PART VII
QUANTITATIVE MEASURES OF CHAOS

Several statistics, when reliably computed, indicate chaos (at least in theory). Moreover, by their magnitude, sign, or trend, they suggest (again, at least theoretically) how chaotic a system is. The most important of such statistics at present are Lyapunov exponents (Ch. 25), Kolmogorov-Sinai entropy (Ch. 26), and mutual information or redundancy (Ch. 27).

Chapter 25
Lyapunov exponents

A Lyapunov[1] ("lee-ah-poo-noff") exponent is a number that describes the dynamics of trajectory evolution. It capsulizes the average rate of convergence or divergence of two neighboring trajectories in phase space. Its value can be negative, zero, or positive. Negative values mean that the two trajectories draw closer to one another. Positive exponents, on the other hand, result from trajectory divergence and appear only within the chaotic domain. In other words, a positive Lyapunov exponent is one of the most important indicators of chaos. A positive Lyapunov exponent measures or quantifies sensitive dependence on initial conditions by showing the average rate (as evaluated over the entire attractor) at which two close points separate with time. Some authors refer to a Lyapunov exponent as a **Lyapunov characteristic exponent** (LCE) or simply a **characteristic exponent**. As we'll see, a dataset can have several Lyapunov exponents.

The origin of this important component of the chaos game seems to go back to a nineteenth century Russian mathematician, Sofya Kovalevskaya (1850-91). According to Percival (1989), Kovalevskaya in 1889 mathematically defined dynamical instability as an average of a measure of the rate of growth of small deviations.[2] Russian mathematician Aleksandr Lyapunov made the definition of dynamical instability more general in a lengthy treatise in 1892 (translated from Russian to French in 1907; both versions subsequently reprinted, e.g. Liapounoff 1949). Oseledec (1968) provided further theoretical basis.

---

[1]. **Lyapunov, Aleksandr Mikhailovich (1857-1918)** Lyapunov was one of the greatest Russian mathematicians. Born in Yaroslavl, he attended St Petersburg University in the late 1870s, where the famed P. L. Chebyshev had created a well known mathematics department. As a university senior in 1880, Lyapunov won a gold medal in a department-sponsored competition for his composition on hydrostatics. That composition was the basis for his first two publications, which appeared in 1881. He was awarded his Master's degree at St Petersburg in 1885, married his first cousin in 1886, and received his PhD in 1892. In 1895 he was appointed to the Chair of Mechanics at Kharkov University and began what was the happiest period of his life. In 1902 he returned to St Petersburg, having been elected in 1901 to the Chair of Applied Mathematics of the Russian Academy of Sciences. He now began a relatively secluded life, in which he placed much emphasis on his scientific work but also found time for his favorite hobby, plants (his apartment was decorated with rubber plants and palms which he had grown himself). During his productive career his general subject was mathematical physics—stability of motion, equilibrium of mechanical systems, potential theory, theory of probability, and the like. He is known as the founder of the modern theory of stability of motion. In 1918 his wife died of tuberculosis. Three days later Lyapunov himself died, leaving a note in which he asked to be buried in the same grave as his wife . . .

Local Convergence or Divergence of Neighboring Trajectories

The simplest way to define a Lyapunov exponent and show its properties is with the one-dimensional map. (As we'll see later, a system has as many Lyapunov exponents as there are dimensions in its phase space. A one-dimensional case provides only one exponent. We'll eventually extend our one-dimensional case to many dimensions [many Lyapunov exponents].) Let's use our faithful friend, the logistic equation (Eq. 10.1): $x_{t+1}=kx_t(1-x_t)$. In particular, we'll examine the progressive displacement or gap, over time, between two trajectories that start close together (Fig. 25.1).

Trajectory convergence

Chapter 10 showed that, with the logistic equation and k < 3 (hence in the nonchaotic domain), all trajectories for a given k value converge to the same fixed point. Let's inspect that convergence more closely. First we'll generate a so-called reference trajectory by choosing an appropriate k value and a value for x_0 then iterating the equation many times. (Some chaologists call a reference trajectory a **fiducial trajectory**; **fiducial** means "used as a standard of reference for measurement or calculation.") Armed with a reference trajectory, we'll then generate a nearby trajectory by starting the iterations over again, using the same k value but a slightly different x_0. Then we'll compare those two trajectories. In particular, for each iteration (each "new time"), we'll compute the difference δ (Greek delta) between the two predicted values of x_{t+1} (the difference between the two converging trajectories). Our goal is to see how the difference or displacement varies with time.

Let's begin with k=0.95. All logistic-equation trajectories for k<1 go to a point attractor at the origin (Ch. 10). Say we start the reference trajectory at x_0=0.40 and the other trajectory at x_0=0.41. The absolute difference (hereafter simply called difference) between those two values, δ_0, is 0.41-0.40=0.01. For our comparison of displacement versus time, therefore, the first datum point is δ_0=0.01 at t=0.

For the first iteration of the logistic equation with k=0.95, the reference trajectory starts at x_0=0.40 and goes to an x_{t+1} of 0.228. For the nearby trajectory, using that same k value but with x_0 = 0.41, x_{t+1}=0.229805. The difference between those two calculated values is 0.229 805-0.228=0.001805. So, the second datum point is a trajectory separation or gap of δ_1=0.001805 at t=1. Next we continue the procedure for subsequent iterations. Still holding k constant at 0.95, the difference in calculated values for the second iteration (t=2) is δ_2 = 0.00092973. For the third iteration, the displacement is δ_3=0.00058704, and so on.

---

2. At the University of Stockholm in 1884, Kovalevskaya became Professor of Mathematics—probably the first woman ever to do so at a European university. For a fascinating account, see Kovalevskaya (1978).

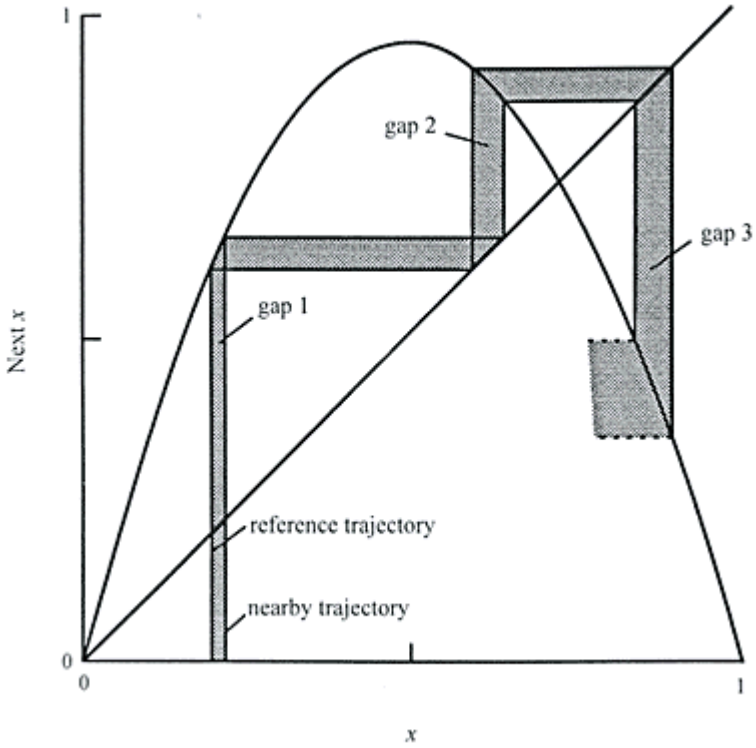


Figure 25.1
Logistic equation (parabola) and hypothetical evolution of two trajectories that start close together. In this example the trajectories diverge with time.

Plotting separation δ on the ordinate against time shows that, except for transients, the calculated differences between two neighboring trajectories for our iterations plot as a straight line on semilog paper (Fig. 25.2). The particular type of semilog paper has the arithmetic scale on the abscissa and the log scale on the ordinate. Any straight line on that type of graph paper has the general form $y = y_0 e^{bx}$, where y_0 is the extrapolated value of the dependent variable y at x=0, e is the base of natural logarithms (a constant, 2.718 . . .), b is a constant representing the slope of the straight line on the graph, and x is the independent or given variable. If the straight line slopes upward, b is positive; if it slopes downward, b is negative. This type of relation is an exponential relation, as defined in the chapter on chaotic equations.

The given variable x here is time (iteration number, n). The dependent variable y is the separation distance δ between the two trajectories. Extrapolating the straight line back to n=0, we'll symbolize the intercept as δ_a and the finite value of δ at a particular time as δ_n ("delta n"). The general form of the exponential equation for the straight-line part of the plot therefore is:

$$\delta_n = \delta_a e^{bn}. \qquad (25.1)$$

As our calculated values show, the gaps between trajectories for k=0.95 become smaller (converge) with time. Consequently, the straight line in Figure 25.2 slopes downward (slope b is negative). Because the plotted points show virtually no scatter at all about the line, we can draw the straight line by eye. Its slope as measured graphically is any arbitrarily chosen ordinate distance divided by the associated abscissa distance. For the example of Figure 25.2, the slope measured in that fashion is about -0.05.

Let's now go to a larger k value (say, 2.8) and again iterate the logistic equation for two neighboring trajectories. We'll start the trajectories at the same two x values as before (0.40 and 0.41). Within the region 1< k<3, all trajectories are still in the nonchaotic domain and go to a fixed-point attractor. However, the fixed point for this range of k isn't at zero. Figure 25.3 plots difference in calculated values (δ_n, on a log scale) versus time. Again the gaps get smaller with time. Comparing Figure 25.3 with Figure 25.2 shows that with k=2.8 the slope of the straight line (the local rate of convergence of the two trajectories) again is negative. In this case the slope is now about -0.22. (On a minor note, with k=2.8 the points began falling on a straight line after only about five or six iterations.)

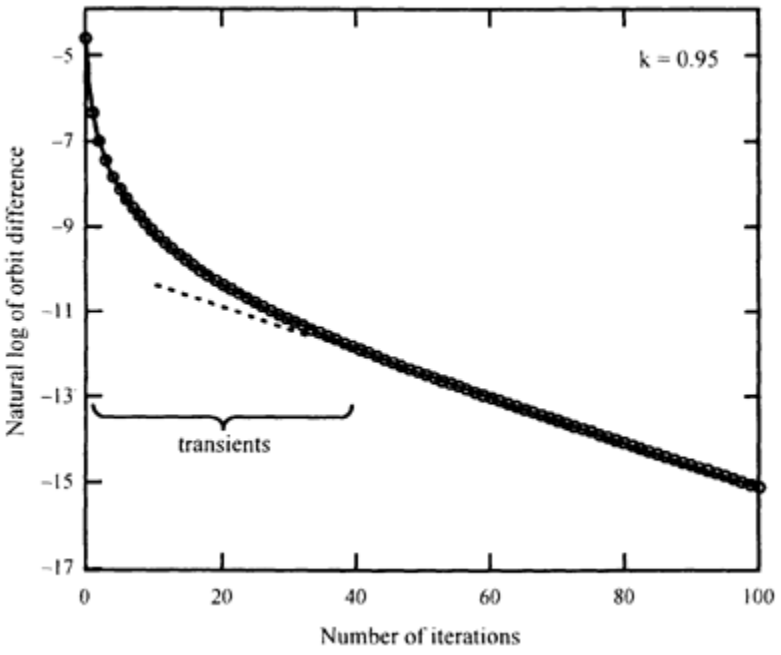


Figure 25.2
Graph of amount of separation of two neighboring
trajectories over time for the logistic equation, with k = 0.95
(hence thenon-chaotic domain with a fixed-point attractor).

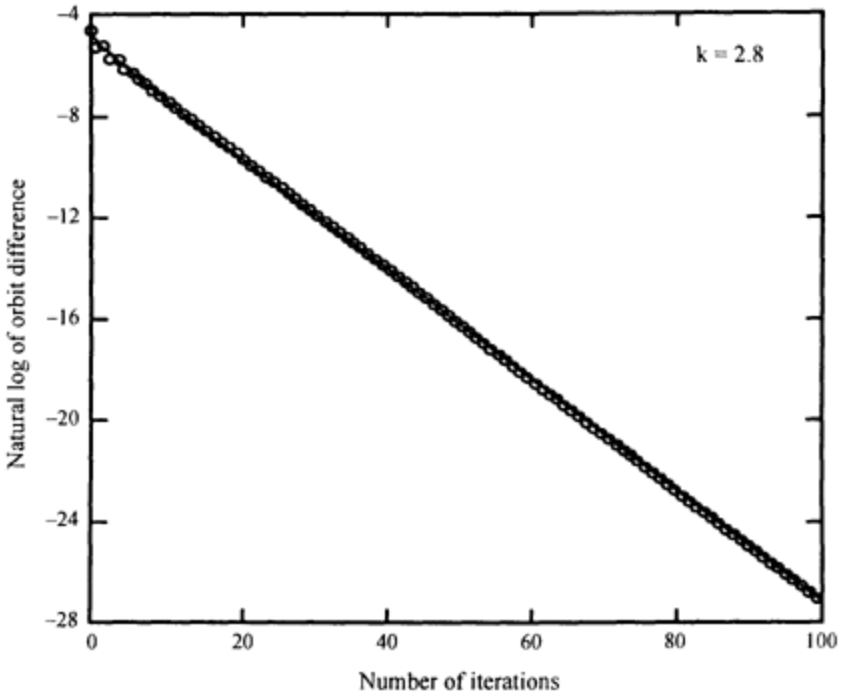


Figure 25.3
Graph of amount of separation of two neighboring
trajectories over time for the logistic equation, with k = 2.8
(hence the nonchaotic domain with a fixed-point attractor).

Now we'll increase k to 3.4 and start the two trajectories at x_0 values of 0.060 and 0.062. Chapter 10 showed that the range 3< k<3.57 produces period-doubling (hence no chaos) and that k=3.4 corresponds to an attractor of period two. The two alternating values of the attractor are approximately 0.45196 . . . and 0.84215 . . . . That is, each trajectory first approximates one attractor value, say 0.45. Then, on the next iteration, they approximate the other (0.84). They also get closer and closer to the true value with time (number of iterations). So, we have two meaningful comparisons at a periodicity of two. One comparison evaluates the gap between the two neighboring trajectories near the lower of the two attractor values (0.45 . . .). The other looks at the difference between the two trajectories near the other attractor value (0.84 . . .). It turns out that the gap associated with the lower attractor value is greater than that associated with the higher attractor value, for any approximate time (any pair of successive iterations). The plot of trajectory separation versus time (Fig. 25.4) therefore yields two separate straight lines—a higher one for the gap near the attractor value of 0.45196 . . ., and a lower one for the gap near the attractor value of 0.84215 . . . . Both lines slope downward at about -0.14.

