statistical assumptions, in particular the threshold autoregressive model of Tong and Lim [75] and the local linear model of Priestley [64,41,63].

The threshold autoregressive model is formally equivalent to local linear approximation with fixed disjoint neighborhoods. A "threshold" corresponds to a partition on a given coordinate. The most important lesson learned from thinking in deterministic terms is the scale of implementation needed to get good results. For example, Tong and Lim discuss a threshold on one of the variables, splitting the state space in half, and using a different linear map for each half. While this is certainly a major improvement over a single linear map, and introduces enough nonlinearity to reproduce phenomena such as limit cycles and chaos, it is clearly inadequate to approximate a general nonlinear transformation with any accuracy. When we use fixed disjoint neighborhoods we typically partition the state space into hundreds or thousands of parts, putting thresholds on all the state space variables, in order to get good results. Also, as seen in Figure (4), using overlapping neighborhoods makes a big improvement.

Priestley's local linear model is a generalization of the threshold autoregressive model, with a different procedure for determining the charts appropriate for a given state. Instead of imposing a metric as we do, or using fixed disjoint neighborhoods as Tong and Lim do, Priestley uses an algorithm that is similar to the Kalman filter to track the free parameters in time. They vary in time as a "random walk", changing according to a linear dynamical equation. His approach has several attractive features, but the use of a linear equation for the parameters is unduly restrictive for modeling general nonlinear transformations. Also, the assumption of continuity in time means that this approach will not perform well for discrete time maps. Priestley's method avoids the arbitrariness of choosing a particular metric, but only through a loss of flexibility in other respects.

The value of assuming that the randomness of a time series is caused by chaotic dynamics rather than a more conventional random process is that it gives a new perspective. The deterministic dynamical systems approach leads naturally to innovations such as trajectory segmenting or the use of an explicit metric. This makes it natural to go to higher order approximation schemes, which can lead to big improvements in accuracy, as our numerical work clearly demonstrates. The dynamical systems viewpoint is also essential for producing the error estimates described in Section 3. Ultimately, of course, all of these methods should be judged on their performance in real-world applications.

3 Scaling of Error estimates

In this section we estimate forecasting errors. The errors depend on properties of the dynamics, such as the attractor dimension D and the Lyapunov spectrum $\{\lambda_i\}$. They also depend on properties of the data set, such as the number of data points, N, and the signal-to-noise ratio S, as well as the extrapolation time T. The resulting scaling laws provide an *a priori* means of estimating the quality of forecasts, and they also suggest improvements for computing nonlinear statistical properties such as the fractal dimension, as discussed in Section 4.

Unless otherwise stated, in this section we will assume that time is discrete, and scaled so that $t = 0, 1, 2, \ldots$. We discuss some of the problems that arise when time is continuous in Section 3.3.

3.1 Dependence on number of data points

The approximation error generally depends on the number of data points. When the accuracy improves as a power law in N, as it generally does for local approximation schemes, the dependence on N is described by the order of approximation, which we defined in Equation (16) as the scaling exponent. It is clear that for good forecasts we want to make the order of approximation as large as possible. One way to achieve this is by using local polynomials. For example, in Figure (4), we use first and second degree local polynomial charts, using data generated by the sine map,

$$x_{t+1} = sin(\pi x_t), \tag{18}$$

where $-1 < x_t < 1$. Similarly, in Figure (5) we show the scaling behavior of the Mackey-Glass delay differential equation [55],

$$\frac{dx(t)}{dt} = -0.1x(t) + \frac{0.2x(t-\Gamma)}{1 + x(t-\Gamma)^{10}}, \tag{19}$$

with $\Gamma = 17$. The resulting slopes are roughly $\frac{(m+1)}{D}$, where m is the degree of the chart, and D is from previous independent calculations [24].

In general it is not always possible to achieve $q = m + 1$. For example, for the delay equation we were unable to improve our results significantly using cubic charts. Improving the order of approximation is the central problem in nonlinear modeling.

3.2 Dependence on extrapolation time

Naturally for a chaotic system errors depend strongly on the time that we attempt to extrapolate into the future. The rate at which errors grow depends on the way we make predictions. There are two choices: We can make *iterative forecasts* by fitting a model for $T = 1$ and iterating to make predictions for $T = 2, 3, \ldots$. Alternatively, we can make *direct forecasts* by fitting a new model for each individual T. On the surface direct forecasting might seem more accurate, since each model is "tailored" for the time it is supposed to predict, and there is no accumulation of errors due to iteration. In fact, as we shall show here, if the model is sufficiently accurate, the opposite is true. Approximation errors for iterative forecasting grow roughly according to the largest Lyapunov exponent λ_{max}, whereas for direct forecasts the errors grow as $q\lambda_{max}$.

