SIAM J. Applied Dynamical Systems
Vol. 12, No. 2, pp. 1085–1113

Metric Entropy and the Optimal Prediction of Chaotic Signals*

Divakar Viswanath†, Xuan Liang‡, and Kirill Serkh§

Abstract. Suppose we are given a time series or a signal $x(t)$ for $0 \le t \le T$. We consider the problem of predicting the signal in the interval $T < t \le T + t_f$ based on 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 $t_f \le \log_2 T/H$ in the limit $T \to \infty$, with H being entropy in a sense that is explained in the text. If $\left|x(T-s) - x(t^*-s)\right|$ is small for $0 \le s \le \tau$, where τ 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 one way or another. 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. 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.

Key words. optimal prediction, metric entropy, Padé approximant

AMS subject classifications. 37N30, 60G25

DOI. 10.1137/110824772

1. Introduction.

We consider the problem of predicting a signal or a time series $x(t)$ in the time interval $T < t \le T + t_f$ assuming knowledge of the signal in the time interval $0 \le t \le T$. The signal is assumed to originate from a deterministic dynamical system, but we assume no knowledge of the physical model. We consider the signal to be known but assume no knowledge of the physical model in order to obtain a mathematically rigorous context for prediction theory. Thus we are able to state precisely what an optimal predictor should do. Unfortunately, the predictors in current use are not optimal.

The restricted setting where the physical model is assumed to be entirely unknown is advantageous in making a connection to the fundamental results of Wyner and Ziv in 1989 [35] and Ornstein and Weiss in 1992 [28] in information theory. A quantity that plays a central role in determining the predictability of a signal is metric entropy. For examples where a physical model is useful for prediction, see [9, 12]. As examples of chaotic signals

*Received by the editors February 16, 2011; accepted for publication (in revised form) by A. Stuart March 5, 2013; published electronically June 26, 2013. This research was supported by NSF grants DMS-0715510, DMS-1115277, and SCREMS-1026317.
http://www.siam.org/journals/siads/12-2/82477.html
†Department of Mathematics, University of Michigan, Ann Arbor, MI 48109 (divakar@umich.edu).
‡Department of Economics, University of California at San Diego, La Jolla, CA 92093 (x5liang@ucsd.edu).
§Department of Applied Mathematics, Yale University, New Haven, CT 06511 (kirill.serkh@yale.edu).

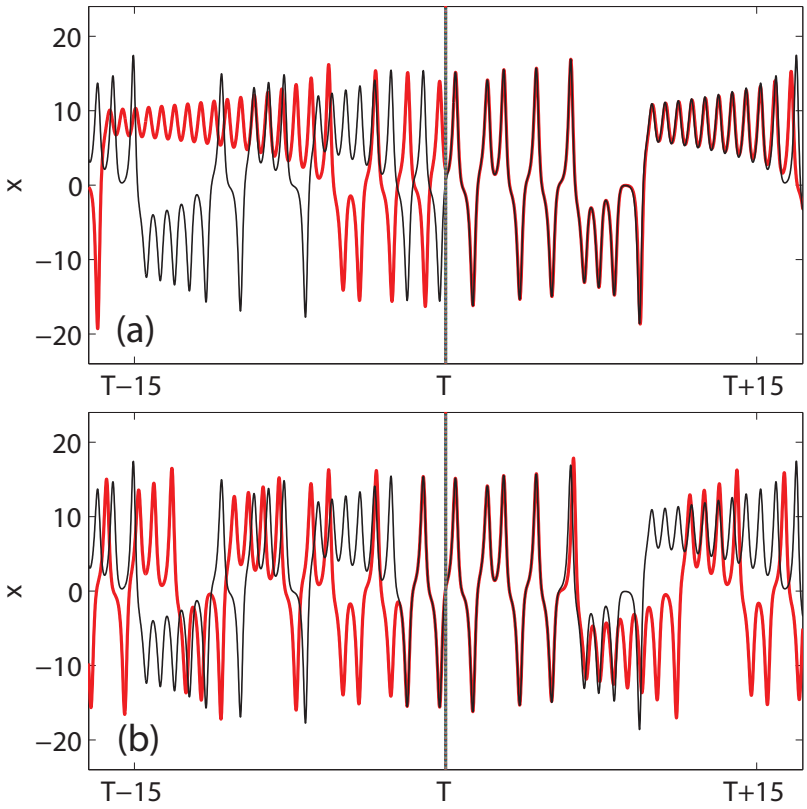


Figure 1. *In both plots, the current time T is 2^{20} symbols. The Lorenz signal, which is shown as a thin black line, is the same in the two plots. The thick red line is* (a) *best fit from the past and* (b) *suboptimal prediction using the embedding predictor.*

whose entropy is not too high, we mention signals obtained from Taylor–Couette flow and Rayleigh–Benard flow [14]. Turbulent signals in atmospheric boundary layers, such as those recorded in [27], have very high entropy.

In the rest of this introduction, we summarize the contributions of this paper and point out connections to other lines of research such as data assimilation.

Optimality and suboptimality of prediction. 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 $0 \le t \le T$. The task is to use the available history, which is $x(t)$ for $0 \le t \le T$, to predict $x(T + t)$ for $0 < t \le t_f$ 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 $t_{return} = 0.7511$ units of time, where t_{return} is the average time from one

"turning point" to another. A turning point is defined as a peak or a trough of 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 [33]. In the plots, T is given as 2^{20} symbols, which means $T = 2^{20} \times 0.7511$ and 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 $t^* \in [0, T - t_{return}]$ such that the length of fit, which is

$$\text{length of fit at } t^* = \text{largest } t_f \text{ such that } |x(T+s) - x(t^*+s)| \le tol \text{ for } s \in [0, t_f], \tag{1.1}$$

is maximized. The maximum value of the length of fit is denoted by t_{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_{best} will be called the best fit from the past. Since the method 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 1(a) follows the signal for $t > T$ for more than 20 symbols. It is not difficult to see why no predictor can follow the signal for a longer time. If we fit the signal starting at $x(t)$, $0 \le t \le T - t_{return}$, 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_{best}, we have $t_f \le t_{best}$. Thus the past has no information about what happens to the signal beyond $T + t_{best}$, and no amount of algorithmic legerdemain can synthesize that information.

Currently available methods for prediction are based on recurrence and embed the signal in phase space in one way or another [1, 7, 10, 14, 30]. All the current predictors known to us use delay coordinates. Suppose the real-valued signal $x(t)$ is obtained as $x(t) = b^T X(t)$, 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 τ and the embedding dimension k [13, 31].