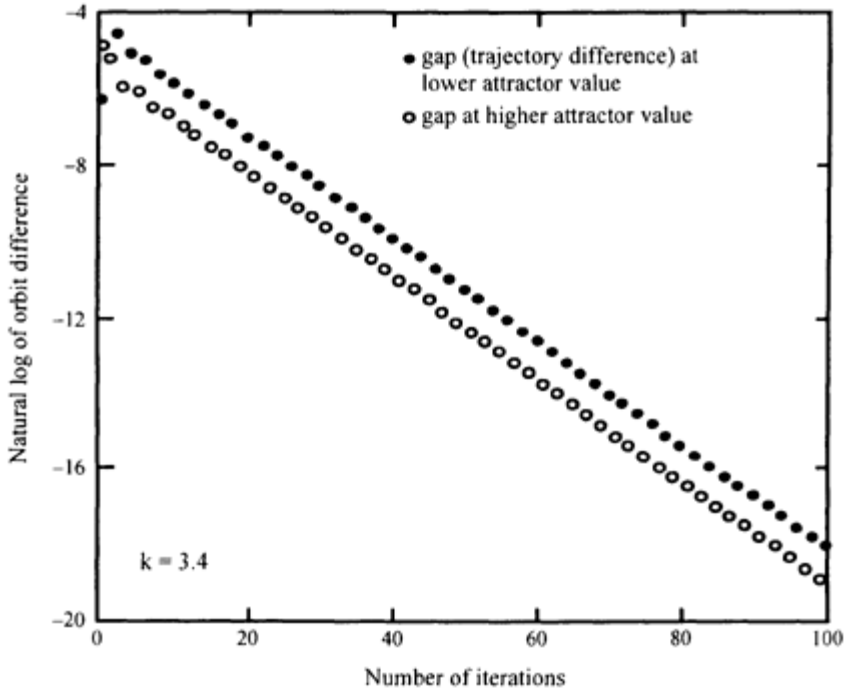


Figure 25.4
Graph of amount of separation of two neighboring trajectories over time for the logistic equation, with k = 3.4 (nonchaotic domain, period two).

Three important features about all of the above nonchaotic examples are:

- The gap between neighboring trajectories has followed an *exponential* law in regard to time.
- The neighboring trajectories have *converged* toward their mutual attractor.
- The convergence means that gaps decrease with time, so *the slope of the fitted line is negative*.

Trajectory divergence

Sticking with the logistic equation as our model, let's now look at trajectory divergence. Divergence means that the gaps get larger with time. Neighboring trajectories with the logistic equation (and in fact with all dynamical systems) diverge only in the chaotic domain. Gaps as the two trajectories *approach* their mutual attractor are now less important; instead, we monitor those differences once the two trajec tories are "on" their mutual attractor. Let's compute some differences and see how they change with time, as before.

We'll choose k and x_0 and begin iterating Equation 10.1, as usual. To make sure of getting "onto the attractor," we first iterate long enough to get past the transients. Let's say it takes about 100 iterations to get onto the attractor. We'll designate the 100th point as the new x_0 (a point on the attractor). We also arbitrarily choose a different but nearby x_0, as the starting point of a neighboring trajectory. (Strictly, we only know with certainty that just the first of those trajectories is approximately on the attractor, assuming our 100 iterations were enough. So, we'll assume that the other trajectory, beginning such a minute distance away, also is approximately on the attractor.) Finally, we determine the differences in the computed x_{t+1} (the amount of separation of the two trajectories) for each successive iteration and plot those differences against time, as before.

Suppose k=3.8 and we begin iterating the reference trajectory at an x_0 of 0.06. On my computer the 100th iteration of the logistic equation (Eq. 10.1) gives x_{t+1}=0.72035971. We assume that the trajectory is now on the attractor and take that value as a new x_0. If we choose a small and arbitrary difference δ_0 of 0.002, then the x_0 for the nearby trajectory begins at 0.72035971+0.002 or 0.72235971. (The number of decimal places here also is arbitrary.) We then calculate the trajectories that start at each of these x_0 values (0.72035971 and 0.72235971). Figure 25.5 shows computed differences in x_{t+1} plotted against time for 100 iterations. Figure 25.5 looks radically different from its trajectory-convergent counterparts (Figs 25.2-4). The three major differences are:

- The computed gap between trajectories changes in a systematic way only for a short time (about ten or so iterations in this case); thereafter, it becomes irregular. So, the divergence rate only applies to an early and restricted range along a trajectory. The irregular behavior beyond that range corresponds to the stage in which sensitivity to initial conditions has had such a cumulative and major effect that the two trajectories go off in totally independent, unrelated directions. (They may occasionally pass close to one another again, but that's purely by chance.)

- The general trend of systematic change (here over the first ten or so iterations) now slopes *upward* (has a positive value) instead of downward (negative value). The upward slope indicates that the gap between the two trajectories increases with time rather than decreases. The reason is sensitivity to initial conditions, the hallmark of the chaotic attractor. (And again, "initial" here means any place on the attractor where we might choose to begin monitoring the divergence of two neighboring trajectories.)

- There's now some scatter about the best-fit line. Hence, the straight line only represents the "average" rate of change.

The last deviation δ_n that adheres to the straight line represents the end of systematic divergence and is important in a practical way: it marks the approximate limit of predictability. For the time over which the systematic (exponential) divergence holds, the measured slope enables us to predict displacements with reasonable accuracy. However, once the regular separation ends we might as well select values indiscriminately from within the range of possible values. In other words, short-term predictions are relatively reliable, but long-term predictions (defined as those for times later than the end of the systematic divergence on the graph) are meaningless. Furthermore, that's true even when we know the evolution equations exactly.

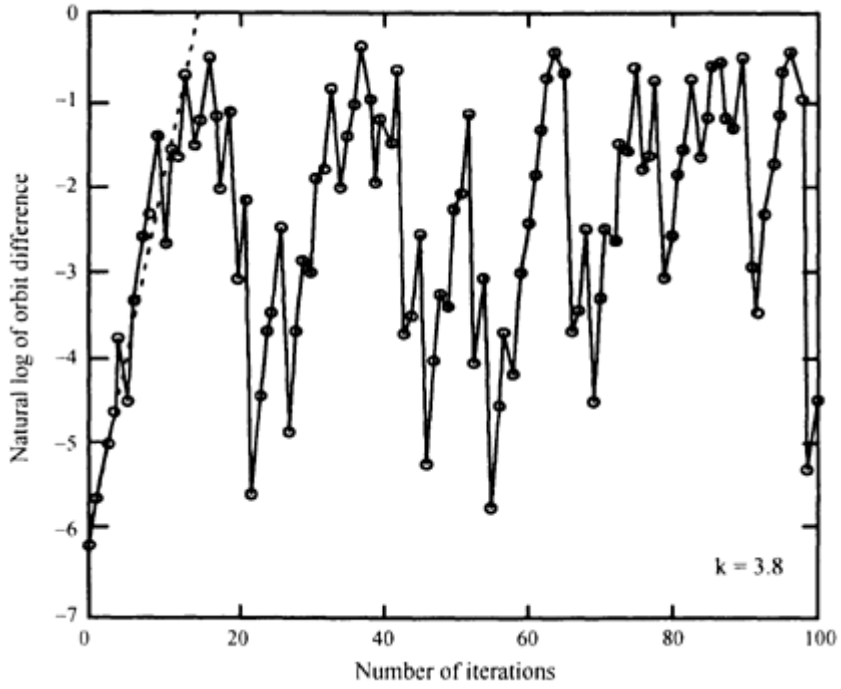


Figure 25.5
Graph of amount of separation of two neighboring trajectories over time for the logistic equation, with k = 3.8 (chaotic domain, transients excluded, starting gap 0.002).

The predictable range for real-world data can vary widely, from a fraction of a second to millions of years. Weather predictability, for example, is on the order of a few days, at best. Predictability of the orbits of planets in the solar system (excluding Pluto) seems to be about 4-5 million years (Laskar 1989, Sussman & Wisdom 1992).

The notion of predictability we're discussing here, as well as most other nonlinear prediction methods, rests on a key assumption. That is that, for whatever set of conditions (values of the variables) a system might be in, it always responds in the same way (always goes to the same new set of values). In other words, the assumption is that phase space trajectories for such a system always follow approximately the same local route.

What if we had chosen a smaller starting gap, δ_0? For the same k (3.8) and with x_0 for the reference trajectory again at 0.72035971, suppose δ_0=0.00000001, instead of 0.002. (Such a radically smaller starting displacement might correspond, for example, to a much better measuring capability.) Using the smaller separation distance makes x_0 for the neighboring trajectory equal to 0.72035972. Subsequent computations now produce the graph shown in Figure 25.6. The smaller δ_0 is the cause of all differences from Figure 25.5. Slopes of the straight lines are not too dissimilar—0.42 on Figure 25.5, 0.48 on Figure 25.6. However, the straight line (range of systematic divergence and hence of somewhat reliable predictability) lasts much longer on Figure 25.6, namely to nearly 40 iterations. In other words, better accuracy on our measurements enables us to predict further into the future.

Other examples where tiny differences increase exponentially (although I didn't describe them as such at the time) appeared in the discussion of "Computer Inaccuracy and Imprecision" (e.g. Figs 14.2, 14.3). There I ignored the concept of transients (i.e. I used the first iterates), and yet the results weren't affected. Discarding transients therefore may not always be necessary.

In summary, the constant b in the equation $\delta_n=\delta_a e^{bn}$ reflects the local exponential rate at which trajectories converge (negative exponent b) or diverge (positive exponent b). The gap between neighboring trajectories can also remain constant with time (as on a torus), in which case the constant b is zero.

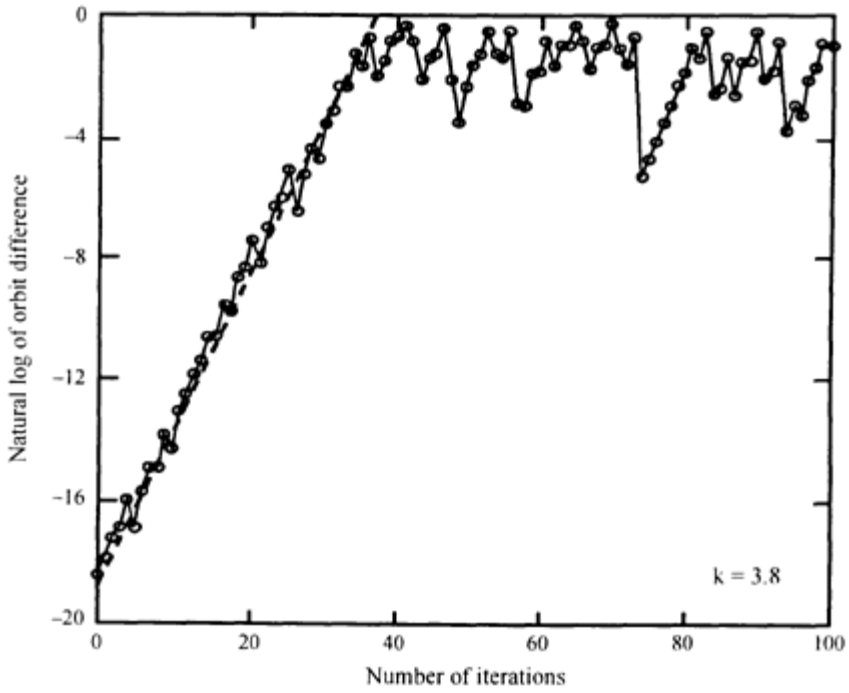


Figure 25.6
Graph of amount of separation of two neighboring trajectories over time for the logistic equation, with k = 3.8 (chaotic domain, transients excluded. starting gap 0.00000001).

Variability of Local Convergence and Divergence Rates Over the Attractor

Does the slope b vary with the two trajectories' location on the attractor? (Because a trajectory visits successive locations as time progresses, examining different locations on the attractor is tantamount to studying effects over time.) Continuing with the logistic equation, we'll fix k at 2.8 and conduct five separate experiments, corresponding to reference-trajectory starting values x_0 of 0.05, 0.20, 0.50, 0.80, and 0.95, respectively. In each case, we'll set δ_0 as 0.01. Starting values of the nearby trajectories therefore become 0.06, 0.21, 0.51, 0.81, and 0.96, respectively.

For all five experiments at k=2.8, the rate of convergence (slope b) turns out to be the same, namely -0.22 (the plots aren't shown here). Therefore, the local rate of convergence seems to be the same, at least for k=2.8 (a nonchaotic case).

Neighboring trajectories on chaotic attractors, on the other hand, don't necessarily behave so consistently. Slopes for the Lorenz attractor (Fig. 18.8), for example, range from -15 to +16, depending on the particular location on the attractor and other factors (Nese 1989). Two adjacent trajectories along the tops of the attractor's lobes can even converge, producing negative values of the slope b for that small zone. In contrast, adjacent trajectories monitored at the bottoms of the two lobes separate rapidly, yielding large and positive values of b. Therefore, we'll have to start referring to a "local slope" (a value of b that pertains to a particular local region on an attractor) from now on.

The divergence or convergence of neighboring trajectories at different rates depending on phase space location complicates chaos analyses. Nicolis et al. (1983) proposed to account for that variability with a **non-uniformity factor** (NUF). The NUF treats local slope as a statistical variable because it varies from place to place on a chaotic attractor (even though constant at any small region, per Eq. 25.1). The many different values can be compiled into a frequency distribution. The Nicolis et al. NUF is simply the standard deviation of that distribution of local-slope values. Thus, the NUF is a global indicator reflecting the spread of the local slopes about the mean local slope.