To show this we must first introduce the new notion of higher order Lyapunov exponents.

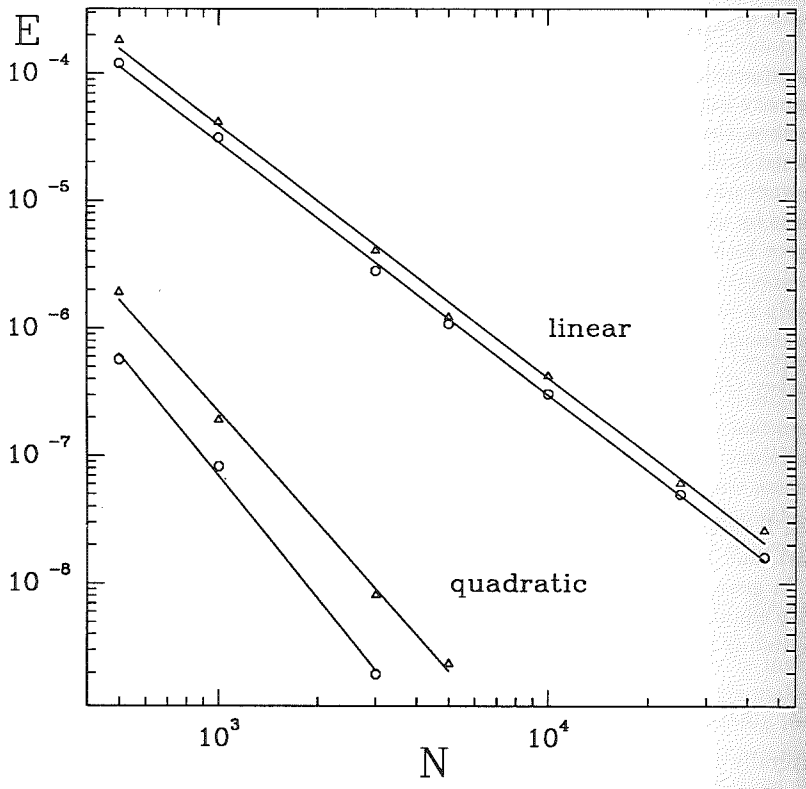

Figure 4: Local approximation by polynomials for the sine map. The slopes are roughly $-q = -(m+1)$, where m is the degree of the polynomial, and q is the order of approximation. For the data points shown with circles we used nearest neighborhoods for the predictions, and for those indicated by triangles we used disjoint neighborhoods, constructed with the k-d tree. .