It should be clear that the basic task of a predictor is to find t^* which maximizes (1.1) or another t^* which nearly maximizes it without looking into the future. The embedding predictors do not accomplish that task in an optimal way. For the discussion here, a brief account of the basic 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 $t^* \in [k\tau, T - t_{return}]$ such that

$$\sum_{p=0}^{k-1} \left(x(T - p\tau) - x(t^* - p\tau)\right)^2 \tag{1.2}$$

is minimized. There is much literature about the choice of the delay parameter τ and the embedding dimension k (see [1], for instance). We will assume that τ and k are suitably chosen (with τk 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 (1.2) work in terms of maximizing (1.1)? Before answering that question, let us ask ourselves why there should be a connection at all between finding t^* to minimize (1.2) and finding it to maximize the length of fit defined by (1.1). When we minimize (1.2), 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. 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 by 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 nearly converge 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).

The situation shown in Figure 1 is typical. Because of the nature of chaotic signals, best fits tend to converge at $t = T$ and diverge rapidly beyond $t = T + t_{best}$. 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 whether the pattern of events leading up to $t = t^*$ and $t = T$ is such that the two patterns will come close to each other in the future and to calculate 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.

What are these stable and unstable components? Ideally, one would like to define a notion of stable and unstable components that uses signals and nothing more. Since no optimal general purpose predictor of chaotic signals is currently known, such a notion cannot be made precise. However, it is clear that such a notion has to correspond in some way with stable and unstable manifolds or with stable and unstable directions associated with local Lyapunov exponents of the underlying dynamical system.

In the limit of $T \to \infty$, the stable and unstable components may be identified with the stable and unstable manifolds. However, for finite T, especially considering the short intervals for which prediction is possible, one has to use a notion of stable and unstable components associated with local Lyapunov exponents. These fixed intervals of time used for defining local Lyapunov exponents can be taken as $\log_2 T/H$.

If we split the distance between the pattern of events leading up to $t = T$ (black line in Figure 1(a)) and the pattern of events leading up to $t = t^*$ for the best fit (thick red line with t^* shifted to T in Figure 1(a)), 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 converge and stay close for an interval of time. The smallness of the unstable component allows the fit between the two signals to persist for the longest interval of time.

Metric entropy. Section 2 states a theorem of Wyner and Ziv [35] and Ornstein and Weiss [28], 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

$$\lim_{T\to\infty} \frac{t_{best}}{\log_2 T} = \frac{1}{H}$$

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 $\left[T, T + t_f\right)$ 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. In sections 6 and 7, we develop a few ideas that take us closer to a general purpose optimal predictor for chaotic signals.

Data assimilation and shadowing filters. Shadowing filters have been proposed as a method for state estimation and data assimilation. Before discussing the connection of this paper to shadowing, we give a brief discussion of data assimilation. This brief discussion has two goals. It has been stated that "the forecast skill of atmospheric models depends not only on the accuracy of the initial conditions and the realism of the model, but also on the instabilities of the flow itself" [23, p. 227]. The current paper is focused exclusively on the instabilities of the flow. Since weather and ocean models [5, 23] are major applications of prediction theory, it is perhaps not out of place to call attention to measurement and modeling errors. Second, the discussion provides some context for shadowing filters.

The following equation provides a framework for many data assimilation techniques [23]:

$$X^a = X^b + W\left(Y^o - H\left(X^b\right)\right).$$

In this equation, X^a and X^b are vectors which correspond to a point in the state space of the physical model. Atmosphere model variables typically include wind velocity components, temperature, moisture, and surface pressure. The observation vector is denoted by Y^0. Observed variables such as satellite radiances and radar reflectivities do not occur in the physical model. The observation operator H maps the state vector of the physical model to observation space. In weather prediction as well as climate modeling, the number of degrees of freedom in the physical model is orders of magnitude greater than the number of observations. Therefore it is impossible to synthesize the current state of the model X^a from observations alone.

The background field X^b obtained from a short-term forecast is used as a starting point for inferring the current state of the physical model.

The essence of data assimilation in this framework is the operator W, which matches the observations against the background field and generates a correction. Techniques such as optimal interpolation, 3DVar, and PSAS fall within this framework. For a mathematical study of such techniques, see [3]. All practical methods must account for the covariance of measurement error. In numerical weather forecasting, this type of data assimilation is performed in six-hour cycles, and information gradually propagates from regions rich in observations to regions poor in observations [23].

Another family of techniques explicitly allows observations to be functions of time [23]. One of these is the extended Kalman filter. The extended Kalman filter updates the covariance matrix of the estimated state vector from time to time using new observations. Propagation and manipulation of the covariance matrix for realistic physical models can be expensive. The ensemble Kalman filter is a cheaper variant which introduces random errors into observations and tracks several trajectories to estimate the covariance matrix. Yet another technique is 4DVar. This technique finds the initial state X_0 to minimize the quantity

$$\left(X_0 - X^b\right)^T B^{-1} \left(X - X_0\right) + \sum_{i=0}^{N} \left(H\left(X_i\right) - Y_i^o\right)^T R_i^{-1} \left(H\left(X_i\right) - Y_i^o\right).$$

Here R_i is the covariance matrix of observations recorded at time t_i. The state X_i at time t_i must be obtained by integrating the physical model assuming the state at t_0 to be X_0. The background field has a significant role in this technique as well. The shadowing filters, to which we now turn, solve a minimization problem that is formally similar to 4DVar.

The trajectory $\tilde{X}(t)$ is an ϵ-orbit of the dynamical system $dX/dt = f(X)$ if $||d\tilde{X}/dt - f(\tilde{X})|| < \epsilon$ for all t. The shadowing lemma states that if the ϵ-orbit stays inside a suitable neighborhood of a hyperbolic invariant set and for ϵ small enough, the ϵ-orbit is δ-shadowed by a true orbit of the dynamical system [8, 24]. A similar result applies to hyperbolic invariant sets of maps. Hammel, Yorke, and Grebogi [18] have shown that the numerically computed orbit of the Henon map $(u_{n+1}, v_{n+1}) = (1 - Au_n^2 + v_n, -Ju_n)$ with $A = 1.4$, $J = -0.3$, and $(u_0, v_0) = (0, 0)$ is δ-shadowed by a true orbit for up to $N = 10^7$ iterations with $\delta = 10^{-8}$, even though the Henon map is not uniformly hyperbolic. A similar result is given for the Ikeda map.

The essence of shadowing is that the error committed in each step of an iteration may be decomposed along stable and unstable manifolds. The error along the stable manifold can be canceled using a very small perturbation at the final point of the trajectory. Similarly, the error along the unstable manifold can be canceled using a very small perturbation at the initial point of the trajectory.