Local Lyapunov Exponents

The exponent b—the exponential rate of trajectory convergence or divergence as derived for any one local region on an attractor—is a so-called local **Lyapunov exponent**. It characterizes a system's dynamics for that particular region, that is, over the associated brief timeframe. Because those local dynamics and hence expo nents can vary with location on the attractor, it's important always to attach the adjective "local" when referring to such exponents. They aren't the same as the standard Lyapunov exponent that we're going to derive in the next section. A few representative papers dealing with local Lyapunov exponents include Fujisaka (1983), Nicolis et al. (1983), Grassberger & Procaccia (1984), Nese (1989), and Abarbanel (1992); see also Abarbanel et al. (1991a) and Abarbanel et al. (1992).

Local Lyapunov exponents can suggest attractor regions of potentially greater short-term predictability. For one thing, the duration over which the exponential relation holds tells us something about the time limits on predictability. Also, the value of the exponent reveals something about the reliability of predictions. For instance, a large and positive local Lyapunov exponent (fast divergence of neighboring orbits) suggests great sensitivity to initial conditions and poor predictability. Negative local exponents, in contrast, suggest convergence of neighboring trajectories and hence relatively good predictability. Measuring—and separately evaluating—local exponents over the entire attractor can tell us how they compare to the global average, their variability (which can be large) over the attractor, the special conditions for which predictions are reliable or unreliable, and many other things.

Definition of Lyapunov Exponent

It's convenient to have a single value that represents all the local slopes over the attractor, that is, over time. Understandably, people adopted an average for that representative number. That average goes by the name of the first or largest Lyapunov exponent.[3] Now we're going to develop the basic exponential equation into an expression that represents an average local slope for the entire attractor. This will take several minutes, but it'll be worth it.

We start with Equation 25.1 ($\delta_n=\delta_a e^{bn}$). For simplicity, let's assume that there are no transients, so that $\delta_a=\delta_0$. Inserting δ_0 for δ_a and dividing both sides by δ_0:

$$\frac{\delta_n}{\delta_0} = e^{bn}. \qquad (25.2)$$

Equation 25.2 highlights the ratio of the *n*th gap δ_n to the starting gap δ_0. That important ratio (the left side of the equation) is the factor by which the initial gap δ_0 gets stretched or compressed over the *n* iterations. Figure 25.7 shows a starting gap (indicated by δ_0 parallel to the abscissa) and a stretching over one iteration, in pseudo phase space. The reference trajectory in Figure 25.7 starts at x_0. The neigh boring trajectory also starts at time zero but at a distance δ_0 away. In symbols, the neighboring trajectory therefore starts at a value of $x_0+\delta_0$, as with our earlier examples. Values as of the next iteration (time one) are plotted on the ordinate. Those values are indicated by subscript 1. At time one the reference trajectory has a value x_1, the gap has become δ_1, and the neighboring trajectory has a value $(x_0+\delta_0)_1$.

---

[3]. "Exponent" applies because each local slope b is an exponent in the basic exponential equation, $\delta_n=\delta_a e^{bn}$. Also, as mentioned early in the chapter, there are usually several Lyapunov exponents. namely as many as there are dimensions in the phase space. We'll get to that topic in a later section.

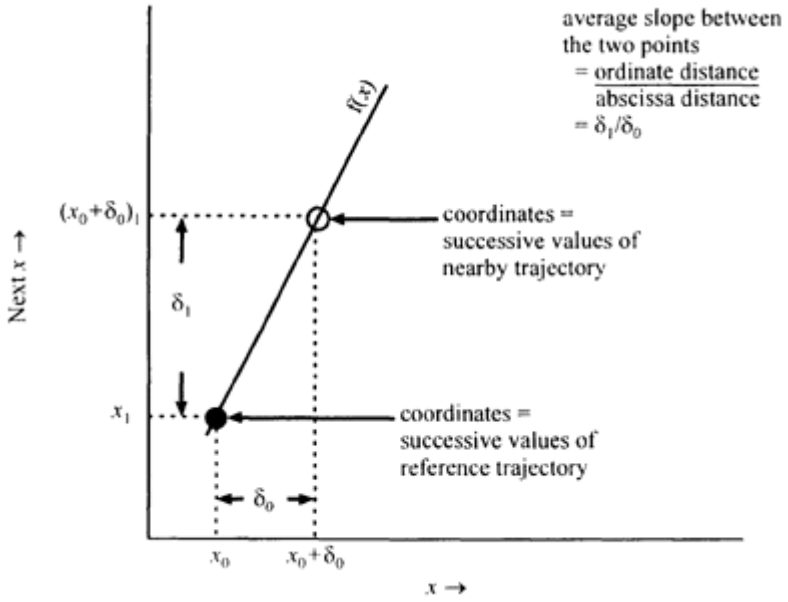


Figure 25.7
One-dimensional map (pseudo phase space plot) showing hypothetical progression of a gap between two adjacent trajectories over one time step. Gap at time 0 is d_0, and gap at time 1 is d_1. The gap changes (here grows) at a rate given by the function f(*x*).

The important quantity is the rate at which the separation distance changes. We'll designate the nature of the change by an undefined general rule that we'll call a "function of *x*." We write such a function in symbols as f(*x*) (called "f of *x*"). In Figure 25.7, the rate at which f(*x*) changes is approximated by the ratio δ_1/δ_0. Graphically, the ratio δ_1/δ_0 is an ordinate distance δ_1 divided by an abscissa distance δ_0 (Fig. 25.7). In other words, it's the average slope of the curve or straight line that connects the two sequentially plotted points. (The slope δ_1/δ_0 isn't the same as our local slope b. However, the two slopes are mathematically related in some way, as Eq. 25.1 implies.)

Mathematicians symbolize the nth iterate of any function f of *x* as $f^n(x)$. (Putting *n* in the position of an apparent exponent is an unfortunate convention, as it looks like a power. Remember, it's not a power when located right after the f; instead, it's just an identifying label.) To show that iterating the function produces *x*, they just put "*x* =" in front of the expression $f^n(x)$, yielding $x=f^n(x)$. Also, the applicable iteration number (1, 2, etc.) replaces *n*.

After n iterations the ratio of gap δ_n divided by original gap δ_0 gives an average rate of change over those n iterations. Now let's make the initial gap (δ_0) smaller and smaller (approaching zero). (In symbols, $\delta_0 \rightarrow 0$.) That means the second plotted point on Figure 25.7 approaches the first point. The computed rate of change for that condition is called a **first derivative** (or simply a derivative).[4] For that special limiting condition ($\delta_0 \rightarrow 0$), mathematicians replace δ with a lower-case d. The d still refers to the gap, just like the δ. To make sure everyone knows that the gap is in terms of x, they write the gap as dx.

Sometimes the gap is expressed in terms of some function such as f or $f^n(x)$, in which case it's written df or $df^n(x)$, respectively. Hence, df or $df^n(x)$ indicates the discrepancy (e.g. the "ordinate gap" in Fig. 25.7) between two successive iterations as δ_0 becomes infinitely small. The relative change for the special condition of δ_0 becoming infinitely small then is no longer simply δ_1/δ_0; instead, that change (ordinate distance/abscissa distance) becomes df/dx. Since our function f is $f^n(x)$, the df/dx is $df^n(x)/dx$. Mathematicians write $df^n(x)/dx$ as

$$\frac{d}{dx} f^n(x).$$

They call it the "derivative of the nth iterate of the function f with respect to x." (Again, a first derivative is merely a slope or relative change, measured at the limit where the starting gap δ_0 becomes infinitely small. It's therefore analogous to ratios such as δ_3/δ_0, δ_n/δ_0, and so on.) Summarizing: the rate of change as $\delta_0 \rightarrow 0$ is

$$\textit{rate of change} = \frac{d}{dx} f^n(x) = \lim_{\delta_0 \rightarrow 0} \frac{\textit{ordinate distance}}{\textit{abscissa distance}} = \lim_{\delta_0 \rightarrow 0} \frac{\delta_n}{\delta_0}. \qquad (25.3)$$

As δ_0 gets smaller, the calculated rate for the first iteration (δ_1/δ_0) applies closer and closer to the point that started at x_0 (Fig. 25.7). Hence, we assign δ_1/δ_0 to the value x_0. Also, we'll adopt another standard symbol and indicate the derivative of the function f (i.e. the value df/dx) by f'(x) (spoken as "f prime of x"). The relative change for the *first* iteration therefore is

$$\textit{rate of change at point } x_0 = \frac{\delta_1}{\delta_0} = f'(x_0). \qquad (25.4)$$

For the next iteration, value x_1 takes the role of x_0, gap δ_1 takes the place of old gap δ_0, and new gap δ_2 takes the place formerly occupied by δ_1. Following the same pattern as in Equation 25.4, we symbolize the relative change associated with point x_1 as f'(x_1) and calculate it as δ_2/δ_1:

---

[4]. A **derivative** is a value that derives from, comes from, or is traceable to a particular source, in this case the point on the curve.

$$\textit{rate of change at point } x_1 = \frac{\delta_2}{\delta_1} = f'(x_1). \qquad (25.5)$$

(So, with derivatives written as f'(x_0), f'(x_1), and so on, the value within parentheses (x_0, x_1, etc.) indicates where, along the curve, to measure the rate of change.)

We now invoke a handy rule known as the **chain rule**. It defines the derivative of the function f, as taken at the value x_0. That is, it defines

$$\frac{d}{dx} f^n(x_0).$$

In particular, it defines that quantity as the product of two terms. One is the derivative of the function taken at the value of the preceding iteration x_{n-1}, which in symbols is f'(x_{n-1}). The other is the derivative of that preceding iterate of the function taken at the value x_0, or

$$\frac{d}{dx} f^{n-1}(x_0).$$

Writing the definition in symbols,

$$\frac{d}{dx} f^{n}(x_0) = f'(x_{n-1}) \cdot \frac{d}{dx} f^{n-1}(x_0). \qquad (25.6)$$

We're going to apply the chain rule to a series of iterates or successive observations. As an example, let's apply it to four sequential observations. That means writing Equation 25.6 four times, namely for n=4, n=3, n=2, and n=1. However, a term for n=1 will be part of the expression for n=2 and so doesn't have to be written separately. Hence, we only have to write Equation 25.6 three times, as follows:

Equation 25.6 with n=4 is

$$\frac{d}{dx} f^{4}(x_0) = f'(x_3) \cdot \frac{d}{dx} f^{3}(x_0). \qquad (25.7)$$

Equation 25.6 with n=3 is

$$\frac{d}{dx} f^{3}(x_0) = f'(x_2) \cdot \frac{d}{dx} f^{2}(x_0). \qquad (25.8)$$

Equation 25.6 with n=2 is

$$\frac{d}{dx} f^{2}(x_0) = f'(x_1) \cdot \frac{d}{dx} f(x_0). \qquad (25.9)$$

The term on the far right in Equation 25.9, namely

$$\frac{d}{dx} f(x_0),$$

is an alternate way of writing f' (x_0), so Equation 25.9 really is

$$\frac{d}{dx} f^{2}(x_0) = f'(x_1) \cdot f'(x_0). \qquad (25.10)$$

Equations 25.7,25.8, and 25.10 have some common terms. Hence, we can make substitutions. For instance, the last term in Equation 25.7 is

$$\frac{d}{dx} f^{3}(x_0),$$

and that term is also defined by Equation 25.8. So, we can insert the right side of Equation 25.8 into the last term of Equation 25.7. Equation 25.7 then becomes

$$\frac{d}{dx} f^{4}(x_0) = f'x_3 \cdot f'x_2 \cdot \frac{d}{dx} f^{2}(x_0). \qquad (25.11)$$

Again the term on the far right,

$$\frac{d}{dx} f^{2}(x_0),$$

has been defined separately (Eq. 25.10). Making that substitution, Equation 25.11 becomes

$$\frac{d}{dx}f^4(x_0) = f'(x_3)\cdot f'(x_2)\cdot f'(x_1)\cdot f'(x_0)$$

$$= \text{the product of the } f'(x_i)\text{'s} \qquad (25.12)$$

where x_i is a general symbol for the successive values of the iterations and i here ranges from zero to n-1. (And so, with our little group of equations, the last term in each equation is linked, as in a chain, to the definition given by the next equation.) Equation 25.12 says that the quantity

$$\frac{d}{dx}f^n(x)$$

is just the product of all the first derivatives, that is, the product of the rates of change $f'(x_i)$, for i ranging from zero to n- 1.

We can write the derivatives on the right-hand side of Equation 25.12 as ratios of gaps, as with Equations 25.4 and 25.5. For instance, $f'(x_0)=\delta_1/\delta_0$, per Equation 25.4. Continuing with our example of n=4, we therefore substitute into Equation 25.12 the ratios δ_1/δ_0 for $f'(x_0)$, δ_2/δ_1 for $f'(x_1)$, δ_3/δ_2 for $f'(x_2)$, and δ_4/δ_3 for $f'(x_3)$:

$$\frac{d}{dx}f^4(x_0) = \frac{\delta_4}{\delta_3}\cdot\frac{\delta_3}{\delta_2}\cdot\frac{\delta_2}{\delta_1}\cdot\frac{\delta_1}{\delta_0}.$$

Cancelling like terms from the top and bottom on the right-hand side leaves

$$\frac{d}{dx}f^4(x_0) = \frac{\delta_4}{\delta_0}$$

or, in general,

$$\frac{d}{dx}f^n(x) = \frac{\delta_n}{\delta_0} \qquad (25.13)$$

where the value of n reflects the number of derivatives or ratios δ_i/δ_{i-1} involved. For instance, with n=4 there are four derivatives, namely the ratios δ_4/δ_3, . . . δ_1/δ_0.

Equations 25.12 and 25.13 together say that

$$\frac{\delta_n}{\delta_0} = \text{the product of the } f'(x_i)\text{'s}. \qquad (25.14)$$

Equation 25.14 brings our story full circle back to Equation 25.2, $\delta_n/\delta_0=e^{bn}$ (with $\delta_a=\delta_0$). The ratio on the left (δ_n/δ_0), according to Equation 25.14, represents the product of a bunch of derivatives. Next we'll rearrange our transientless version of Equation 25.2 to isolate the exponent b, the separation rate of two neighboring trajectories. To get b, we first put Equation 25.2 into log form (base e):

$\log_e(\delta_n/\delta_0)=bn\log_e e$.

