | 1                 | 16                |
| 12        | 11                | 0                 | 4                 | 24        | 18                | 9                 | 18                |
| 13        | 10                | 2                 | 4                 | 25        | 20                | 8                 | 19                |
| 14        | 8                 | 0                 | 6                 | 26        | 20                | 6                 | 18                |
| 15        | 11                | 2                 | 8                 | 27        | 21                | 6                 | 19                |
| 16        | 13                | 6                 | 9                 | 28        | 23                | 7                 | 21                |
Figure 5: In each of the three plots, the black dots are part of a signal obtained from the iterates of the automorphism \((8.2)\) of the three dimensional torus \(T^3\). Here \(T = 2^{28}\). The bigger red dots are: (a) the best fit from the past; (b) suboptimal prediction using the embedding method; (c) optimal prediction using the Padé method.

The figures and tables of this section give a good sense of how much is lost when a predictor fails to account for the unstable components of the distance between segments of the signal. They also suggest that a predictor which subjects the signal to more delicate analysis should be able to approach optimality.

10 Time series embedding and prediction

Current predictors of chaotic signals use delay coordinates. It has even been stated that “for prediction problems past based coordinates are unavoidable” [6]. While it is a tautology to say that a predictor must be based on information from the past, we are not convinced that delay coordinates are the best or the only way to extract information from the history of a signal.

The idea behind delay coordinates is mainly topological in nature. Consider a complete graph \(K_n\) with \(n\) vertices and with \(n \geq 5\). If such a graph is drawn on the plane, two edges must intersect as a consequence of the Jordan curve theorem. One the other hand, if each vertex is assigned to points in \(\mathbb{R}^3\) that are “random” or in general position, no two edges will intersect. Two typical lines in \(\mathbb{R}^3\) do not intersect. Any topological manifold of dimension \(m\) can be embedded in \(\mathbb{R}^{2m+1}\) using points in general position and partitions of unity [14].

The theoretical justification of attractor reconstruction using delay coordinates is due to Takens [21]. The key idea is that delay coordinates generically give points in
general position. At several points the proof of the embedding theorem of Takens is similar to the proof of the classical embedding theorem by Hurewicz [14]. Fixed points and short periodic orbits require special treatment.

Delay coordinates are used by methods that infer attractor dimension from time series data [8, 11, 12]. Here we note that dimension is a topological quantity.

With regard to prediction of signals or time series, purely topological information will not be sufficient. The embedding predictors attempt to estimate the distance between points in phase space using delay coordinates. A more systematic study of metric information may be useful.

Suppose that \( x(t), t \geq 0, \) is a chaotic signal. Let the points in phase space that correspond to \( t = t_1, t = t_2, \) and \( t = t_3 \) be \( A, B, \) and \( C, \) respectively. Suppose that the signal is known for \( 0 \leq t \leq T \) and that \( t_1, t_2, t_3 \) are all less than \( T. \) We ask the following question: how accurately can the angles of the triangle \( ABC \) be estimated using the time series data? This is an example of the type of question which tries to get at metric information rather than the topological kind.

11 Conclusion

The central point of this paper is that a good predictor of chaotic signals must not simply try to find a pattern of events that is as close as possible to the pattern of events leading up to the current time. The distance between the two patterns of events must be resolved into stable and unstable components. The magnitudes of the unstable components must be small and delicately balanced for optimal prediction. The stable components on the other hand are typically as large as the tolerance for correct prediction permits.

This conclusion has a counter-intuitive consequence. Because the stable components are typically not small, the known pattern of events which is best suited for predicting the current pattern of events will not resemble the current pattern particularly closely.

We have made this point using theoretical arguments and examples whose simplicity made them suitable for illustration. We certainly feel that this point is of very general applicability. Wherever signals or patterns need to be predicted using a database of previously recorded patterns, this point is very likely to be of much importance. This should be true whether the signal is noisy or not, although the type of validity and manner of prediction of noisy signals will differ. This should be true for temporal patterns as well as spatial patterns, although we have not treated spatial patterns explicitly.