As already mentioned, shadowing filters have been proposed for noise reduction and state estimation [17, 15, 20, 21, 29]. For example, if s_t are noisy observations of the state x_t for $t = 1, \ldots, T$, one may attempt to calculate the noise δ_t by minimizing $\sum_{t=1}^{T} ||e_t||^2$, where $e_t = s_{t+1} - \delta_{t+1} - f(s_t - \delta_t)$ [20]. Here the physical model f is assumed to be known. Other versions of the shadowing filter assume only partial knowledge of f [21].

Judd and Smith [20] have considered various gradient descent methods for denoising and

stated a "dictum" based on numerical experience. Their dictum is that the end point of the estimated trajectory will lie close to the unstable manifold of the end point of the true trajectory. In other words, much of the error near the end point is along the unstable direction. Unfortunately, this is the exact opposite of what an optimal predictor must do. This situation results because the shadowing filter uses gradient descent to match the entire segment of the trajectory as nearly as possible and is therefore biased to fitting the past. The asymmetry between fitting the past and predicting the future is not broken in favor of the latter.

Work of Lalley. Lalley and others (see [25, 26]) have subjected the problem of denoising deterministic signals to an incisive mathematical investigation. Lalley has found that an effective denoising algorithm must increase the width of the matching window at a sublogarithmic rate. Some of the considerations that led to that finding could be relevant to optimal prediction. Lalley's work provides a useful contrast to more applied work on denoising. Judd and Smith [21] refer to an earlier paper of theirs and state, "we showed, that contrary to what might be expected, collecting more and more data will not provide a continually improving estimate of the true state of a chaotic system." Here it must be understood that Judd and Smith are referring to shadowing filters based on gradient descent. In fact, Lalley [25] has used very general assumptions on additive noise to prove that his algorithm can recover the state of a chaotic system by collecting more and more data.

2. Theorem of Wyner and Ziv and Ornstein and 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 and Ziv and Ornstein and 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 real-valued 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 $T : \Omega \to \Omega$ be a measurable transformation that preserves the probability measure μ on Ω. 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 1 (Poincaré recurrence [24]). *Assume X_0 to be μ-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 Ω 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 10 characters of the text. The theorem then asserts that the sequence formed by the first 10 characters will repeat again and again. The origin of the sequence X_0 can be taken anywhere in the text.

If Ω 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 (Kac's theorem [22]). *Suppose that the sequence $X_0, X_1, \ldots$ is stationary and ergodic. Let $B \subset \mathbb{R}$ with $\mathbb{P}(B) = \mathbb{P}(X_0 \in B) > 0$. Let $n \geq 1$ be the smallest integer such that $X_n \in B$. Then $\mathbb{E}\left(n \middle| X_0 \in B\right) = 1/\mathbb{P}(B)$.*

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.

Since X_n is assumed to take values in a finite alphabet $\mathcal{A}$ 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 $|\mathcal{A}|^m$ different values. Suppose the probabilities of the different possibilities are $p_1, p_2, \ldots, p_M$, where $M = |\mathcal{A}|^m$. Then

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

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 $|\mathcal{A}|^m$ possible values of the sequence $X_0, \ldots, X_{m-1}$. Some possible sequences are more probable and some are less probable. How many 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 fewer. An optimal compression of the source will use mH bits to encode m symbols asymptotically.

Theorem 3 (Ornstein and Weiss [28]). *Let $X_0, X_1, \ldots$ be a stationary and ergodic sequence,*

in which each X_n takes values in a finite alphabet $\mathcal{A}$. Let t_{best} be the greatest integer such that $X_{T+1}, \ldots, X_{T+t_{best}}$ occurs as a contiguous subsequence of $X_0, \ldots, X_T$. Then

$$\lim_{T\to\infty} \frac{t_{best}}{\log_2 T} = \frac{1}{H}$$

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

Theorem 2 tracks the recurrence of an event associated with X_0 for some X_n with $n > 0$. Theorem 3 checks whether an event that follows the current symbol X_T repeats a past event. We will refer to either scenario as a recurrence.

Theorem 3 is a remarkable sharpening of the Poincaré recurrence theorem. If we regard T as current time and assume that observations begin at 0, as we do throughout this paper, we have a perfect characterization of the extent to which the pattern that follows T will repeat some pattern of events we have seen in the past. The $1/H$ bound was first stated by Wyner and Ziv [35], who were able to prove the convergence of $t_{best}/\log_2 T$ to $1/H$ in probability. Almost sure convergence of the type asserted by Theorem 3 was proved by Ornstein and Weiss [28].

The distinction between convergence in probability and almost sure convergence is pertinent to the prediction of chaotic signals. If predictions of weather or hurricane tracks or 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 [14, 30]. Our discussion of predictability of chaotic signals will be framed in terms of almost sure predictability.

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 3.

Theorem 3 interprets entropy in terms of the maximum segment following X_T that occurs as a contiguous subsequence of the segment preceding it. The entire segment following 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 3. This convention saves us the trouble of using four names—Wyner, Ziv, Ornstein, and Weiss—every time 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 independent and identically distributed (i.i.d.) sequence made up of

± 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 an 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 nor the Y_n sequence is predictable in the manner we consider. For notions of prediction pertinent to such signals, see [16].

Suppose $X_{n+1} = f(X_n)$ is a dynamical system. The phase space of the dynamical system can be any Riemannian manifold, but for convenience we will assume it to be a subset of $\mathbb{R}^d$. Let μ be a probability measure that is invariant with respect to the dynamical system (in other words, $\mu(A) = \mu(f^{-1}(A))$ for Borel sets A). If X_0 has μ 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 μ 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 of 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:

$$\begin{aligned} \frac{dx}{dt} &= 10(y - x), \\ \frac{dy}{dt} &= 28x - y - xz, \\ \frac{dz}{dt} &= -8z/3 + xy. \end{aligned}$$

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 holes in the two wings of the butterfly-shaped attractor. 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 way may be assumed to be μ distributed, with μ 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 [24] or [36]. Let $f : \mathbb{R}^d \to \mathbb{R}^d$ be a smooth dynamical system, and let $\mathcal{A}$ be an invariant set. Let μ be a probability measure on $\mathcal{A}$ that is invariant with respect to f. Assume that f is an ergodic transformation of $\mathcal{A}$ with respect to the measure μ. In this setting, the definition of metric or Kolmogorov–Sinai entropy is quite simple. Let $\mathcal{P}$ be a finite partition of the set $\mathcal{A}$. We can generate a finite-valued stationary ergodic process as follows. Pick X_0 according to μ, and take $X_{n+1} = f(X_n)$ for $n = 0, 1, \ldots$. Let Y_n be the partition in $\mathcal{P}$ to which X_n belongs. Then the finite-valued process Y_n is stationary and ergodic and as such has a Shannon entropy which we denote by $h_\mu(f, \mathcal{P})$. In general, $h_\mu(f, \mathcal{P})$ can depend upon the partition $\mathcal{P}$. The metric entropy $h_\mu(f)$ is defined as the maximum over all finite partitions $\mathcal{P}$.