The ingredient on the far right, $\log_e e$, equals 1, so

$\log_e(\delta_n/\delta_0)=bn$.

Finally, dividing both sides by n gives an expression for a local slope, b:

$$b = \frac{1}{n}\log_e(\delta_n/\delta_0). \qquad (25.15)$$

Equation 25.15, by the way, reveals the units of b. The ratio δ_n/δ_0 is dimensionless, since both δ's are in the same units (units of x). The log of that ratio also is dimensionless (log units). Iterations usually represent time, so we let n have units of time. Thus, b has dimensions of 1/time (e.g. 1/seconds, 1/years, etc.). In practice, people often give the numerator a name that depends on the base of the logs. The most popular base of the logs is either e or 2. If taken to base e, local slope b is in **nats** per unit time; if to base 2, it's in bits per unit time (Ch. 6).

We're almost done. This next step is important. First, a word about the philosophy behind it. Local slopes on an attractor, as mentioned, can vary greatly. If we measure only one local slope, we've only examined one locality on the attractor— a locality that might be unrepresentative. In many cases we want to cover the entire attractor. The limitation of the exponential scaling zone (the straight-line relation, as in Figs 25.5 and 25.6) means that we can't use the same two trajectories, from one place to another along the attractor. The usual procedure is to track the same reference trajectory over the entire attractor but choose different "nearby" trajectories at different locations on the attractor. That gives us a bunch of local slopes that represent trajectory divergence (or convergence, or whatever) at different locations on the attractor.

Equation 25.15 describes the separation rate of one pair of neighboring trajectories at some specified location on the attractor. We'll now rewrite the equation in a form whereby it applies to a *group* of local slopes. In other words, it applies to the separation rates of many pairs of neighboring trajectories. We assume that Equation 25.1 and its log-transformed and rearranged version, Equation 25.15, still apply. However, the interpretation is slightly different when a *group* of local slopes is involved. For one thing, δ_n/δ_0 according to Equation 25.14 equals the product of the $f'(x_i)$ values. Hence, we can substitute the product of the $f'(x_i)$ values for δ_n/δ_0 in Equation 25.15. Equation 25.15 uses the log of (δ_n/δ_0), so we need the log of the product of the $f'(x_i)$ values. The log of a product, such as the log of $[f'(x_3)\, f'(x_2)\, f'(x_1)\, f'(x_0)]$ (our example from above with n=4), is the *sum* of the logs of each component (Appendix). In our example, that sum is log of $f'(x_3)$+log of $f'(x_2)$+log of $f'(x_1)$+log of $f'(x_0)$. The standard and general way of writing such a sum in concise symbol form (here we'll take logs to base e) is

$$\sum_{i=0}^{n-1} \log_e f'(x_i).$$

For instance, the summation for our values ranges from i=0 for $f'(x_0)$ to i=3 for $f'(x_3)$, and 3 is n-1 or 4-1. Finally, although gaps or distances can be positive or negative depending on which value is subtracted, we're interested only in the absolute values of the gaps; it's not possible to take the log of a negative number. Hence, we want logs of *absolute values* of derivatives. In symbol form, that's $\log_e|f'(x_i)|$. With those revisions, $\log_e(\delta_n/\delta_0)$ of Equation 25.15 (applicable to just one local slope) becomes

$$\sum_{i=0}^{n-1} \log_e|f'(x_i)|$$

(applicable to a group of local slopes). Making that substitution into Equation 25.15:

$$b = \frac{1}{n} \sum_{i=0}^{n-1} \log_e|f'(x_i)|. \qquad (25.16)$$

Equation 25.16 is just a fancy and shorthand way of telling us to add up the logs of the absolute values of the local slopes and then to get the average log by multiplying that sum by 1/*n*. To show that the answer isn't just one local slope but is instead the average of many local slopes, we'll adopt the symbol λ_1 (Greek lambda, here with subscript 1) in place of b. Finally, to indicate that the slopes should come from all over the attractor (i.e. that the time span should be very long), we qualify the right-hand side of Equation 25.16 with the term . Those changes bring us to our overall goal—a symbol expression for the first, or largest Lyapunov exponent λ_1 (e.g. Wolf 1986: 275):

$$\lambda_1 = \lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} \log_e |f'(x_i)| . \qquad (25.17)$$

Equation 25.17 shows that the first Lyapunov exponent λ_1 is a logarithm—the average log of the absolute values of the *n* derivatives or local slopes, as measured over the entire attractor. In mathematical terminology, a Lyapunov exponent is also the logarithm of a so-called **Lyapunov number**. In other words, a Lyapunov number is the antilog of the Lyapunov exponent, or the number whose logarithm, to a given base, equals the Lyapunov exponent.

Because it averages local divergences and/or convergences from many places over the entire attractor, *a Lyapunov exponent is a global quantity*, not a local quantity. We can interpret it in three ways.

- In *n* dimensions, λ_1 quantifies, in a single number, the average rate at which the fastest growing phase space dimension grows.
- It quantifies average predictability over the attractor.
- Because neighboring trajectories represent changes in initial conditions of (or perturbations to) a system, λ_1, is an average or global measure of how sensitive the system is to slight changes or perturbations. A system isn't sensitive at all in the nonchaotic regime, since any two nearby trajectories converge. In contrast, a system is highly sensitive in the chaotic regime, in that two neighboring trajectories separate, sometimes rapidly.

The value of λ_1 doesn't depend on where we start monitoring, because we're going to cover the entire attractor regardless. Therefore, λ_1 is a basic characteristic for the attractor and is invariant for any attractor. (Invariant here means invariant in regard to starting time and also "unaffected by transformations of the data.")

Lyapunov Spectrum

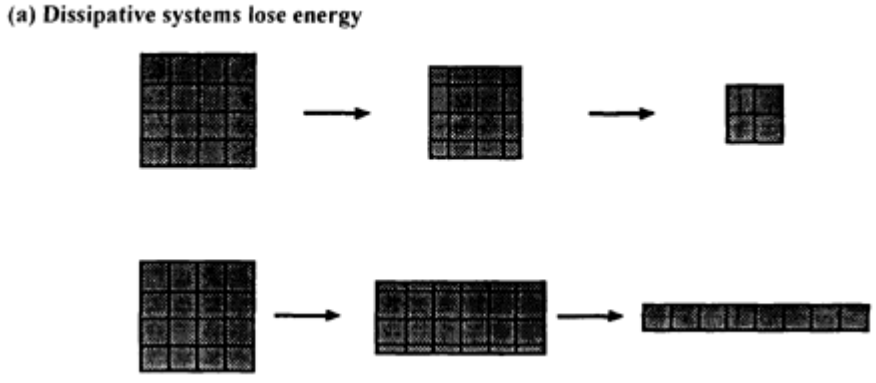
A common approach in visualizing phase space motion is to imagine how a small length, area, volume or higher-dimensional element might evolve in time. Examples of such elements are a square or circle (in two-dimensional phase space) and a cube or sphere (in three dimensions). The center of such an element is an observed datum point. There's an elaborate terminology for describing such an element. Originating at every measured datum point, and making up the associated element's skeletal framework, are one or more equal-length orthogonal axes, called **principal axes**. Each axis ends at another phase space point, although not necessarily an observed (measured) point. Because those axes start at a common center point (a datum point) and each axis ends at some other point in phase space, chaologists call each axis a *vector* (Ch. 5), **axis vector**, or **perturbation vector**. The number of principal axes corresponds to the dimension of the phase space. (However, as we'll see, their directions continually change and have no relation to the directions of the phase space axes.) People usually normalize the orthogonal axes (make them orthonormal) so that the axes have unit (1.0) length.

In any dissipative (energy-losing) system, whether chaotic or not, a length, area, volume or higher-dimensional element shrinks over time, in phase space (Fig. 25.8a). The frictional pendulum and its point attractor of Figure 11.b are a good example. On a phase space plot of velocity versus position, the dynamics describe a spiral. The size of the spiral gets smaller and smaller with time. Eventually the pendulum stops swinging altogether. The spiral then reaches its smallest possible size (a dot), representing a constant velocity and position. Conservative systems, in contrast, don't lose energy with time. In the graphical illustration of Figure 25.8b, the overall total area of the box remains constant over time, even though there may be some change in shape (because of unequal changes in axes lengths).

The phase space orientation of a set of orthonormal axes is arbitrary. In other words, the axes aren't necessarily parallel to the *x*, *y*, and *z* coordinates of the phase space plot. Each axis vector must simply be orthogonal to all other axis vectors.

Now we let the system evolve forward in time, that is, we go to the next measurement. In the process, the imaginary axes or vectors move. Because of the basic nonlinear nature of the system, their lengths change. The amount of change varies from one axis to another. In other words, our idealized element gets distorted. Its orientation in the phase space also may change. Principal axes have numerical labels according to how much change or distortion they undergo. The principal axis that's stretched the most (or reduced the least) is the **first principal axis**. That axis, in other words, corresponds to the direction in which the flow is most unstable. The axis that undergoes the next greatest amount of growth (or least shrinkage) is the second principal axis, and so on.

Now, what does all of this have to do with Lyapunov exponents? The two endpoints of each principal axis are considered to be neighboring points in phase space. We monitor the change in their gap, over time. That is, we measure the growth or shrinkage of each principal axis over the entire attractor, according to whether its endpoints get closer together or farther apart. That means *we get a Lyapunov exponent for each principal axis*. The largest Lyapunov exponent measures the rate of expansion of the first principal axis—the one that shows the largest amount of growth (or the slowest rate of shrinkage) over the attractor. The second Lyapunov exponent measures the rate of change of the second principal axis, and so on down to the smallest Lyapunov exponent.



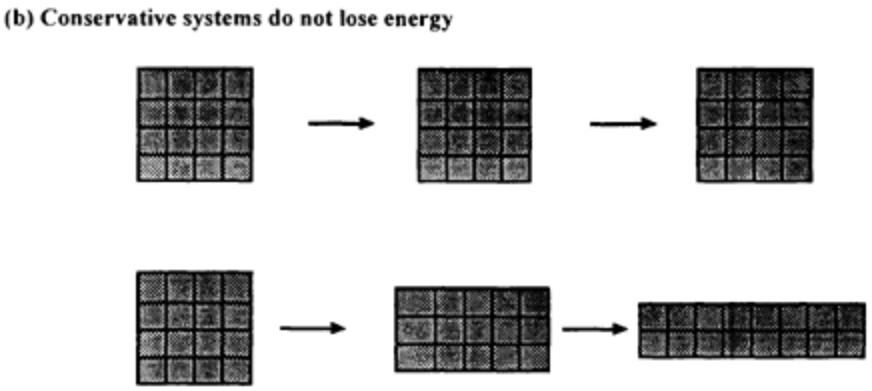


Figure 25.8
Phase space area and some possible changes over time (after Bergé et al. 1984: 23). Size (area) of main box (the one on the left, in each row) represents an amount of energy. In dissipative systems (the two examples in part (a)), area of box decreases over time. In conservative systems (the two examples in part (b)), area of box remains constant over time. With both systems, shape of area-element can change over time, and trajectories can diverge whether area stays constant or not.

For a system to be chaotic, at least one principal axis must show an overall *divergence* of neighboring trajectories (Figs 25.5, 25.6). Therefore, at least one Lyapunov exponent has to be *positive* on a chaotic attractor. The Lyapunov exponents of an attractor, considered as a group and ordered by magnitude, are known as the attractor's *spectrum of Lyapunov exponents*.

A different Lyapunov exponent corresponds to each principal axis, and the number of principal axes equals the number of phase space dimensions. Consequently, the number of Lyapunov exponents equals the number of phase space dimensions. However, all principal axes continually change direction or orientation in phase space, depending on attractor configuration and on flow direction. Hence, a particular Lyapunov exponent doesn't correspond to a particular phase space direction or to a particular axis of the phase space graph.

Chaos can happen in dynamical systems of any number of dimensions. Thus, there can be any number of positive Lyapunov exponents (almost as many as the dimensions of the phase space itself).

Besides positive and negative exponents, a "zero" exponent also is possible. An exponent of zero corresponds to an axis along which the gap between trajectories either stays constant or increases at a rate less than exponential. (Some authors call the zero-exponent case **marginally stable**. "Marginally stable" means that any perturbation stays at about the same level, as the system evolves.) Mathematicians have proven that, except for the fixed-point attractor, at least one Lyapunov exponent is always zero (cf. Parker & Chua 1989: 321). Hence, there are three main types of trajectory behavior in dynamical systems:

- asymptotically unstable (exponential divergence of neighboring trajectories, positive Lyapunov exponent)
- marginally stable (zero exponent)
- asymptotically stable (exponential convergence of neighboring trajectories, negative exponent).

Authors describe a system's dynamics qualitatively in shorthand form by simply listing the signs of the Lyapunov exponents. For instance, (+,0,-) means a three-dimensional phase space in which neighboring trajectories as averaged over the entire attractor diverge exponentially along the first principal axis, are marginally stable along the second, and converge along the third. The Lorenz attractor (Fig. 18.8) is an example.

In three-dimensional phase space, the various attractor types have the following spectrum of Lyapunov exponents:

- A fixed-point attractor is (-,-,-). That label tells us that convergence goes on along all three principal axes of any phase space volume.

- A limit cycle is (0,-,-), because a volume element stays constant along the direction of the trajectory but contracts (the gaps converge) in the directions of the other two principal axes.

- Orbits on a torus, once they reach the attractor, keep the same distance thereafter. A volume element therefore stays constant for two principal axes but contracts along the third (0,0,-).

- Finally, on a chaotic attractor the two trajectories diverge in the direction of the first principal axis, yielding a positive exponent. The exponent for the second principal axis has to be zero because at least one axis must have a zero exponent. The exponent for the third principal axis has to be negative (and larger in absolute value than that of the positive exponent), because the over-all sum of the exponents has to be negative. So, the classification code for a chaotic attractor is (+,0,-).