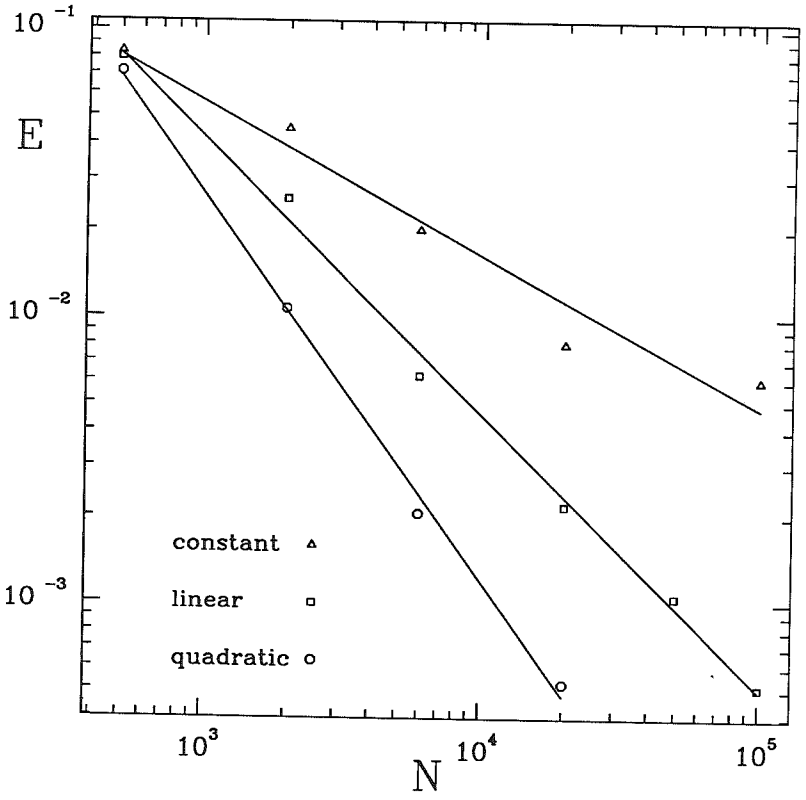

Figure 5: Local approximation by polynomials for the Mackey-Glass delay differential equation with $\Gamma = 17$, $\Delta t = 1$, $\tau = 6$, and $d = 4$ for a fixed extrapolation time $T = 85$. (The characteristic time is roughly 50.) Each value of $\bar{E}$ is based on an average of 500 forecasts. Using independently computed values for the attractor dimension D [24], we find that the measured value of $q \approx m + 1$, within the expected statistical error.

3.2.1 Higher order Lyapunov exponents

Since the accuracy of an approximation scheme depends on higher derivatives, for direct forecasting the growth rate of errors with time depends on the average growth rate of higher derivatives under iteration. This can be described in terms of a generalization of the Lyapunov exponents, which we describe in this section. For simplicity, we will only discuss the one dimensional case here, leaving the general case for a future paper [19].

Consider a one dimensional map, $x_{t+1} = f(x_t)$. Assume that f is analytic. We will define the q^{th} order Lyapunov exponent as

$$\lambda^{(q)} = \lim_{t\to\infty} \frac{1}{t} \log |x_t^{(q)}|, \tag{20}$$

where $x_t^{(q)}$ is the q^{th} derivative of the t^{th} iterate, $x_t^{(q)} = \frac{d^q f^t}{dx^q}(x_0)$. This reduces to the usual Lyapunov exponent when $q = 1$. Equation (20) can also be expressed in terms of the q^{th} order Lyapunov *number*

$$\Lambda^{(q)} = \lim_{t\to\infty} |x_t^{(q)}|^{\frac{1}{t}}.$$

For the remaining discussion we will assume that f is analytic. Furthermore, we will assume that f is ergodic with a natural measure and corresponding probability density function $P(x)$, so that we can rewrite Equation (20) as

$$\lambda^{(q)} = \lim_{t\to\infty} \frac{1}{t} \langle \log |x_t^{(q)}| \rangle = \lim_{t\to\infty} \frac{1}{t} \int \log |\frac{d^q f^t}{dx^q}(x)| P(x) dx \tag{21}$$

Using the ensemble average above, it is possible to interchange the average and the logarithm, to define a new set of exponents $l^{(q)}$[13]

$$l^{(q)} = \lim_{t\to\infty} \frac{1}{t} \log \langle |x_t^{(q)}| \rangle \tag{22}$$

It is clear that $l^{(q)} \geq \lambda^{(q)}$. For many examples, as a rough approximation we should find

$$l^{(q)} \approx \lambda^{(q)} \tag{23}$$

We will use this approximation to give us a rough idea of the relationship between first and higher order Lyapunov exponents, and also to estimate the scaling of our error estimates, acknowledging in advance that there are certainly examples for which this approximation breaks down.

When f is analytic the behavior of higher order Lyapunov exponents is at least approximately related to the first order Lyapunov exponent. To demonstrate this we will prove an equality for $l^{(2)}$ in terms of $l^{(1)}$.

[13]When $q = 1$ this exponent is what Fujisaka [32] calls the -1 order characteristic exponent. He generalizes the Lyapunov exponents in a different way, analogous to the definition of generalized dimensions [37].

We begin by recursively differentiating the map.

$$x_{t+1} = f(x_t) \tag{24}$$
$$x'_{t+1} = f'(x_t)x'_t \tag{25}$$
$$x''_{t+1} = f'(x_t)x''_t + f''(x_t)(x'_t)^2 \tag{26}$$

It is clear that the behavior of higher Lyapunov exponents is more complicated than that of the first order exponent. For example, $\lambda^{(1)} = \langle \log |f'(x)| \rangle_x$, but in general $\lambda^{(q)} \neq \langle \log |f^{(q)}(x)| \rangle_x$.

To get an intuitive feeling for Equation (26), suppose we neglect the first term on the right. This implies that for large t, x''_t grows roughly as the square of x'_t. However, since by definition $\langle |x'_t| \rangle$ grows as $e^{\lambda^{(1)}t}$, this suggests that the exponent for the second derivative is roughly twice that of the first.