At first glance, it might seem as if the dependence of metric entropy on $\mathcal{P}$ could be a problem. However, this dependence is not as severe as one might think. For example, one may modify $\mathcal{P}$ to the finer partition $\mathcal{P} \vee \mathcal{P}$, where the finer partition keeps track of the partitions in $\mathcal{P}$ to which x and its iterate $f(x)$ belong. Even though $\mathcal{P} \vee \mathcal{P}$ is a finer partition, $h_\mu(f, \mathcal{P} \vee \mathcal{P}) = h_\mu(f, \mathcal{P})$ because it is readily evident that combining the nth and the $(n+1)$st symbols into a single symbol in the nth position will neither increase nor decrease the information per symbol. In fact, $h_\mu(f) = h_\mu(f, \mathcal{P})$ if the partition $\mathcal{P}$ is *generating*. Generating partitions are defined using conditional entropy [24]. If the partitions in $\mathcal{P} \vee \cdots \vee \mathcal{P}$ become fine enough to closely approximate any given partition $\mathcal{Q}$ of $\mathcal{A}$, the partition $\mathcal{P}$ is generating.

Later the theoretical discussion will focus on hyberbolic attractors $\mathcal{A}$. 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 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 [32, 33].

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 $\mathcal{A}$ grow or decay. For a definition, see [24]. 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 counting multiplicities. With probability 1 with respect to the measure μ, these are the only possible rates of growth or decay.

Table 1

Recurrences of a Lorenz signal. The subsequence extending from position $T+1$ to position $T+t_{best}$ is matched by a subsequence extending from position t^* to t^*+t_{best}, where t^* is a position in the past with $t^*+t_{best} \le T$.

$\log_2 T$	t_{best}	Matching sequence	$\log_2 T$	t_{best}	Matching sequence
2	2	BA	12	13	AAAABAAABBAAA
3	3	BBB	13	12	BBBBABAAABAA
4	7	ABABBBB	14	16	ABBBABBAAAAAAABA
5	7	BBAABAA	15	17	BBBABBBABABBAABAB
6	9	BBBBBBBAA	16	16	BAABABBBAAAAAABB
7	10	AAAAAAAAAA	17	20	AAABBBAAABABAAABAAAA
8	14	AABAABAAAAAAAA	18	24	BABBABBBABBBBABBABBABABB
9	9	BBBAABBAA	19	20	BBBBABABBBAABAAAABAA
10	10	BABABBBBAB	20	16	ABBBAABABAAAABA
11	9	ABBBAAAAA	21	30	BABBBBAABABBAAAABBBBAAAAAABABB

Table 2

Recurrences of flips of a fair coin calculated in the same manner as in the previous table.

$\log_2 T$	t_{best}	Matching sequence	$\log_2 T$	t_{best}	Matching sequence
2	6	AABAAA	12	12	AAABAABAABBB
3	4	AAAB	13	12	BABBAAAAAABA
4	8	ABAAAABB	14	13	BBAABBBABABBB
5	7	BBABBAB	15	14	BBAAAAAAAAAAAB
6	6	BABAAA	16	16	BAAABBBABBBABABB
7	11	AABAAABBBBB	17	20	AAAABABABAABAABABBAB
8	5	BABBB	18	17	BAABAAAABAABAABBB
9	11	BBBBBBBABBA	19	22	BABBAAAAAABBBBAAABAAAB
10	12	BAAAABABAABA	20	19	BABAAAAABABABBABBAB
11	10	BABBABAABB	21	20	ABABBBBBBBBBABABBAAA

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

$$h_\mu \le \sum_{\lambda_i > 0} \lambda_i. \tag{3.1}$$

This is Ruelle's inequality [36] (the same logarithm must be used in defining h_μ and the Lyapunov exponents λ_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 quarter-planes A or B to another is $t_{return} = 0.7511$. By Ruelle's inequality (3.1), the entropy of the sequences of A's and B's is bounded above by $0.905 \times 0.7511/\log 2 = 0.98$. The entropy appears to be close to 0.98 [32, 33]. Table 1 appears to be in agreement with this estimate of the entropy.

Table 2 tabulates t_{best} (defined as in Theorem 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_{best} are more pronounced for the Lorenz signal. For the special case of i.i.d. sequences such as coin tosses, Theorem 3 was proved by Erdős and Renyi.

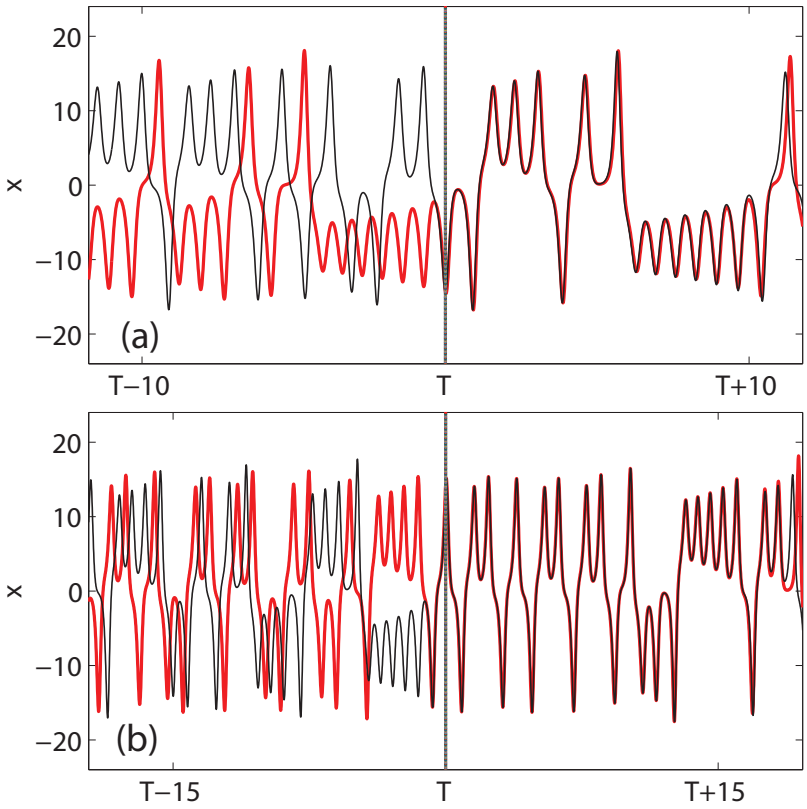


Figure 2. *Best fits from the past (thick red lines) to a Lorenz signal (thin black lines).* (a) $T = 2^{14}$ *symbols.* (b) $T = 2^{21}$ *symbols.*

The intersections with the quarter-planes A and B are recorded using the symbols A and B. For continuous time Lorenz signals $x(t)$, one may use the average time between symbols $t_{return} = 0.7511$ as the unit. Following that usage, the values of the current time T for the two plots in Figure 2 are reported as 2^{14} and 2^{21} symbols.