You'll often see the statement that the length of any phase space element (i.e. the first principal axis of our N_d axes) grows as $c^{\lambda_1 t}$, where c is a constant (usually 2 or 2.718 . . .). Similarly, the cross-sectional area of any element grows as $c^{(\lambda_1+\lambda_2)t}$; its volume grows as $c^{(\lambda_1+\lambda_2+\lambda_3)t}$; and so on. Here's a brief explanation of where those relations come from. They arise from the assumed (or real) exponential growth rate, as follows. As with the earlier derivation, we'll deal with local slope b, with the understanding that the same principle applies globally. With time *t* as the independent variable and length y as the dependent variable, the exponential relation for the length of the principal axis at time one (y_1) is

$$y_1 = y_0 c^{b_1 t} \qquad (25.18)$$

in which y_0 is the length at time zero and b_1 is the slope associated with the first principal axis. The ratio y_1/y_0 then is the relative growth of the axis over one time unit. Equation 25.18 shows that y_1/y_0 grows as

Now for the two-dimensional case. The expression for growth of a cross-sectional area stems from the relation that the area of a disk or circle is π times the radius squared. Let's designate two orthogonal radii, *y* and *z*, as the first and second principal axes. The area of the circle then is πyz. That area at time zero is $\pi y_0 z_0$. By the next time increment (time one), the circle deforms into an ellipse. An expression similar to πyz applies to the cross-sectional area of the ellipse for time one, so the area at time one is $\pi y_1 z_1$. The relative growth of the area during the time interval is area at time one divided by area at time zero=$\pi y_1 z_1/\pi y_0 z_0 = y_1 z_1/y_0 z_0$. Next we insert expressions for y_1 and z_1 into this last ratio. Assuming that the axis y_1 grows exponentially, its value at time one is $y_1 = y_0 c^{b_1 t}$. Similarly, assuming that axis z_1 grows exponentially, its value at time one is $z_1 = z_0 c^{b_2 t}$, where b_2 is the slope associated with the second principal axis. Inserting these expressions into the relative-growth ratio $y_1 z_1/y_0 z_0$ gives:

$$relative\ growth = y_1 z_1/y_0 z_0 = y_0 c^{b_1 t} \cdot z_0 c^{b_2 t}/y_0 z_0 = c^{b_1 t} \cdot c^{b_2 t}$$

or

$$relative\ growth = c^{(b_1+b_2)t}. \qquad (25.19)$$

In other words, cross-sectional area grows as $c^{(b_1+b_2)t}$. In the same way, volume grows as $c^{(b_1+b_2+b_3)t}$, and so on.

Practical Calculation of Lyapunov Exponents

Some of the proposed methods for calculating Lyapunov exponents from measured data are those of Wright (1984), Eckmann & Ruelle (1985) (explained in more detail in Eckmann et al. 1986), Sano & Sawada (1985), Wolfet al. (1985), Briggs (1990), Bryant et al. (1990), Stoop & Parisi (1991), Zeng et al. (1991), Nychka et al. (1992), and probably others. Wright's (1984) method applies only to relatively limited conditions, so I won't elaborate on it here. The basic idea behind all other methods is to follow sets of trajectories over short time-spans and compute their rates of separation, then average those rates over the attractor. These other methods have at least five features in common.

- They begin by reconstructing the attractor, often with the standard time-delay embedding procedure. If the attractor can't be reconstructed, then the game is over already; we can't measure a Lyapunov exponent. (And, as mentioned, there are few reliable criteria to tell us when we've successfully reconstructed the attractor.)

- Working in the reconstructed phase space, all methods focus on a fiducial trajectory over the attractor.

- They directly or implicitly treat each fiducial point as the center of a little area, sphere or hypersphere that has principal axes. As mentioned earlier, evolution of the system disrupts the orthogonality of such principal axes. For instance, a circle in phase space gets distorted into an oval; a sphere evolves into an ellipsoid. (In chaotic systems, the axes in fact tend to align themselves in the direction of most rapid local growth.) The loss of orthogonality is undesirable. For one thing, if the axes collapse toward a common direction they become indistinguishable. In that case we can only measure one trajectory gap. For another, probing the phase space is more thorough if the vectors are orthogonal. Thirdly, mutually perpendicular axes are much easier and more familiar to work with than axes that aren't mutually perpendicular. Because of these considerations, most methods periodically reset the axes to where they are again mutually perpendicular and of unit length (orthonormalized).

There are several ways to reset principal axes to an orthogonal orientation. Different techniques for determining Lyapunov exponents use different resetting methods. Probably the two most common methods are Gram-Schmidt orthogonalization and QR decomposition (Zeng et al. 1992). Adding the normalization step completes the impressive-sounding procedure called **reorthonormalization** (*re*: again; *ortho*: right angle; *normal*: of unit length). The set of principal-axis vectors travels around the attractor, anchored to the reference trajectory, and every so often we reorthonormalize those vectors.

- They're much better at measuring positive exponents than negative ones. At least with chaotic attractors, nearby trajectories diverge over much of the attractor. Most of such chaotic data lend themselves to positive exponents. Of course, one positive exponent is enough to show chaos, so if our only goal is to find out whether or not the system is chaotic, the problem becomes easier. The Wolf et al. technique, in fact, provides only the single largest positive exponent.[5] The other methods, not necessarily successfully, try to find the entire spectrum of exponents.

As mentioned, however, negative exponents also are difficult to measure. They reflect trajectory convergence. Such convergence on chaotic attractors can take place normal to the flow, in a direction where the attractor is relatively thin and information relatively scarce. In addition, measurement of negative exponents depends on their magnitudes and on the relative amount of noise in the data (Sano & Sawada 1985). If trajectories converge to the extent that attractor thickness is less than the resolution of the data (the noise magnitude), then there's no way to get any further information about contraction rates. Measuring a Lyapunov exponent for those conditions produces an incorrect zero exponent, because noise on average neither expands nor contracts. Finally, in other situations, negative exponents for chaotic or nonchaotic attractors measure the rate at which neighboring trajectories that are off the attractor get onto the attractor (Abarbanel et al. 1990), as explained early in this chapter. We may need to determine such rates by analyzing transients, and transients usually last over a very short time or may not have been measured.

- Most methods produce extra (false or spurious) exponents that you have to identify and throw out. (The methods of Wolf et al. [1985] and Stoop & Parisi [1991] may be exceptions.) The extra exponents appear because the methods normally yield the same number of exponents as the number of dimensions of the phase space, and the phase space here is an artificial one whose embedding dimension necessarily is greater than that of the dynamical system it represents. Authors use various approaches, sometimes apparently successfully, to recognize false exponents. Bryant (1992) gives a good discussion with recommendations for how to identify them.

So much for what the methods have in common. Now let's take a closer look at three of them—those of Wolf et al., Eckmann & Ruelle, and Sano & Sawada.

The popular Wolf et al. (1985) method begins by taking the first point of the reconstructed (lagged) phase space to represent the fiducial trajectory (Fig. 25.9). The authors choose a nearby point as the first observation along the neighboring trajectory. That nearby point must be from a different temporal section of the dataset. They monitor the gap between the two trajectories over time. The length of time over which they do so can be until the occasion of the very next datum value, or until the gap exceeds an arbitrary distance e, or until another trajectory becomes closer. When such a predesignated condition occurs, they consider that particular neighboring trajectory to have migrated a maximum allowable distance from the fiducial trajectory. At that stage, they compute a local slope as [final separation distance minus starting separation distance] divided by the associated time interval. Then they abandon that first neighboring trajectory.

The last point just analyzed on the fiducial trajectory stays on for additional duty as a new reference point. A new nearby point is chosen to represent a new nearby trajectory (Fig. 25.9). Then they measure the rate at which the new nearby trajectory diverges from the fiducial trajectory, as before. They repeat the process until the fiducial trajectory gets to the end of the data. At this stage, we assume that the data cover the attractor. Averaging the logs of the absolute values of the various local divergence rates gives the largest Lyapunov exponent (Eq. 25.17).

---

[5]. Those authors do, however, also propose a more elaborate and cumbersome technique for the two largest exponents.

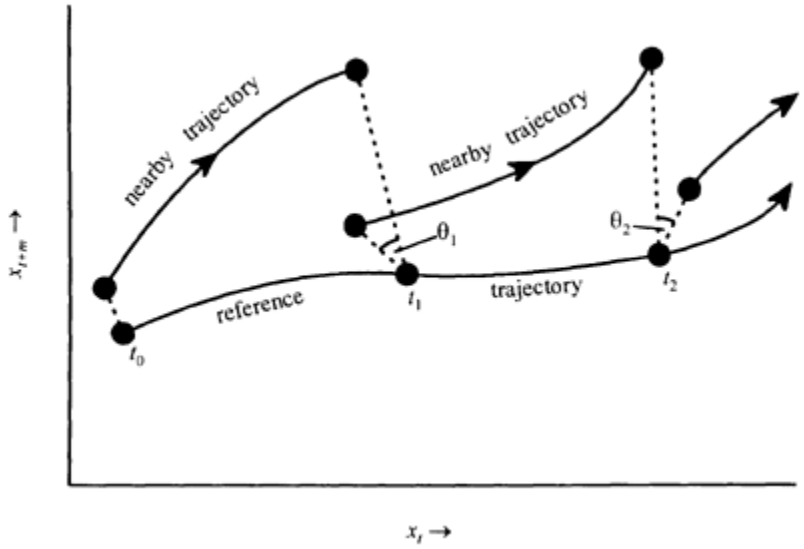


Figure 25.9
Wolf et al. (1985) approach to estimating largest Lyapunov exponent from measured data. At each replacement time, angle θ (between old length element and new element) as well as new length element are minimized.

When a nearby trajectory becomes obsolete and is replaced with a new one in the Wolfet al. method, two criteria determine the starting point for the new nearby trajectory. First, the starting point should be as close as possible (in the reconstructed phase space) to the new fiducial point. Secondly, the spatial orientation of the last (now abandoned) separation vector should be preserved as closely as possible. That means making the angle θ on Figure 25.9 as small as possible. In practice, fulfilling those two criteria can be difficult, and the computer often must look at more than one nearby datum point to find the best compromise replacement point. Fortunately, such orientation problems usually don't have a major effect on the Lyapunov exponent. Also, divergence rates of replacement trajectories must be monitored closely. If they grow suspiciously quickly, the fiducial and nearby points probably aren't on the same part of the attractor and are going in different directions because of a fold in the attractor. In that case (very rapid divergence), go back and choose a different second point. Abarbanel et al. (1990) give further discussion.

The amount of data needed for calculating Lyapunov exponents with the Wolf et al. system depends on three factors (Wolfet al. 1985: 305). First is the number of replacement points that will be needed; second is the number of orbits around the attractor needed to assess the stretching adequately; and third is how well the attractor can be reconstructed with delay coordinates. In practice, the amount of data increases approximately exponentially with the dimension *D* of the attractor; the authors give 10^D-30^D as a general range of that desirable amount. About 10000-100000 points are typical requirements. Eckmann & Ruelle's (1992) rule of thumb yields one of the largest estimates: the required number of points is about the *square* of that needed to reliably estimate the correlation dimension. (Ch. 24 mentions their rule for estimating the required size of the dataset for determining the correlation dimension.) For example, if by their rule we need 1000 points to get the correlation dimension, we need 1000000 points to estimate a Lyapunov exponent!

Eckmann & Ruelle (1985) (also Eckmann et al. 1986) and Sano & Sawada (1985), working independently, proposed methods that are philosophically very much alike. Hence, in most of the following general description I'll treat them as one method. This technique follows *entire neighborhoods* (groups of points) near the fiducial trajectory, over a short time. In a stripped-down, simplified version, we begin by taking the first datum point as the first fiducial point. Then we locate all other points (x_{j1}, x_{j2}, etc.) within an arbitrarily prescribed radius ε (Fig. 25.10).

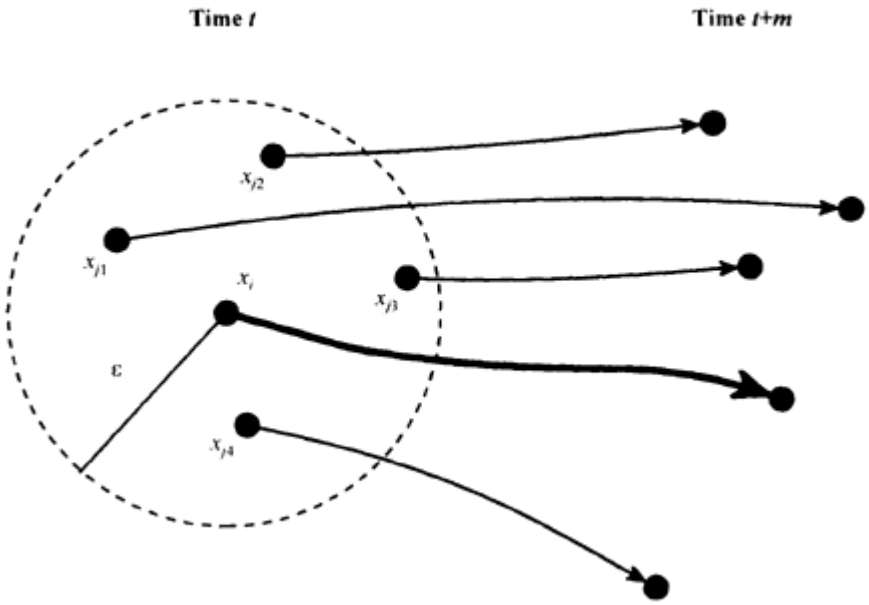


Figure 25.10
Sketch of Eckmann et al. concept for estimating a Lyapunov exponent. A circle of radius ε is centered at fiducial point x_i; four qualifying neighboring points are shown.

Those qualifying points are called "neighbors." Each neighbor has its own separation distance from the fiducial point. Moving along the fiducial trajectory to the second datum point, we next find how many of the original neighbors still qualify as neighbors. Those surviving neighbors have a new gap relative to the fiducial trajectory. We assume that the time between steps one and two is short enough that the change in gaps is linear rather than nonlinear. That is, we assume a linear relation between the gaps at time one and those at time two. In other words, the group relation is that gaps at time 1 equal a constant times gaps at time zero. Fitting a straight line to those gaps (Eckmann & Ruelle used least squares) gives the value of that constant. The same general procedure is repeated at each successive point on the fiducial trajectory. From the many values of the constant determined in the regressions, we compute the Lyapunov exponents from certain mathematical equations.