Suppose we follow a storm system which crosses from Canada to Minnesota and then to Wisconsin and makes landfall at Michigan after crossing Lake Michigan. If we want to predict the path of the storm after it makes landfall, it is reasonable to look for earlier storms which followed a similar path across Lake Michigan. The arguments and computations of this paper suggest strongly that such pattern matching is not the right idea. Instead, we have to look for an earlier storm whose eastward path across Lake Michigan may in fact look different. The key property is that the unstable components that separate the two paths must be small so that the paths of the two storms are converging. One has to analyze the unstable components to determine for how long the two paths are likely to remain close after landfall. Although the method for such
analysis is unknown, we have offered a few hints.

12 Acknowledgments

The authors thank Emery Brown, John Gibson, Jeff Humphreys, Charles Li, Steve Lalley, and Roddam Narasimha for useful discussions.

References

[1] H.D.I. Abarbanel. Analysis of observed chaotic data. Springer Verlag, 1996.

[2] G.A. Baker. Essentials of Padé approximants. Academic Press, 1975.

[3] H.W. Bode and C.E. Shannon. A simplified derivation of linear least square smoothing and prediction theory. Proceedings of the IRE, 38:417 – 425, 1950.

[4] E.N. Brown, L.M. Frank, D. Tang, M.C. Quirk, and M.A. Wilson. A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. Journal of Neuroscience, 18:7411, 1998.

[5] M. Casdagli. Nonlinear prediction of chaotic time series. Physica D: Nonlinear Phenomena, 35:335–356, 1989.

[6] M. Casdagli, S. Eubank, J.D. Farmer, and J. Gibson. State space reconstruction in the presence of noise. Physica D: Nonlinear Phenomena, 51:52–98, 1991.

[7] S. Chernyshenko and M. Bondarenko. Master-modes in 3D turbulent channel flow. Arxiv preprint arXiv:0809.2896, 2008.

[8] J.P. Eckmann and D. Ruelle. Fundamental limitations for estimating dimensions and Lyapunov exponents in dynamical systems. Physica D: Nonlinear Phenomena, 56:185–187, 1992.

[9] J.D. Farmer and J.J. Sidorowich. Predicting chaotic time series. Physical review letters, 59:845–848, 1987.

[10] H. Furstenberg. Stationary processes and prediction theory. Princeton Univ Pr, 1960.

[11] P. Grassberger and I. Procaccia. Characterization of strange attractors. Physical review letters, 50:346–349, 1983.

[12] P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. Physica D: Nonlinear Phenomena, 9:189–208, 1983.

[13] R.W. Hamming. Numerical methods for scientists and engineers. Dover Publications, 1986.
[14] W. Hurewicz and H. Wallman. *Dimension theory*. Princeton, 1941.

[15] M. Kac. On the notion of recurrence in discrete stochastic processes. *Bull. Am. Math. Soc.*, 53:1002–1010, 1947.

[16] A. Katok and B. Hasselblatt. *Introduction to the modern theory of dynamical systems*. Cambridge Univ Pr, 1997.

[17] S.P. Lalley. Beneath the noise, chaos. *The Annals of Statistics*, 27:461–479, 1999.

[18] R. Narasimha, S.R. Kumar, A. Prabhu, and SV Kailas. Turbulent flux events in a nearly neutral atmospheric boundary layer. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365:841, 2007.

[19] D.S. Ornstein and B. Weiss. Entropy and data compression schemes. *Information Theory, IEEE Transactions on*, 39:78–83, 1992.

[20] G. Sugihara and RM May. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature*, 344(6268):734–741, 1990.

[21] F. Takens. Detecting strange attractors in turbulence. *Dynamical systems and turbulence, Warwick 1980*, pages 366–381, 1981.

[22] D. Viswanath. The fractal property of the Lorenz attractor. *Physica D: Nonlinear Phenomena*, 190:115–128, 2004.

[23] N. Wiener. *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*. The MIT Press, 1949.

[24] AD Wyner and J. Ziv. Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. *IEEE Transactions on Information Theory*, 35:1250–1258, 1989.

[25] L.S. Young. What are SRB measures, and which dynamical systems have them? *Journal of Statistical Physics*, 108:733–754, 2002.