When we think of the Lorenz signal as a sequence made up of the symbols A and B, it is natural to define t_{best} as in the entropy theorem (Theorem 3). However, for continuous time signals the definition of t_{best} which follows (1.1) is more natural. We take

$$tol = 5 \tag{3.2}$$

to be the tolerance for Lorenz signals throughout this paper. Table 3 reports t_{best} with $tol = 5$. The t_{best} numbers with $tol = 5$ are somewhat smaller than the t_{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 later compared to the best fits in Table 3, the length of match is reported in symbols but not as a real number.

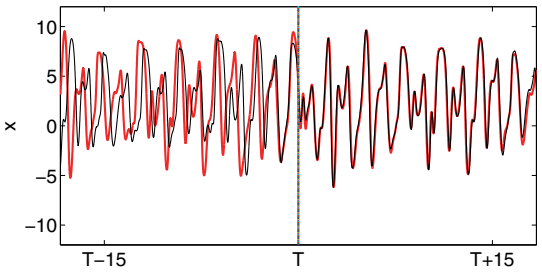
For another example of the applicability of the entropy theorem, we turn to the following equations:

$$\frac{dx_j}{dt} = x_{j-1}\,(x_{j+1} - x_{j-2}) - x_j + f \tag{3.3}$$

Table 3

Best fits to a Lorenz signal, where t_{best} in symbols equals t_{best} as a real number divided by $t_{return} = 0.7511$.

$\log_2 T$ (in symbols)	t_{best} (in symbols)	t_{best} (as a real)	$\log_2 T$ (in symbols)	t_{best} (in symbols)	t_{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



T	t_f
10^2	3.6
10^3	7.4
10^4	14.2
10^5	18.4

Figure 3. *Best fit from the past for a chaotic signal obtained from (3.3) using $f = 8.17$, and a table showing the logarithmic dependence of t_f on T.*

with $j = 0, 1, 2, 3, 4$ and with the arithmetic in the subscripts being modulo 5. When $f = 8.17$, this system is chaotic [2]. As shown in Figure 3, the entropy theorem applies to this chaotic system (the signal is from x_0). As expected, the best fit into the future diverges from the signal as we walk back in time. This example was introduced in [2] to show that a single signal cannot be used to synchronize a chaotic physical model. In this case, the system has two positive conditional Lyapunov exponents, and two signals are needed to synchronize the physical model. The same point has come up in the theory of the Navier–Stokes equations. For example, less than 5% of the modes suffice to synchronize turbulent channel flow, but less than 1% will not do [12]. The master modes or the determining modes must be sufficiently numerous to capture the entire system, a point we alluded to at the beginning of the introduction.

4. Recurrence of chaotic signals and limits of predictability. Suppose we are trying to predict a signal $x_0, \ldots, x_T$. The entropy theorem says that $t_{best} \approx \log_2 T/H$ for large T. Thus it appears that 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 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 $\mathcal{A}$. We assume that f is transitive on $\mathcal{A}$. 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

$$P\left(X_0, \ldots, X_T\right) = \left(\tilde{X}_{T+1}, \tilde{X}_{T+2}, \ldots\right).$$

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. A prediction $\tilde{X}_{T+s}$ is deemed to be valid if $\left|X_{T+s} - \tilde{X}_{T+s}\right| \leq tol$ 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,

$$\limsup_{T \to \infty} \frac{t_f}{\log_2 T} < \frac{1+\epsilon}{H} \tag{4.1}$$

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 μ on $\mathcal{A}$. 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 μ-distributed, $X_1 = f\left(X_0\right)$, $X_2 = f\left(X_1\right)$, and so on.

In order to explain why every prediction algorithm must satisfy the bound (4.1), we turn to another notion of entropy, namely, topological entropy $h_{top}(f)$ [24]. To begin 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 δ balls required to cover $\mathcal{A}$ in the metric d_n, topological entropy is defined using the relation $N(\delta, n) \approx C 2^{n h_{top}}$ for small δ. It is independent of the metric. In general, $h_\mu(f) \leq h_{top}(f)$ (see Theorem 4.5.3 of [24]). With the assumptions we have made on $\mathcal{A}$ and μ, $h_\mu = h_{top}$.

Suppose we are given the sequence $X_0, \ldots, X_T$. This 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 μ-distributed [36]. To predict n iterates of X_T with tolerance $tol = \delta$ from that information, we require one of the points $X_0, \ldots, X_{T-n}$ to be within δ of X_T in the d_{n+1} metric. For such a thing to be possible, we require $T - n \geq C 2^{n h_{top}}$ or $n \leq \log_2 T / h_{top}$ asymptotically.

It may seem that one may extract some more information about $f^n\left(X_T\right)$ 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 $T < C2^{nh_{top}/(1+\epsilon)} = C2^{n/(1+\epsilon)}$, 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.

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 $0 \le t \le T$ 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 continuation. 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 [34].

A prediction algorithm P is optimal if

$$\liminf_{T\to\infty} \frac{t_f}{\log_2 T} \ge \frac{1-\epsilon}{H} \tag{4.2}$$

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 [35] for an example.

5. The embedding predictor, related predictors, and their suboptimality. With regard to Wiener–Kolmogorov predictors, Wiener wrote [34, p. 71] the following: *geometrical 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 for decomposing 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 α is a small fraction.

Phase space reconstruction using delay coordinates is used by all existing predictors. We refer to the most basic of these predictors as the embedding predictor [10, 14]. Given a signal $x(t)$ for $0 \le t \le T$, the embedding predictor finds t^* to minimize (1.2), 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 $\left|x\left(t_1\right) - x\left(t_2\right)\right|$ 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 whether X_1 and X_2 are close to each other or not.

Although the choices of the delay parameter τ 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 τ 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 dimension k, it is stated that it should be at least as large as the dimension of the underlying chaotic set. For the validity of Taken's embedding theorem [31], $k \ge 2m+1$ is required with m being the dimension of the chaotic set.