In practice, we might follow the neighbors over several data values before doing the linear fit. Also, the radius ε doesn't necessarily have to be constant from one fiducial point to the next. The selected value of ε is a compromise between the two conflicting requirements that ε be small enough that the flow can be considered linear (a limitation resulting from nonlinearity) and large enough to include enough neighbors to satisfy certain statistical requirements (a limitation caused by noise). The optimum choice often appears only by trial and error and linear fitting. Another problem is that finding neighbors can be difficult. The number of neighbors within a given radius depends on the size of the dataset, noise, and the fractal nature of the attractor.

Probably the main difference between the Eckmann & Ruelle (1985) approach and that of Sano & Sawada (1985) is the way of doing the orthonormalization. Abarbanel et al. (1991b) review other differences. Other methods mentioned above are largely variations or improvements of the Eckmann-Ruelle/Sano-Sawada approach.

Calculating Lyapunov exponents for an attractor reconstructed from a time series of discretely measured, noisy real-world data is a problem that has only partly been resolved. For instance, Theiler et al. (1992) found that it is possible to get positive Lyapunov exponents for data known to be nonchaotic. (In fact, even data that are essentially random conceivably can have a positive Lyapunov exponent.) Along the same line, Vassilicos et al. (1993) analyzed dollar to Deutsche mark exchange rates. They found a "marginally positive" largest Lyapunov exponent (and hence possible chaos) for a subsample of 15000 data points but a marginally negative exponent using the full sample of 20000 data points. In general, all the various calculation methods also have problems with quality of data (noise), how best to fit the function that describes local dynamics, and the size of the local neighborhoods (Glass & Kaplan 1993). In addition, the methods can't always reliably recognize false exponents, as mentioned. Finally, they have difficulty in high-dimensional systems, possibly because trajectories on those attractors rarely return to a given locality (a problem mentioned earlier in regard to dimension estimates). But then again, most chaos tools don't work well in high dimensions.

Summary

A Lyapunov exponent is a number that reflects the rate of divergence or convergence, averaged over the entire attractor, of two neighboring phase space trajectories. Trajectory divergence or convergence has to follow an exponential law, for the exponent to be definable. In other words, the separations must plot as a straight line with trajectory gap on the ordinate (log scale) and time on the abscissa (arithmetic scale). The general procedure is to determine the logarithms of local slopes of such straight-line relations for all regions of the attractor and then average those logs to derive the Lyapunov exponent. A negative Lyapunov exponent indicates an average convergence of trajectories, a positive exponent an average divergence. Convergence and hence negative Lyapunov exponents typify nonchaotic attractors. Divergence (and hence at least one positive Lyapunov exponent) usually happens only on chaotic attractors. However, convergence can also take place on chaotic attractors. Hence, local rates can vary considerably at different regions on the attractor.

A "local Lyapunov exponent" evaluates trajectory separation over short finite time periods (i.e. over local regions of the phase space attractor). Local Lyapunov exponents can be useful in the practical world in prediction over short timescales and in assessing how accurate a forecast might be.

The number of Lyapunov exponents for a set of data is numerically the same as the number of phase space axes. However, each exponent corresponds to a principal axis of a length, area, volume or higher-dimensional element, and the orientation of such principal axes varies with location on the attractor. Hence, Lyapunov exponents don't correspond to the phase space axes nor to any particular phase space orientation. The exponents of an attractor collectively are known as the spectrum of Lyapunov exponents and are represented in a shorthand form as a string of plusses, minuses, and zeros. There isn't yet a completely reliable way to compute Lyapunov exponents for attractors constructed from real-world data; the two most common techniques are the Wolf et al. (1985) method or a refinement of the Eckmann & Ruelle (1985)/Sano & Sawada (1985) method.

Chapter 26
Kolmogorov-Sinai entropy

We come now to a very elegant quantitative measure of chaos. To date, it's been useful mostly with computer-generated data, because it requires unusually large amounts of relatively noise-free values. It also needs enormous computer capabilities. It's popular in the chaos literature and involves many important principles of chaos theory.

The notion of **entropy** crops up in physics, mathematics, statistics, economics, computer science, literature, earth sciences, biology, and other fields. To many people, however, it's always been an abstract and confusing idea. In fact, mathematician John von Neumann supposedly recommended using the term because "no one knows what entropy really is, so in a debate you'll always have the advantage" (Tribus & McIrvine 1971: 180). Part of the confusion is because people apply "entropy," like "dimension," to many different concepts or measures.[1] Be that as it may, one particular variant of the group is an important indicator of chaos.

Selected Milestones in the History of Entropy

Clausius's thermodynamic entropy

Based in part on French physicist Sadi Camot's groundwork of the 1820s, the German physicist Rudolf Clausius[2] introduced entropy in 1865 in his work on heat-producing engines. The general idea is that it's impossible to direct all of a system's energy into useful work, because some of that energy isn't available (escapes, for example). Entropy in this original thermodynamic (heat movement) sense is a measure or calculation of the inaccessible energy. High entropy means that much of a system's energy can't be used for work. In other words, only a small part is available for work. Low entropy means that only a small proportion of the system's energy is unobtainable, that is, most of it's available.

---

[1]. Within science, literature and other fields, this plethora of uses of "entropy" causes much unhappiness and hand-wringing. Denbigh & Denbigh (1985), Schneider (1988), and Hayles (1990) give more extensive discussions.

Boltzmann's statistical entropy

Austrian physicist Ludwig Boltzmann[3] contributed a key development in the "career" of entropy when he devised a statistical or probability measure of entropy *H:*

$$H = -\mathrm{K} \sum_{i=1}^{N_s} P_i \log P_i \qquad (26.1)$$

where K is Boltzmann's constant (and depends only on the units) and P_i here is the ordinary probability of an element being in any one of the N_s phase space states. Equation 26.1 is very similar to our Equation 6.17c, reproduced below. (In fact, the rest of this chapter relies heavily on the probability material we discussed in Ch. 6.)

Shannon's information entropy

At least as early as the 1870s, some researchers (possibly including even Boltzmann himself) began looking upon entropy as a measure of information. In the mid-twentieth century, Nyquist, Hartley, Shannon and others in electrical communications, and Wiener in cybernetics, made major progress toward a numerical measure of information. Claude Shannon,[4] an engineer with the Bell Telephone Laboratories, built on earlier work on mathematical approaches to information. In the late 1940s (summarized in Shannon & Weaver 1949) he largely created the then-new discipline of information theory. In particular, by considering probabilities, Shannon arrived at Equation 6.17c:

[2]. **Clausius, Rudolf (1822-88)** Rudolf Clausius, one of the greatest of theoretical physicists, made outstanding contributions to thermodynamics and kinetic theory during the 1850s and 1860s. After graduating from the University of Halle in 1848 he became a professor at the Imperial School of Artillery and Engineering in Berlin. Later he was professor at Zürich, Würzburg, and finally Bonn. His scientific accomplishments include formulating or restating fundamental laws of thermodynamics, putting the theory of heat on a sound basis, and developing the kinetic theory of gases. Much of this work began when he was in his late twenties and climaxed in his early forties. Although regarded as the founder of thermodynamics, he gave much credit to such predecessors as Sadi Carnot, Henri Clapeyron, Robert Mayer, and others. Max Planck, another great German physicist, perhaps best described Clausius's impact: "One day, I happened to come across the treatises of Rudolf Clausius, whose lucid style and enlightening clarity of reasoning made an enormous impression on me, and I became deeply absorbed in his articles, with an ever increasing enthusiasm. I appreciated especially his exact formulation of the two laws of thermodynamics, and the sharp distinction which he was the first to establish between them . . ."

[3]. **Boltzmann, Ludwig E. (1844-1906)** Austrian physicist Ludwig Boltzmann received his doctorate at the University of Vienna in 1866 and later held professorships in mathematics and physics at Vienna, Grad, Munich, and Leipzig. In 1894 he became Professor of Theoretical Physics at Vienna. where he spent most of the rest of his life. He was one of the principal founders of statistical mechanics. In the 1870s he published several papers in which he showed that the second law of thermodynamics was explainable by applying probability theory and laws of mechanics to the motions of atoms. His work was vigorously attacked and long-misunderstood, but he was eventually proved right. He himself was physically strong, sensitive, and had a fine sense of humor. However, he tended to have periods of silence and depression. During one of his depressed periods, he took his own life at age 62.

$$I_w = -\sum_{i=1}^{N_s} P_i \log P_i . \qquad \text{(6.17c, repeated)}$$

Any total (N, N_r, etc.) can replace N_s in that equation. (Norbert Wiener developed essentially the same expression for his own field, at about the same time.) The formula is, of course, nearly the same as Boltzmann's long-established equation for entropy (Eq. 26.1) as used in statistical mechanics (with Boltzmann's K=1). The greatest strength of Shannon's formula is its broad applicability; whereas Boltzmann's formula was limited to statistical mechanics, Shannon's applies to any field and to many types of probability.

In producing his equation, Shannon simultaneously provided both a gigantic advance and massive consternation. The consternation arose because of the name he gave the equation. On the one hand, he said it actually measures information, choice, or uncertainty. However, to the dismay and confusion of many, he opted to call it "entropy" because his equation was essentially identical to Boltzmann's. Physicists today are still arguing over whether Shannon's entropy (dealing with the generation of averages of various statistical options) really measures the same thing as Boltzmann's entropy (dealing with the probabilities of accessible molecular states). In any event, many people began to treat entropy and information (in the Shannon sense) as the same thing. Chaos theory treats them as the same thing.

Important examples of equations that quantitatively measure both information and entropy are $I=\log_2(1/P)$ (our Eq. 6.16a) and

[4]. **Shannon, Claude E. (1916-)** Claude Shannon is generally acknowledged as the father of information theory. After receiving his PhD in the mathematics of genetic transmission from Massachusetts Institute of Technology (MIT) in 1940, he worked at Bell Laboratories from 1941-56. During the Second World War he was part of a team that developed digital codes, including a code that Roosevelt and Churchill used for transatlantic messages. That work led to his theory of communication and to his introduction of entropy into information theory. His now-classic 1948 paper on that subject wasn't widely understood or accepted at the time and even received a negative review from a prominent mathematician. Today, however, "Shannon's insights help shape virtually all systems that store, process, or transmit information in digital form—from compact discs to supercomputers, from facsimile machines to deep-space probes such as Voyager" (Horgan 1992). In 1956 he returned to MIT as professor of communications science. His career includes many national and international honors and awards but is also noteworthy for his many side interests. Juggling is but one example; he reportedly rode through the halls of Bell Laboratories on a unicycle while juggling four balls. (He claims that his hands are too small to juggle more than four balls at once!) Shannon's home near Boston (called, naturally, "Entropy House") contains entire rooms devoted to his many inventions and hobbies. Those include a rocket-powered frisbee, a mechanical mouse that finds its way out of a maze, a juggling robot of the comedian W. C. Fields, and a computer called Throbac (Thrifty Roman Numeral Backward Computer) that calculates in roman numerals.

$$I_w = \sum_{i=1}^{N_s} P_i \log (1/P_i).$$ (6.17b, repeated)

Since we're now talking about entropy, we'll substitute the symbol H for entropy, in place of I, and H_w (for I_w) for the weighted sum. Equation 6.16a (the equal-probability case) therefore becomes:

$$H = \log_2(1/P) \qquad (26.2)$$

and Equation 6.17b (the unequal-probability case) becomes:

$$H_w = \sum_{i=1}^{N_s} P_i \log (1/P_i). \qquad (26.3)$$

As explained in Chapter 6, the weighted sum H_w is an average value. Some authors call the entropy of Equations 26.2 or 26.3 the **information entropy**.

All relations developed in Chapter 6 between information and probability also apply to entropy and probability. Whether the context is one of information or entropy, *probabilities* are the foundation or essence of it all; they are the required basic data. As Equations 26.2 and 26.3 indicate, the entropy or information varies depending on the distribution of those probabilities. In other words, entropy doesn't depend at all on the actual values of the variable; rather, it's just a statistic that characterizes an ensemble of probabilities. For that reason, the entropy H_w properly is called the "entropy of the probability distribution P_i." An important feature of Equation 26.3 and entropy H_w is that the probabilities P_i can be any type of probability (conditional, etc.), not just the ordinary ones. We'll come back to that feature later.

Two variants or extensions of information entropy that are important in chaos theory are **Kolmogorov-Sinai entropy**[5] and mutual information. Both variants (discussed later) use, or are based on, Equations 26.2 and 26.3.

Interpretations of Entropy

Information is but one of many interpretations of entropy (Table 26.1). Chronologically, it was a relatively late interpretation. There's some debate and confusion in the literature as to whether all interpretations are correct. Opinions also differ about the sign that the value computed from Equations 26.2 or 26.3 should have, for a particular interpretation.

The first entry in Table 26.1 is Clausius's thermodynamic interpretation. Other scientists developed the second interpretation—disorder—soon after Clausius's initial work. The rationale was as follows. Energy is dissipated and entropy therefore increases not only with heat loss but also with other actions. Examples are the change from solid to liquid to gas, chemical reactions, and mixing, as when hot and cold liquids mix or when a gas expands into a vacuum. Those processes involve not only an increase in entropy but also an accompanying decrease in the orderly arrangement of constituent atoms (an increase in disorder). For example, atoms and molecules supposedly are ordered (have low entropy) in a solid, less ordered in a liquid, and quite disordered (have high entropy) in a gas. Consequently, scientists began to look upon a measurement of entropy as a measurement of the system's degree of disorder or disorganization.

Table 26.1 Extreme or opposite states of entropy (modified and expanded from Gatlin 1972: 29).