We can prove a related inequality by making certain assumptions. Divide Equation (26) by $(x'_t f'(x_t))^2$, and rewrite it in terms of $y_t = x_t''/(x_t')^2$. After taking absolute values, averaging over x_0, and making the assumption that y_t and $f'(x_t)$ are uncorrelated, the result is

$$\langle |y_{t+1}| \rangle \leq \langle \frac{1}{|f'(x_t)|} \rangle \langle |y_t| \rangle + \langle \frac{|f''(x_t)|}{|f'(x_t)|^2} \rangle. \tag{27}$$

Let $a = \langle |f'(x_t)|^{-1} \rangle$, and $b = \langle \frac{|f''(x_t)|}{|f'(x_t)|^2} \rangle$, Equation (27) can be solved to give

$$\langle |y_t| \rangle \leq a^t \langle |y_0| \rangle + (\frac{1-a^t}{1-a}) b.$$

If we also assume that $|x'_t|^{-2}$ and $|x''_t|$ are uncorrelated, then we can write $\langle |y_t| \rangle = \langle |x''_t| \rangle \langle |x'_t|^{-2} \rangle$. Dividing and taking logarithms gives

$$\log \langle |x''_t| \rangle \leq -\log \langle |x'_t|^{-2} \rangle + \log(a^t \langle |y_0| \rangle + (\frac{1-a^t}{1-a}) b)$$

Since $\langle |x'_t|^{-2} \rangle > \langle |x'_t| \rangle^{-2}$ and there is a minus sign in front of the logarithm, we can make this substitution and preserve the inequality. Dividing by t gives

$$\frac{1}{t} \log \langle |x''_t| \rangle \leq \frac{2}{t} \log \langle |x'| \rangle + \frac{1}{t} \log(a^t \langle |y_0| \rangle + (\frac{1-a^t}{1-a}) b) \tag{28}$$

By definition for a chaotic mapping $\langle \log |f'(x_t)| \rangle = \lambda^{(1)} > 0$. Since $\log \langle |f'(x_t)| \rangle \geq \langle \log |f'(x_t)| \rangle$, this also implies that $\log \langle |f'(x_t)| \rangle > 0$. If we assume that $a < 1$, when we take the limit as $t \to \infty$ the term on the right vanishes, giving

$$\lim_{t\to\infty} \frac{1}{t} \log \langle |x''_t| \rangle \leq \lim_{t\to\infty} \frac{2}{t} \log \langle |x'| \rangle$$

Applying the definition of $l^{(q)}$ from Equation (22) gives

$$l^{(2)} \leq 2 l^{(1)} \tag{29}$$

A similar relationship holds for general q. If we assume that the approximation of Equation (23) is valid, then we have

$$\lambda^{(q)} \approx q\lambda^{(1)}. \tag{30}$$

For the numerical examples we have studied so far, this gives rough agreement. (See Figure (6) for example.)

There seems to be an analogy between higher order Lyapunov exponents and the generalized Renyi dimensions and entropies [37]. We intend to investigate this in more detail in the future [19], as well as possible connections to the higher order characteristic exponents of Fujisaka [32].

3.2.2 Direct forecasting

In this section we estimate the rate of growth of errors for direct forecasting, *i.e.*, constructing a new approximant $\hat{f}_T$ to approximate f_T at each time T. The central assumption is that the rate of growth of errors for q^{th} order approximation is dependent on the average rate of growth of the q^{th} derivative, which can be described in terms of the q^{th} order Lyapunov exponent.

To make this a little more concrete, it is probably worth explicitly demonstrating this for an example. Suppose that we use linear interpolation, and approximate a function F over an interval $[x_1, x_2]$ as

$$F(x) \approx \hat{F}(x) = F(x_1) + \left[\frac{F(x_2) - F(x_1)}{x_2 - x_1}\right](x - x_1). \tag{31}$$

The error that we make with this approximation is $E(x) = F(x) - \hat{F}(x)$. This can be estimated by expanding F in a Taylor's series about x_1, to get

$$E(x) = \frac{F''(x_1)}{2}(x - x_1)(x - x_2) + O(\epsilon^3),$$

where $\epsilon = x_2 - x_1$, and $O(\epsilon^n)$ indicates that the remaining terms are of order ϵ^n or smaller. The average absolute error is

$$\langle |E(x)| \rangle = \frac{1}{\epsilon}\int_{x_1}^{x_2} |E(x)|dx = \frac{1}{12}|F''(x_1)|\epsilon^2 + O(\epsilon^3) \tag{32}$$