Figure 4 shows suboptimal predictions of Lorenz signals using the embedding predictor and $\tau = .03$, $k = 5$. Our choice of the delay parameter at $\tau = .03$ is much smaller than what the mutual information criterion would imply. The mutual information criterion would imply a τ 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 extant 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. [11]. 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 4 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

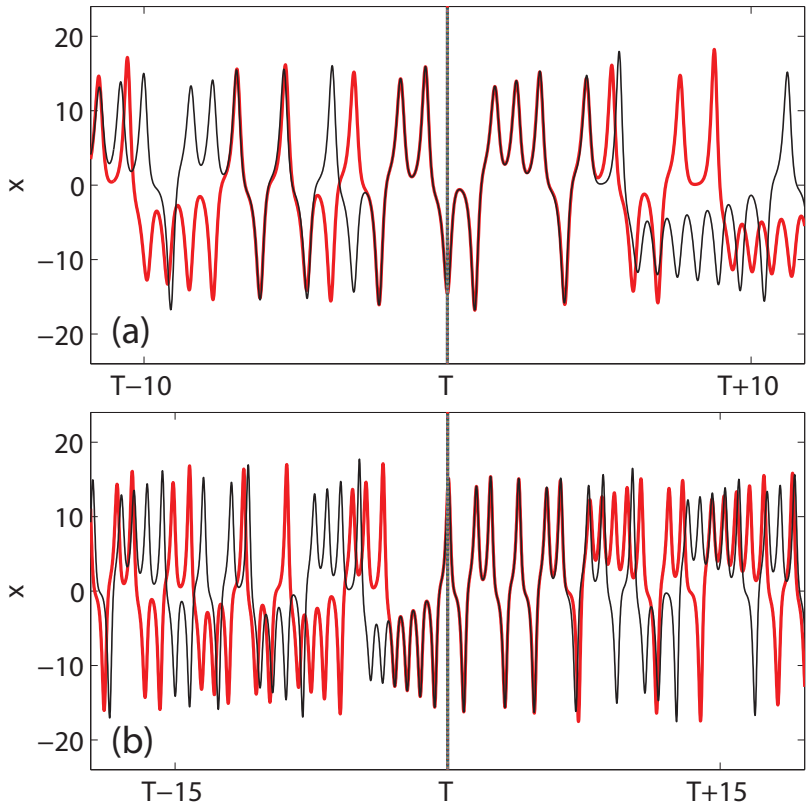


Figure 4. *Suboptimal predictions (thick red lines) using the embedding predictor to a Lorenz signal (thin black lines).* (a) $T = 2^{14}$ *symbols.* (b) $T = 2^{21}$ *symbols.*

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 4, 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 that 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_{embed} column of Table 4 is obtained as follows. The metric (1.2) 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_{embed}: $|x(t^* + s) - x(T + s)| \leq tol$ for $0 \leq s \leq t_{embed}$ but not for $0 \leq s \leq t$ with $t > t_{embed}$. Comparison of t_{embed} in Table 4 with t_{best} in Table 3 shows that the embedding predictor does not approach optimality.

In the rest of this section, we consider a number of extant ideas for improving the basic embedding predictor. All these ideas have merit. However, to be fully effective, they need to

Table 4

Length of suboptimal predictions of a Lorenz signal (t_{embed}) using the embedding predictor.

$\log_2 T$ (in symbols)	t_{embed} (in symbols)	$\log_2 T$ (in symbols)	t_{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

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 [14]. To predict $x(T+s)$, the basic embedding predictor picks a single $t^* \in [k\tau, T-s]$ using the metric (1.2). Instead, a predictor may pick p different instants $t_1^*, \ldots, t_p^*$ 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 $t = T$ by fitting $x(t_i^* + s)$ as a linear combination of the delay coordinates at $t = t_i^*$ for $i = 1, \ldots, p$, 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 ($t_{return} = 0.7511$) from $t = T$. For sufficiently large T, the signal will have delay coordinates at $t = t_i^*$ close to those 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 $i = 1, \ldots, p$. Therefore extrapolation using least squares will improve the order of accuracy (see Figure 2 of [14]).

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 $i = 1, \ldots, p$ 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 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 extrapolation 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 [14], to trap the delay coordinates at $t = T$ within a simplex in reconstructed phase space [30], or to weight close recurrences using the closeness of the approach [1]. 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 ϵ 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 that 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

$$x = c + \sum_{i=1}^{u} a_i v_i^+(c) + \sum_{i=1}^{s} b_i v_i^-(c). \tag{6.1}$$

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 λ_i^+ be the characteristic multiplier corresponding to v_i^+, and

similarly let λ_i^- correspond to v_i^- (if l is a Lyapunov exponent defined using natural logarithms, $\exp(l)$ is the corresponding characteristic multiplier). We have

$$f^n(x) \approx c_n + \sum_{i=1}^{u} a_i \left(\lambda_i^+\right)^n v_i^+(c_n) + \sum_{i=1}^{s} b_i \left(\lambda_i^-\right)^n v_i^-(c_n), \quad \text{where} \quad c_n = f^n(c). \tag{6.2}$$

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

To eliminate some linear algebra from the discussion, we will assume that $v_i^+(x)$, $1 \le i \le u$, and $v_i^-(x)$, $1 \le i \le s$, 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 [24].

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 $x_T = c$ and 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 $1 \le i \le u$ and $b_i = B_i\delta$ for $1 \le i \le s$ with A_i and B_i fixed to determine the shape of the box around c and with as small a δ as possible. Kac's theorem (Theorem 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 δ that leaves the box large enough to be likely to include a point from the available history is given by $A_1 \ldots A_u B_1 \ldots B_s \delta^{s+u} \approx 1/T$. 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

$$A_1 \ldots A_u \delta^u \approx \frac{C}{T} \tag{6.3}$$

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:

$$\|f^n(x) - f^n(c)\| \approx \sqrt{\sum_{i=1}^{u} a_i^2 \left(\lambda_i^+\right)^{2n}}. \tag{6.4}$$

Here we have neglected the μ_i^- components because $\left|\mu_i^-\right| < 1$, 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 δ implied by (6.4) is

$$\delta^u \approx \frac{C\left(\prod \lambda_i^+\right)^n}{T}, \tag{6.5}$$

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 δ, we get

$$n \leq \frac{\log_2 T + u \log_2 tol - (u/2)\log_2 u - \log_2 C}{\sum \log_2 \lambda_i^+}. \tag{6.6}$$

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 whether $t = t^*$ will give a long fit to the segment following $x(T)$, with the length of fit defined as in (1.1). 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 (1.2). It does not even attempt to resolve the close recurrences into s, f, and u components as an optimal predictor should.

7. Optimal prediction of toral automorphisms. Let A be a $d \times d$ matrix with integer entries and $\det A = \pm 1$. The map $X_{n+1} = AX_n \bmod 1$ 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 [24]. 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.