High entropy	Low entropy
1. Large proportion of energy unavailable for doing work	Small proportion of energy unavailable for doing work
2. Disorder, disorganization, thorough mix	Order, high degree of organization, meticulous sorting or separation
3. Equally probable events, low probability of a selected event	Preordained outcomes, high probability of a selected event
4. Uniform distribution	Highly uneven distribution
5. Great uncertainty	Near certainty, high reliability
6. Randomness or unpredictability	Nonrandomness, accurate forecasts
7. Freedom (wide variety) of choice, many possible outcomes	Narrowly constrained choice, few possible outcomes
8. Large diversity	Small diversity
9. Great surprise	Little or no surprise
10. Much information	Little information
11. Large amount of information used to specify state of system	Small amount of information used to specify state of system
12. High accuracy of data	Low accuracy of data

[5]. **Kolmogorov, Andrei Nikolaevich (1903-87)** It's difficult to name anyone who has matched Russian mathematician Andrei N. Kolmogorov's scientific, pedagogical, and organizational contributions. His scientific career alone was outstanding. In his approximately 500 publications, he made important advances in probability theory, Fourier analysis and time series, automata theory, intuitionistic arithmetic, topology, functional analysis, Markov processes, turbulence, theory of fire. statistics of mass production, information theory, random processes, and dynamical systems. A significant part of his scientific work dealt with applying mathematics to natural sciences, life sciences. and Earth sciences. In addition, he wrote papers on philosophy, history, methodology of science, mathematical education, and other topics. After being awarded his PhD by Moscow University in 1929, he was (besides his teaching and research duties) head of various departments and institutes at Moscow State University from 1933 until his death in 1987. Over the years he established several institutes and laboratories in Moscow. On top of all that, he founded and/or edited several technical mathematical journals. He received many national and international awards and honors. Equally important were his contributions in education. He was keenly interested in teaching, especially young people, throughout his life. In fact, school mathematics education was his main (but not only) concern during the last third of his career. During his life he organized national mathematical competitions for students, gave popular lectures to schoolchildren. led national committees that analyzed teaching methods, helped launch a boarding school (elementary school) for gifted physics and mathematics students, helped initiate and publish about 200 special (mathematical) issues of a student-oriented magazine, helped train teachers on new course content (by lecturing and writing articles), served on the editorial boards of pedagogical publications, chaired many committees on education improvement, and so on. Finally, he had a particularly broad range of other interests in life—art (especially Old Russian painting and architecture), nature (hiking and mountain skiing), travel (including trips to at least 20 foreign countries), human company, Russian poetry, and gardening.

The third common interpretation in Table 26.1 is in terms of probability (Eq. 26.1). Boltzmann's probabilities dealt with particle motion, deviations from equilibrium, and similar notions; we'll take a simpler analogy. Say we dump the pieces of a picture puzzle on the table and completely stir them up so that they are in great disorder (high entropy). Selecting a piece at random, there's an equal chance of picking any particular piece. Similarly, rolling a six-sided die has an equal chance of producing any of the six numbers; the spin of a roulette wheel can produce one particular number as well as another. Just as entropy is a maximum for disordered conditions, it's also a maximum for equally probable events. That is, when all possible outcomes are equally likely, the probability of any one outcome is low, and entropy is high. In contrast, when some scoundrel rigs the roulette wheel or loads the die so that a particular number will come up, the outcome is far from equi-probable, and entropy is lowest. So, there's an inverse relation between entropy and probability, as Equation 26.2 shows.

A stripped-down example of two possible states shows how entropy varies with probability distribution. Equation 26.3 for two states is $H_w=P_1\log_2(1/P_1)+P_2\log_2(1/P_2)$. Choosing P_1 automatically determines P_2, since the probabilities always sum to 1. For instance, if $P_1=0.1$, then $P_2 = 0.9$, and $H_w=0.1\ \log_2(1/0.1)+0.9\ \log_2(1/0.9)=0.332+0.137=0.469$ bits. Choosing any P_1 (and hence P_2) and computing the entropy, for different combinations of P_1 and P_2, produces the curve shown in Figure 26.1. Probabilities are most *unequal* when one of them is zero and the other is 1.0. Computed entropy then is zero (the lowest possible value). As the two probabilities become closer in value (i.e. as P_1 increases from 0 to 0.5), entropy increases. When the two probabilities become equal ($P_1=P_2=0.5$), entropy reaches its maximum value. (And so, as mentioned above, any value for entropy applies only to a particular distribution of the probabilities P_i.)

The fourth interpretation, based on the probability notion, is that of uniformity of distribution of data. If the data are uniformly distributed among a certain number of compartments, the probability of getting each compartment in a sample or trial is equal. In that sense, a uniform distribution is a high-entropy condition (Fig. 26.1 at peak of curve, where $P_1=P_2=0.5$). Conversely, a very nonuniform distribution (e.g. all observations in one compartment) means low entropy, because one bin has a probability of 1 and the other bins a probability of zero (Fig. 26.1 at either end of the parabola).

A fifth notion of entropy is uncertainty. The uncertainty can pertain to the outcome of an experiment about to be run, or it can pertain to the state of a dynamical system. For instance, when rolling a die, we really have no idea which number is going to result. In other words, our uncertainty is greatest. Therefore, we might as well assume complete disarray, equal probabilities, and a uniform distribution.

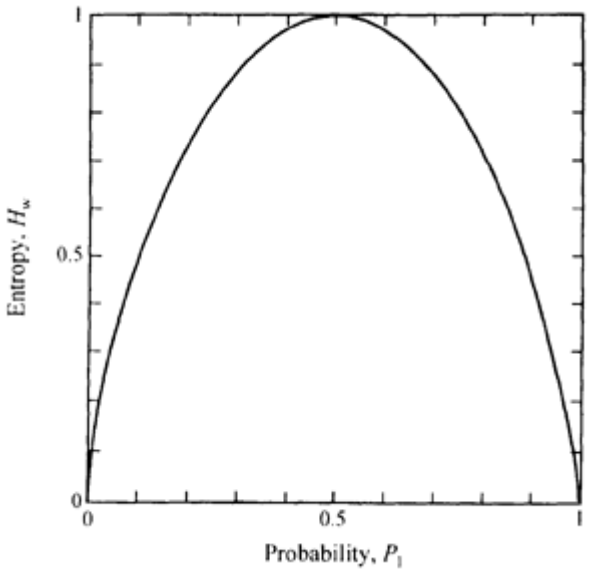


Figure 26.1
Variation of computed entropy H_w with probability distribution for the two-bin (binary) case. Probability for bin one (P_1) is plotted on the abscissa. Specifying P_1 also specifies P_2.

Thus, just as complete disarray and equal probabilities represent greatest entropy, so too does maximum uncertainty. The other extreme is when we're absolutely sure what an outcome will be. Then there's zero uncertainty and zero entropy. (For example, absolute certainty means probability P=1, for which case the entropy in Equations 26.2 and 26.3 computes to zero.) Hence, entropy reflects degree of uncertainty—the lower the entropy, the less the uncertainty, and vice versa.

In concert with the above five variations is a sixth idea: that of randomly distributed observations versus reliable predictability (and so in this sense "random" means unpredictable). Where there's disorder (thorough mixing) and great uncertainty (high entropy), predictions can't be based on any known structure or pattern and can only be done probabilistically. In contrast, something well organized or nearly certain (low entropy) is usually very predictable.

Freedom of choice, or large number of possible outcomes or states, is a seventh way of looking at entropy. When constrained to one choice or outcome, there's no disorder, uncertainty, or unpredictability whatsoever. Probability is highest (1.0) and entropy is lowest (zero). As more choices become available, entropy increases. Maximum entropy comes about when the outcome can happen in the largest number of ways. The easiest way to see this is to use the identity derived in Chapter 6, namely the equal-probability case where $1/P$=number of possible outcomes, N_s. Inserting N_s for $1/P$ in Equation 26.2 above shows that entropy $H=\log N_s$. So, entropy increases with N_s.

The idea of many possible outcomes suggests diversity, the eighth variation in Table 26.1. Conceptually, the analogy seems valid. However, please be careful when adapting it to practical problems. Some ecologists, for example, use Equation 26.1 to measure the diversity of species in a sample. Unfortunately, as Table 26.1 shows, high entropy can reflect not only large diversity but also uniform distribution. A single value of entropy doesn't distinguish between the two. For instance, a high entropy can indicate both a large number of species and a uniform distribution of those species within the sample. Looking at it another way, a sample that has only a few species but uniform distribution can yield the same value of entropy as a sample with many species that are very unevenly distributed (Pielou 1969: 222). Entropy therefore isn't a suitable measure of diversity because it can be ambiguous.

A notion that follows probability, uncertainty, and variety of choice is that of surprise, a ninth interpretation. In this sense, entropy (the number obtained by calculating Eqs 26.2 or 26.3) reflects the amount of surprise you might feel upon learning the outcome of an event or measurement. When the likelihood of a particular result is high and uncertainty therefore is lowest, you aren't surprised when the result comes out as you expected. Entropy for such a near-certain little-surprise situation is low (as with the no-snow-in-Miami-on-4-July analogy of Ch. 6). Conversely, where any event is as likely as another out of many possible events, uncertainty before the event is highest (the high-entropy condition). You are then more surprised at the outcome, whatever it is.

I've already mentioned the idea of interpreting entropy as information (entry 10 in Table 26.1). In this sense, the number given by Equations 26.2 or 26.3 is an information value that characterizes a particular group of probabilities. We say that the given ensemble or probability distribution represents an information value of so many bits. Furthermore, a relatively large number means a relatively large amount of information, and vice versa. For example, Chapter 6 showed that the two-choice equal-probability case (e.g. whether the baby will be a boy or girl) carries an information of one bit; that's not very much information, because we knew beforehand that there were only two possibilities. In contrast, Equation 26.2 tells us that rolling a six-sided die is characterized by the relatively larger amount of 2.59 bits of information (Ch. 6). To the extent that a new measurement changes the probability in Equation 26.3, entropy is the average amount of new information we've gained by making the measurement.

The 11th concept in Table 26.1 interprets entropy as the amount of information needed to specify the state of the system to a particular accuracy. The grids and computed entropies of Table 26.2[6] help explain this notion. The table pertains to cases where all possibilities are equally likely. Each grid represents a subdivision of phase space, using boxes of size e. Suppose a system is in a particular state (a particular box within a phase space grid), and we want to guess that state or box. The procedure we'll follow in zeroing in on the target box is to ask yes-or-no questions (Schuster 1988: 236ff.). In particular, the game will be to locate the target box by asking the fewest possible questions about where it is. The most economical approach is to ask questions that subdivide the eligible boxes into two equal groups. In other words, we methodically narrow down the possible locations of the target box by halving the remaining group of eligible boxes. A typical yes-or-no question might be, is the box in the lower half of the grid?

---

[6]. Table 26.2 is merely Table 6.1, reproduced here for convenience.

Let's see what happens to the value of information entropy as the required number of yes-or-no questions increases. For the top row of Table 26.2, we don't need to ask any questions at all because there's only one box to locate. In other words, to specify the state of the system to that very crude accuracy, the required number of questions is zero. Here's the important thing: the computed entropy (col. 4 of the table) is also zero. Now let's move to the next step in complexity: two boxes (second row of table). Suppose the variables that define the state of our system plot in the left-most of those two boxes. Just one question is enough to locate that box (i.e. to specify that condition). It makes no difference whether the answer is yes or no. For instance, we can ask "Is it the right-hand box?" The answer is no, so we'd know we want the left-hand box. One question suffices. Also, the computed value of H (assuming equal probabilities for the two boxes) is $\log_2(1/P)=\log_2(1/0.5)=\log_2 2=1$ bit. (Or, alternatively, $H=\log_2 N_s=\log_2 2=1$ bit.) In other words, again the required number of questions equals the computed entropy (here 1). The next level of complexity is the middle row of the table. The grid now has four boxes. Locating any particular target box takes two questions:

1. Is our mystery box in the upper half of the grid? (The answer tells us whether our target is one of the upper two boxes or one of the lower two.)

2. Is it the left-hand box?

1 Grid	2 Number of compartments or states N_S	3 Probability P $=1/N_S$	4 Entropy $H = \log_2(1/P)$ $= \log_2 N_S$ (in bits)
	1	1	0
	2	0.5	1
	4	0.25	2
	8	0.125	3
	16	0.062	4

Table 26.2
Probabilities and information for
equally
likely outcomes (after Rasband 1990:
196).*
* This table contains the same
information as Table 6.1.

Also, the computed entropy *H* is 2 bits. Again, therefore, the number of yes-or-no questions needed to locate the box is the same as the entropy. The same principle holds for the other grids in the table. In other words, the computed entropy *H* (in bits) tells how many yes-or-no questions it takes to guess the state of the system.

Guessing the state of the system is tantamount to specifying it to a particular accuracy. Entropy increases proportionally. To specify the system to the lowest possible accuracy, namely to within just the entire grid, requires zero bits; to the still somewhat crude accuracy of half the grid, 1 bit; to the greater accuracy of one-quarter of the grid, 2 bits; and so on. The higher the entropy, the greater the accuracy to which we can determine the state of the system.

The converse of this last statement leads to the final (12th) interpretation of entropy: entropy indicates the accuracy of the data. For instance, in Table 26.2 each grid pattern might represent the degree of accuracy to which our tools let us measure the system. A certain computed entropy (col. 4) pertains to each of those grids or accuracies. If we can only measure to within the entire largest box (row 1—the low-accuracy case), the associated entropy is low (zero). On the other hand, if we can measure to within a relatively small box (row 4—a relatively high accuracy), entropy is considerably higher. So, high entropy corresponds to high accuracy in the measurements and vice versa.

In practice the probabilities usually aren't equal and hence are weighted in arriving at the total entropy. However, the various interpretations of entropy as just described are still valid.

Kolmogorov-Sinai entropy defined

Shannon's entropy (information entropy), regardless of what types of probabilities we use in it, can't by itself identify chaos. Its value is always positive and finite (except when P=1 and entropy is zero) and can vary by a lot, depending on the control parameter, number (size) of compartments, and other factors. However, the equations for information entropy (Eqs 26.2, 26.3) form the basis for another type of entropy—Kolmogorov-Sinai entropy—that at least theoretically can identify chaos.[7] For brevity, I'll usually just refer to it as K-S entropy. It has several aliases in the literature, including "source entropy," "entropy of the source," **measure-theoretic entropy**, "metric-invariant entropy," and **metric entropy**.

K-S entropy has the following three important features:

- First, K-S entropy requires *sequence probabilities* (the probabilities that the system will follow various routes over time; Ch. 6). The entropy equations (Eqs 26.2, 26.3) are very general and don't specify any particular type of probability. K-S entropy deals with the information or uncertainty associated with a *time sequence* of measurements or observations.