Consider a one dimensional chaotic map, $x_{t+1} = f(x_t)$, (where the subscript now represents time). Again, assume this map is ergodic. Suppose we want to approximate the T^{th} iterate f^T using linear interpolation. If we use uniform knots ($\tilde{x}_i = \epsilon i$ in Equation (31)), from Equation (32) we get

$$\langle |E(x_{t+T})| \rangle = \frac{\epsilon^2}{12}\int |\frac{d^2 f^T}{dx^2}(x_t)|P(x_t)dx_t + O(\epsilon^3) = \frac{\epsilon^2}{12}\langle |x_T''| \rangle + O(\epsilon^3), \tag{33}$$

where $P(x_t)$ weights the errors on each individual forecast according to the frequency with which they occur. Taking logarithms and referring to Equations (22) and (23) shows that the mean error must scale as

$$\log \langle |E(x_{t+T})| \rangle_{x_t} \sim ql^{(1)}T \approx q\lambda^{(1)}T \tag{34}$$

For the numerical examples that we have studied this is a fairly good approximation.

3.2.3 Iterative forecasting

In this section we derive error estimates for iterated forecasting, *i.e.*, constructing an approximant for $T = 1$ and then iterating to predict $T = 2, 3, \ldots$. Assume that at $T = 1$ we approximate f by a map g. Define the error of approximation as

$$\delta(x) = g(x) - f(x) \tag{35}$$

where $|\delta(x)|$ is small and bounded. Similarly, define the approximation error at time T as

$$\Delta_T(x) = g^T(x) - f^T(x),$$

so by definition $\Delta_1 = \delta$. Assume that for any given T our approximation is good enough so that $\Delta_T(x)$ is small and bounded by $\Delta_{max} > |\Delta_T(x)|$ for all x. Thus, the scaling derived here will only be valid in the limit that the approximation is quite accurate.

From the previous two equations we get

$$\Delta_T(x) = f(g^{T-1}(x)) + \delta(g^{T-1}(x)) - f^T(x).$$

Expand $f(g^{T-1}(x))$ and $\delta(g^{T-1}(x))$ to first order in a Taylor series about $f^{T-1}(x)$. Assume that δ and f are both smooth, so that $f'' = O(1)$ and $\delta'' = O(1)$, where again $O(x)$ means "of the same order as x", and make use of $f' + \delta' = g'$. This implies

$$\Delta_T(x) = g'(f^{T-1}(x))\Delta_{T-1}(x) + \delta(f^{T-1}(x)) + O(\Delta_{max}^2) \tag{36}$$

By expanding this expression for $T-2, T-3, ...,$ it is clear that

$$\Delta_T(x) = \sum_{j=0}^{T-1}\prod_{i=j+1}^{T-1} g'(f^i(x))\delta(f^j(x)) + O(\Delta_{max}^2) \tag{37}$$

If we also assume that $\delta' = O(\Delta_{max})$, then we can approximate g' by f', and from the chain rule we can write the product of derivatives as the derivative of the iterate, so that this becomes

$$\Delta_T(x) = \sum_{j=1}^{T}\frac{df^{T-j}}{dx}(x_j)\delta(x_{j-1}) + O(\Delta_{max}^2). \tag{38}$$

Take absolute values and average over x, to get

$$\langle|\Delta_T(x)|\rangle = \sum_{j=1}^{T}\langle|\frac{df^{T-j}}{dx}(x_j)\delta(x_{j-1})|\rangle + O(\Delta_{max}^2). \tag{39}$$

It simplifies matters if we assume for the moment that $f(x)$ is a one dimensional map, and that $\frac{df^{T-j}}{dx}(x_j)$ and $\delta(x_{j-1})$ are uncorrelated. $\langle|\delta(x_j)|\rangle$ is independent of j. By definition (22) the average derivative $\langle|\frac{df^j}{dx}(x_0)|\rangle \approx L^j$, where $L = e^{\lambda^{(1)}}$. We can then rewrite this as

$$\langle|\Delta_T(x)|\rangle \approx \langle|\delta(x)|\rangle \sum_{i=0}^{T-1} L^i \tag{40}$$

Note that this is a natural result. It says that the cumulative error amplification from the first step is L^{T-1}, from the second step is L^{T-2}, *etc.* Summing the series gives

$$\langle|\Delta_T(x)|\rangle \approx \frac{L^T - 1}{L - 1}\langle|\delta(x)|\rangle. \tag{41}$$