We will 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 $X_{n+1} = AX_n \bmod 1$ for $n \geq 0$. The first toral automorphism that is considered is

$$X_{n+1} = AX_n \bmod 1, \quad A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}. \tag{7.1}$$

This matrix A has eigenvalues $\frac{3+\sqrt{5}}{2} \approx 2.61803$ and $\frac{3-\sqrt{5}}{2} \approx 0.381966$, and its entropy is $\log_2 2.61803 = 1.3885$. The second toral automorphism that is considered is

$$X_{n+1} = AX_n \bmod 1, \quad A = \begin{pmatrix} 0 & -1 & 0 \\ 1 & -2 & 1 \\ 2 & -3 & 3 \end{pmatrix}. \tag{7.2}$$

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

Before considering the optimal prediction of signals derived from toral automorphisms, 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 automorphisms 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.

In the previous section, we have emphasized that close recurrences must be resolved into stable and unstable components and the quantities $a_i(\lambda_i^+)$ that appear in (6.2) must be estimated. For hyperbolic toral automorphisms, the stable and unstable directions split in exactly the same way at every point on the torus. The optimal predictor based on Padé approximation that we derive takes advantage of this fact and limits itself to estimating a_i. The difficulty of estimating stable and unstable directions near close recurrences, which must be tackled by a general purpose predictor, are sidestepped by the Padé predictor.

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

$$s_n = \sum_{k=1}^{d} c_k \lambda_k^n, \quad n = 0, 1, \ldots. \tag{7.3}$$

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 λ_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 [19] 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 [4].

Define $f(z) = \sum_{n=0}^{\infty} s_n z^n$. Using (7.3), we get

$$f(z) = \sum_{k=1}^{d} \frac{c_k}{1 - \lambda_k z} =: \frac{\alpha_0 + \alpha_1 z + \cdots + \alpha_{d-1} z^{d-1}}{1 + \beta_1 z + \cdots + \beta_d z^d}.$$

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

$$\begin{aligned}\alpha_0 + \cdots + \alpha_{d-1}z^{d-1} &= \left(1 + \beta_1 z + \cdots + \beta_d z^d\right) \sum_{n=0}^{\infty} s_n z^n \\ &= \sum_{n=0}^{\infty} z^n \left(s_n + \sum_{j=1}^{\min(d,n)} s_{n-j}\beta_j \right).\end{aligned}$$

Equating coefficients of z^n for $n = d, \ldots, 2d-1$, we get the d equations

$$\sum_{j=1}^{d} s_{n-j}\beta_j = -s_n. \tag{7.4}$$

This Toeplitz system must be solved to determine β_j. Its solvability is a necessary condition for exponential extrapolation. Once the β_j are determined, (7.4) is used with $n = 2d, 2d+1, \ldots$ to determine s_{2d}, s_{2d+1}, and so on.

The analogy of this process to the Wiener–Kolmogorov predictor described in [34] 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. The Toeplitz system that comes up in exponential extrapolation is unsymmetric in general, while the Toeplitz system that comes up in the Wiener–Kolmogorov predictor is symmetric.

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 Δx_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 Δx_i sequence. For each i with $0 \le i \le 2d-1$, if $\Delta x_i > 1/2$, we replace Δx_i by $\Delta x_i - 1$. On the other hand, if $\Delta x_i \le -1/2$, we replace Δx_i by $\Delta x_i + 1$. After these operations, we will have $|\Delta x_i| \le 1/2$ 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 $X_{n+1} - X_{m+1} = A(X_n - X_m) \bmod 1$. Therefore if $X_{t^*-2d-1} - X_{T-2d-1}$ is small enough, the sequence Δx_i, $i = 0, \ldots, 2d-1$, can we written as a linear combination of exponentials like the s_i sequence in (7.3). The λ_i will be the eigenvalues of A. We use a tolerance to check whether the Δx_i are small enough to permit sensible exponential extrapolation.

Table 5

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 (7.1) of the two-dimensional torus $\mathbb{T}^2$.

$\log_2 T$	t_{best}	t_{embed}	t_{pade}	$\log_2 T$	t_{best}	t_{embed}	t_{pade}
5	2	1	1	17	12	4	8
6	7	0	0	18	10	4	10
7	3	0	1	19	13	6	10
8	5	1	0	20	15	6	11
9	4	2	2	21	15	6	13
10	4	2	2	22	14	1	12
11	8	2	3	23	19	6	19
12	6	2	5	24	15	9	13
13	13	1	5	25	17	7	17
14	8	6	6	26	16	9	15
15	8	4	6	27	18	7	18
16	11	0	6	28	18	1	16

Using exponential extrapolation, we compute Δx_{2d}, Δx_{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 5, we list t_{best} (the best fit from the past defined as in (1.1)), 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 approaches optimality.

From Figure 5, 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 rapidly converges to the signal starting at T and closely tracks 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 converges rapidly to the signal at $t = T$ and then tracks it for a number of iterates.

Table 6 and Figure 6 refer to the toral automorphism defined by (7.2). 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 6.

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

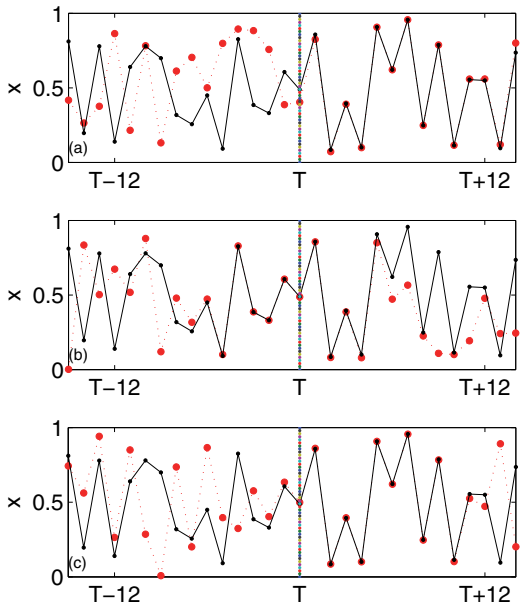


Figure 5. *In each of the three plots, the black dots are part of a signal obtained from the iterates of the automorphism (7.1) of the two-dimensional torus $\mathbb{T}^2$. Here $T = 2^{21}$. The bigger red dots are* (a) *the best fit from the past;* (b) *suboptimal prediction using the embedding method;* (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 (7.2) of the three-dimensional torus $\mathbb{T}^3$.

$\log_2 T$	t_{best}	t_{embed}	t_{pade}	$\log_2 T$	t_{best}	t_{embed}	t_{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

be able to approach optimality.