In the limit of large T the asymptotic rate of growth is

$$\lim_{T\to\infty}\left(\frac{L^T - 1}{L - 1}\right)^{\frac{1}{T}} = L \tag{42}$$

To get an alternative view that gives a feeling for what these assumptions mean, we could return to Equation (36) and assume that

$$\delta(f^{T-1}(x)) \ll g'(f^{T-1}(x))\Delta_{T-1}(x). \tag{43}$$

Again approximating g by f and recursively substituting for Δ_{T-i}, this leads to $\Delta_T(x) = \frac{df^T}{dx}(x)\delta(x)$. Taking logarithms and absolute values, and assuming that $\delta(x)$ is uncorrelated with $\frac{df^T}{dx}(x)$, we get

$$\langle\log|\Delta_T(x)|\rangle = \langle\log|\frac{df^T}{dx}(x)|\rangle + \langle|\log\delta(x)|\rangle = T\lambda^{(1)} + \langle|\log\delta(x)|\rangle, \tag{44}$$

This agrees with previous estimates of Lorenz [54]. This also shows that the assumption of Equation (43) is roughly equivalent to the other assumptions we made leading up to Equation (42). It also suggests that the correspondence with Lyapunov exponents is more exact if we use $\langle\log E\rangle$ rather than $\log\langle E\rangle$ to evaluate the errors.

Either of these derivations depends on the assumption that $f' = O(1)$, $\delta' = O(\Delta_{max})$, and $\delta'' = O(1)$. This is not surprising: If the derivatives of the approximation are sufficiently different from those of the true dynamics, then in general the approximation will have different Lyapunov exponents. This should not be a problem for smooth approximation schemes, but this comes under question for methods such as linear interpolation, for which g is only C^1. For linear interpolation with a uniform knot spacing ϵ, it is fairly easy to show that $\delta' = O(\epsilon)$, which is larger than δ (which is $O(\epsilon^2)$), but still small. δ'' may cause problems, however, since it is a delta function. Intuitively, since we are taking an average, and the integral of the second derivative is $O(\epsilon)$, problems due to this seem unlikely, but this should be investigated in more detail. The safe case is when g and f are both smooth.

So far, we have restricted our discussion to one dimension, which is especially simple because there is only one Lyapunov exponent. In more than one dimension there can be more than one positive Lyapunov exponent, and the situation is more complicated. However, for long times, a displacement between an approximate trajectory and a true trajectory will typically line itself up along the most unstable direction, so that, as long as the largest Lyapunov exponent is sufficiently larger than the others, it will asymptotically dominate. Thus in higher dimensions we expect that the growth of errors will be dominated by the largest Lyapunov exponent. Note that this is revised from our previous paper [23]; although the metric entropy is related to the short term rate of loss of information, it does not seem to be the relevant quantity here. [14]

To conclude, for an iterative forecasting scheme satisfying the assumptions above we conjecture that the errors grow according to

$$\bar{E} \sim N^{-\frac{q}{D}}e^{\lambda_{max}T}, \tag{45}$$

where λ_{max} is the largest Lyapunov exponent. Comparing this to Equation (34), the difference is that the errors for iterative forecasts grow at an exponential rate given by λ_{max}, in contrast to $q\lambda_{max}$ for direct forecasts.

Intuitively, the superiority of iterative estimates must come from the fact that they make use of the regular structure of the higher iterates. The time series that we are trying to approximate are generated by iterating a dynamical system, and so iterative approximations are more natural. The power of the iterative procedure is reminiscent of Barnsley's methods for constructing complicated fractals from the recursive application of simple affine mappings [2]. Some preliminary results indicate the advantages of iterated forecasting in neural nets [49].

The validity of these estimates for a simple example is demonstrated in Figure (6), where we show the approximation error as a function of the extrapolation time T. As predicted, the error grows roughly according to Equation (34) for direct approximation and according to Equation (45) for iterative approximation.

For iterated forecasts we should expect that the distribution of errors will have long tails. As we have seen here, the cumulative error after many iterations is dominated by the product of the errors along the way. As long as the second moment exists, a corollary of the central limit theorem is that the probability density function of the product of many random numbers is log-normal. We expect that this will also be true for iterated forecasts, at least in the limit where we have a very accurate model and iterate many times. Log-normal distributions have long tails, corresponding to occasional very poor forecasts. Thus, if we are concerned with bounds on the worst case error rather than the mean square error, we might expect that iterated forecasts

[14] We would like to thank Martin Casdagli for pointing out this mistake.