8. Conclusion.

Matching the pattern of events leading to the present moment is a natural idea for predicting nonlinear signals. In 1950, Bode and Shannon [6] expressed that idea 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

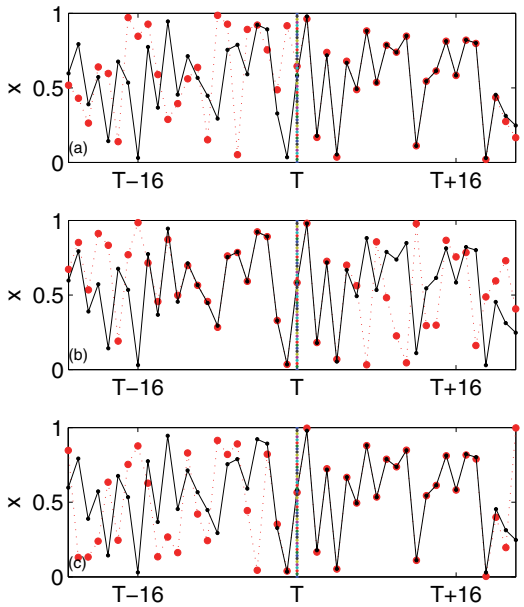


Figure 6. *In each of the three plots, the black dots are part of a signal obtained from the iterates of the automorphism* (7.2) *of the three-dimensional torus* $\mathbb{T}^3$. *Here* $T = 2^{28}$. *The bigger red dots are* (a) *the best fit from the past;* (b) *suboptimal prediction using the embedding method;* (c) *optimal prediction using the Padé method.*

> 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."

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 counterintuitive 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.

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

REFERENCES

[1] H. D. I. Abarbanel, *Analysis of Observed Chaotic Data*, Springer-Verlag, New York, 1996.

[2] H. D. I. Abarbanel, D. R. Creveling, R. Farsian, and M. Kostuk, *Dynamical state and parameter estimation*, SIAM J. Appl. Dyn. Syst., 8 (2009), pp. 1341–1381.

[3] A. Apte, C. Jones, A. M. Stuart, and J. Voss, *Data assimilation: Mathematical and statistical perspectives*, Int. J. Numer. Methods Fluids, 56 (2008), pp. 1033–1046.

[4] G. A. Baker, Jr., *Essentials of Padé Approximants*, Academic Press, New York, London, 1975.

[5] A. F. Bennett, *Inverse Modeling of the Ocean and Atmosphere*, Cambridge University Press, Cambridge, UK, 2002.

[6] H. W. Bode and C. E. Shannon, *A simplified derivation of linear least square smoothing and prediction theory*, Proc. I.R.E., 38 (1950), pp. 417–425.

[7] G. Boffetta, M. Cencini, M. Falcioni, and A. Vulpiani, *Predictability: A way to characterize complexity*, Phys. Rep., 356 (2002), pp. 367–474.

[8] R. Bowen, *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*, 2nd revised ed., Springer-Verlag, Berlin, New York, 1975, 2008.

[9] 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*, J. Neurosci., 18 (1998), pp. 7411–7425.

[10] M. Casdagli, *Nonlinear prediction of chaotic time series*, Phys. D, 35 (1989), pp. 335–356.

[11] M. Casdagli, S. Eubank, J. D. Farmer, and J. Gibson, *State space reconstruction in the presence of noise*, Phys. D, 51 (1991), pp. 52–98.

[12] S. Chernyshenko and M. Bondarenko, *Master-Modes in 3D Turbulent Channel Flow*, preprint arXiv:0809.2896v1 [physics.flu-dyn], 2008.

[13] J. P. Eckmann and D. Ruelle, *Fundamental limitations for estimating dimensions and Lyapunov exponents in dynamical systems*, Phys. D, 56 (1992), pp. 185–187.

[14] J. D. Farmer and J. J. Sidorowich, *Predicting chaotic time series*, Phys. Rev. Lett., 59 (1987), pp. 845–848.

[15] J. D. Farmer and J. J. Sidorowich, *Optimal shadowing and noise reduction*, Phys. D, 47 (1991), pp. 373–392.

[16] H. Furstenberg, *Stationary Processes and Prediction Theory*, Princeton University Press, Princeton, NJ, 1960.

[17] S. M. Hammel, *A noise reduction method for chaotic systems*, Phys. Lett. A, 148 (1990), pp. 421–428.

[18] S. M. Hammel, J. A. Yorke, and C. Grebogi, *Numerical orbits of chaotic processes represent true orbits*, Bull. Amer. Math. Soc. (N.S.), 19 (1988), pp. 465–469.

[19] R. W. Hamming, *Numerical Methods for Scientists and Engineers*, Dover, New York, 1986.

[20] K. Judd and L. Smith, *Indistinguishable states* I: *Perfect model scenario*, Phys. D, 151 (2001), pp. 125–141.

[21] K. Judd and L. A. Smith, *Indistinguishable states* II: *The imperfect model scenario*, Phys. D, 196 (2004), pp. 224–242.

[22] M. Kac, *On the notion of recurrence in discrete stochastic processes*, Bull. Amer. Math. Soc., 53 (1947), pp. 1002–1010.

[23] E. Kalnay, *Atmospheric Modeling, Data Assimilation and Predictability*, Cambridge University Press, Cambridge, UK, 2003.

[24] A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge University Press, Cambridge, UK, 1995.

[25] S. P. Lalley, *Beneath the noise, chaos*, Ann. Statist., 27 (1999), pp. 461–479.

[26] S. P. Lalley and A. B. Nobel, *Denoising deterministic time series*, Dyn. Partial Differ. Equ., 3 (2006), pp. 259–279.

[27] R. Narasimha, S. R. Kumar, A. Prabhu, and S. V. Kailas, *Turbulent flux events in a nearly neutral atmospheric boundary layer*, Philos. Trans. R. Soc. A, 365 (2007), pp. 841–858.

[28] D. S. Ornstein and B. Weiss, *Entropy and data compression schemes*, IEEE Trans. Inform. Theory, 39 (1992), pp. 78–83.

[29] T. Stemler and K. Judd, *A guide to using shadowing filters for forecasting and state estimation*, Phys. D, 238 (2009), pp. 1260–1273.

[30] G. Sugihara and R. M. May, *Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series*, Nature, 344 (1990), pp. 734–741.
[31] F. Takens, *Detecting strange attractors in turbulence*, in Dynamical Systems and Turbulence, Springer-Verlag, Berlin, New York, 1981, pp. 366–381.
[32] D. Viswanath, *Symbolic dynamics and periodic orbits of the Lorenz attractor*, Nonlinearity, 16 (2003), pp. 1035–1056.
[33] D. Viswanath, *The fractal property of the Lorenz attractor*, Phys. D, 190 (2004), pp. 115–128.
[34] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*, The MIT Press, Cambridge, MA; John Wiley & Sons, New York; Chapman & Hall, London, 1949.
[35] A. D. Wyner and J. Ziv, *Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression*, IEEE Trans. Inform. Theory, 35 (1989), pp. 1250–1258.
[36] L. S. Young, *What are SRB measures, and which dynamical systems have them?*, J. Stat. Phys., 108 (2002), pp. 733–754.