---

[7]. A. N. Kolmogorov (1959) proposed applying Shannon's entropy to dynamical systems. Later in 1959 Y. Sinai gave a refined definition and a proof.

A brief example shows how to use sequence probabilities in computing information entropy. We'll compute information entropy for longer and longer overall periods—first for a period of one time interval (e.g. week), then for a period that lasts over two consecutive intervals, and so on. Such computations are easy and straightforward. Equation 26.3, namely

$$H_w = \sum_{i=1}^{N_r} P_i \log (1/P_i),$$

is the equation. At each successive time we just plug in the first of the various probabilities (an example is given below), compute the associated product $P\log(1/P)$, repeat for each other probability in turn, and add up the products, as Equation 26.3 says. The probabilities to use (sequence probabilities) are the likelihood that the system will follow each of the various possible routes that finish at the chosen time. The summation then is over all routes represented in the data, or N_r. Changing the symbols slightly helps keep those distinctions in mind. $H_{\Delta t}$, replaces H_w, to show that we're computing entropy over a particular window of time (duration). (Here that duration usually will be in terms of number of observations rather than actual time units.) Also, P_s replaces P_i to help us remember that the probabilities are sequence probabilities. Our entropy relation (Eq. 26.3) in new clothing therefore becomes

$$H_{\Delta t} = \sum_{i=1}^{N_r} P_S \log (1/P_S) \qquad (26.4)$$

in which N_r is the number of possible phase space routes.

Our example will have just two possible states, A and B. Figure 26.2 (mainly a reproduction of Fig. 6.7) shows that at "time 1" there are only two possible "routes," namely into bin A or into bin B. (And so entropy computations for time 1 are an exception to using sequence probabilities, because the only probabilities available are ordinary ones.) In Figure 26.2, the routes into bins A and B have ordinary probabilities of 0.60 and 0.40, respectively. For each route, Equation 26.4 says to compute $P\log(1/P)$, then add the two results. For bin A, $P\log_2(1/P)=0.60\log_2(1/0.60)=0.442$. For B, $P\log_2(1/P)=0.40\log_2(1/0.40)=0.529$. (For brevity, I'll drop the units of "bits" that actually go with each product.) According to Equation 26.4, entropy is the sum for all routes: $H_{\Delta t}=0.442+0.529=0.971$.

Entropies for two or more successive time intervals are computed in the same way but with sequence probabilities. Each time gets its own entropy—its own sum for the several products of $P_s\log_2(1/P_s)$. Let's compute the entropy for "time 2" of Figure 26.2. To get to time 2, there are four possible routes and associated probabilities: AA ($P_s=0.60\times 0.70=0.420$), AB ($P_s=0.60\times 0.30=0.180$), BA ($P_s=0.324$), and BB ($P_s=0.076$). Using those sequence probabilities, we compute the quantity $P_s\log_2(1/P_s)$ for each route.

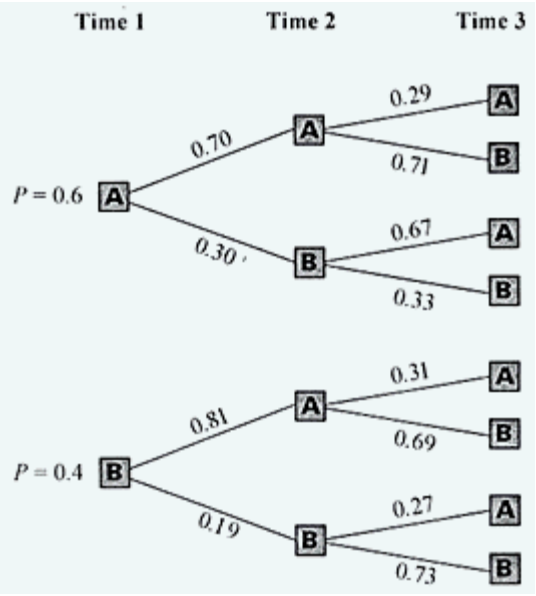


Figure 26.2
Branching pattern and hypothetical probabilities for two bins (A and B) over three time steps. Numbers along paths are conditional probabilities.

For instance, route AA has $0.420 \log_2(1/0.420)=0.526$. We then add up the four resulting numbers. That gives $H_{\Delta t}=0.526+0.445+0.527+0.283=1.781$, the entropy for time 2.

Over three consecutive measurements with two possible states there are eight possible routes (AAA, AAB, etc.). For each of the eight sequence probabilities (0.122, 0.298, 0.121, etc. in the example of Fig. 26.2) we calculate $P_s \log_2(1/P_s)$, then add up the resulting eight values. That sum is the information entropy for time 3.

In practice, the computer estimates sequence probabilities from the basic data. We need the probability of every string that leads to each successive time. That is, any one sequence probability applies only to a unique route. For example, for time 2 we count the number of occasions the system followed each of the four possible routes. Then we add up the occurrences for all routes to get a grand total. Finally, we divide the number for each route by the grand total. That produces the sequence probabilities to use in computing the entropy for that time. Then we repeat that exercise for each subsequent time, in turn, to get its associated sequence probabilities.

The total number of required sequence probabilities can be very large. Therefore, we may well need a large dataset and much computer time. Efficient histogramming methods, such as a bin sort on a binary tree (Gershenfeld 1993), can reduce computer time.

Sequence probabilities in chaos analyses usually are based on lagged values of a single variable. That means pseudo phase space. As with attractor reconstruction, the two aspects that we can vary are the lag itself and the number of phase space dimensions.

- Secondly, K-S entropy represents a *rate*. The distinctive or indicative feature about a dynamical regime isn't entropy by itself but rather the entropy rate. A rate (actually an average rate) is any quantity divided by the time during which it applies. Hence, any computed entropy divided by the associated time gives an average entropy rate—an average entropy per unit time:

$$\text{average entropy rate} = H_{\Delta t}/\text{time} = \sum_{i=1}^{N_t} P_S \log (1/P_S) / (\text{time}). \quad (26.5)$$

Claude Shannon, in the classic treatise *The mathematical theory of communication* (Shannon & Weaver 1949: 55, theorem 5), seems to have been the first to articulate the concept of entropy rate.

- Thirdly, K-S entropy is a *limiting value*. For discrete systems or observations, two limits are involved, nam time lengthens to infinity ($\lim_{t\to\infty}$) and bin size ε (width of classes, as in a histogram) shrinks to zero ($\lim_{\varepsilon\to 0}$).

Based on the above three features, **K-S entropy** (H_{KS}) is the average entropy per unit time *at the limiting conditions of time going to infinity and of box size going to zero*. In symbols,

$$\begin{aligned} H_{KS} &= \lim_{\varepsilon\to 0}\lim_{t\to\infty} \text{of entropy rate} \\ &= \lim_{\varepsilon\to 0}\lim_{t\to\infty} (H_{\Delta t}/time) \\ &= \lim_{\varepsilon\to 0}\lim_{t\to\infty}\left[\sum_{i=1}^{N_t} P_S \log (1/P_S)\right]/time. \end{aligned} \qquad (26.6)$$

There are several ways to interpret K-S entropy. All of them stem from Table 26.1. For instance, K-S entropy measures the average amount of new information or knowledge gained per measurement or per unit time (Farmer 1982). In other words, it's the average temporal rate of production, creation or growth of information contained in a series of measurements (Fraser 1986, Atmanspacher & Scheingraber 1987, Gershenfeld 1988). Applied to predictions, K-S entropy is:

- the average amount of uncertainty in predicting the next *n* events (Young 1983)
- the average rate at which the accuracy of a prediction decays as prediction time increases, that is, the rate at which predictability will be lost (Farmer 1982)
- the average rate at which information about the state of the system is lost (Schuster 1988).

Estimating K-S Entropy

Equation 26.6 tells us to do two general types of operations. It says to first arbitrarily choose a box size ε and, holding ε constant, see what happens to the entropy rate as time gets bigger and goes to infinity. That gives the so-called "inner limit" of Equation 26.6. The first steps or computations to make for that purpose are:

1. Choose a bin width ε.
2. Use the basic data to estimate sequence probabilities for all possible routes, for successive times.
3. For each time, compute entropy as

$$\sum_{i=1}^{N_t} P_S \log (1/P_S).$$

possible routes that aren't represented all the way to the time in question have a probability of zero (based on the data on hand) and so can be neglected.

There are two ways to use those results to get the inner limit of Equation 26.6. One way is to divide each entropy by the associated time to get entropy rate, plot entropy rate versus time, and estimate the asymptotic entropy rate as time increases. That method takes many time events to reach the asymptotic limit, that is, to converge. In fact, for practical purposes it's computationally impossible in most cases because of the huge amount of data needed.

The other way needs fewer time steps to show the asymptotic limit (i.e. it converges faster). It uses an alternative relation that Shannon gave. That alternative expression (Shannon & Weaver 1949: 55, theorem 6) defines an entropy difference:

$$\textit{entropy difference} = t(H_t/t)-(t-1)(H_{t-1}/[t-1]) \qquad (26.7)$$

where H_t is entropy as of time t, and t-1 refers to the observation just prior to time t. Within the first term on the right, namely $t(H_t/t)$, the two t's cancel out, leaving just H_t. In the same way, t-1 cancels out in the second product on the right, leaving just H_{t-1}. Equation 26.7 therefore reduces to simply

$$\textit{entropy difference} = H_t\text{-}H_{t-1} \qquad (26.8)$$

That result is nothing more than our friend, a first-differencing. It uses just the entropies themselves, not the entropy rates. At time 1, there isn't any entropy difference because there's no preceding entropy to subtract. At time 2, H_t-H_{t-1} is the entropy at time 2 minus the entropy at time 1; at time 3, H_t-H_{t-1}=entropy at time 3 minus that at time 2; and so on.

Shannon said that both the average entropy rate $H_{\Delta t}/t$ and the entropy difference H_t-H_{t-1}, when plotted against time, converge to the same asymptotic value of $H_{\Delta t}/t$. He also stated that H_t-H_{t-1}, "is the better approximation." For our purposes, what's important is that H_t-H_{t-1} becomes asymptotic sooner (Fig. 26.3). It therefore requires fewer data, less computer power, and so on. So, in practice it's probably better to use Equation 26.8, plot H_t-H_{t-1} versus time and take the asymptotic H_t-H_{t-1} as the asymptotic $H_{\Delta t}/t$.

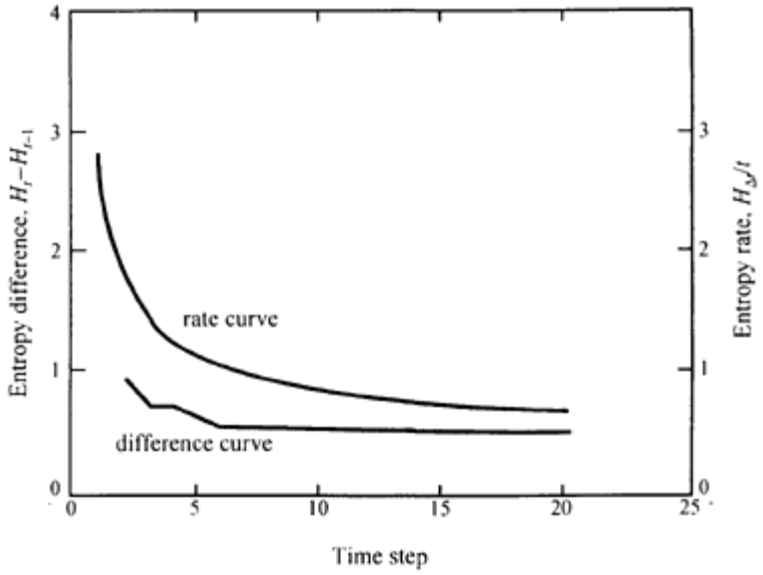


Figure 26.3
Decrease of entropy difference H_t-H_{t-1} (the "difference curve") and entropy rate H_{Dt}/t (the "rate curve") with time. Data are for logistic equation with k = 3.7 (hence chaotic regime), using 10 bins and 1000000 values.

An asymptotic $(\lim_{t\to\infty})$ entropy rate applies only to the particular bin size chosen. The second major operation that Equation 26.6 requires is to repeat the entire procedure and get estimates of the inner limit for smaller and smaller bin sizes $(\lim_{\varepsilon\to 0})$. The formula for K-S entropy with this approach therefore is:

$$H_{KS} = \lim_{\varepsilon\to 0}\lim_{t\to\infty}(H_t - H_{t-1}). \qquad (26.9)$$

In the next chapter I'll mention other ways to estimate K-S entropy.

Some Features that Influence Computed Values of Entropy

As we've seen, computing entropy consists largely of just adding up a bunch of $P \log(1/P)$ values. However, anything that affects those P values can also affect the value of $P \log(1/P)$ and hence also the entropy. For instance, Figure 26.1 shows an example of how entropy can vary with the distribution of probabilities. Four other influential factors are the number (width) of bins, size of dataset, lag, and number of sequential events (or embedding dimension). Let's look at each of those:

- *Bin width (number of possible states)* Table 26.2 (cols 2 and 4) shows, for the case of equal probabilities, the different entropies that result from different partitionings or numbers of bins. Figure 26.4, based on the logistic equation with k=3.7 and a lag of one, shows the different curves of H_t-H_{t-1} versus time for different numbers of possible states (bin widths). Here there isn't a smooth progression toward an asymptotic limit as bin size decreases. Instead, for these particular data the relations for 2, 10, and 20 bins all seem to become asymptotic to roughly the same value of H_t-H_{t-1} (about 0.47-0.48). (Extrapolating the values is somewhat uncertain.) However, the relation for 5 bins in Figure 26.4 becomes asymptotic to a noticeably lower value, namely about 0.43.

  Histogramming actually may not be the best procedure for estimating probabilities. Other methods (Ch. 6) are becoming popular.

- *Size of dataset* Sample size influences probabilities (and hence entropies) in that a short record (small dataset) probably won't adequately sample the system's long-term evolution. Probabilities (all types) for scanty datasets won't be representative. They change as we collect more measurements. Some important bins may not get sampled at all.

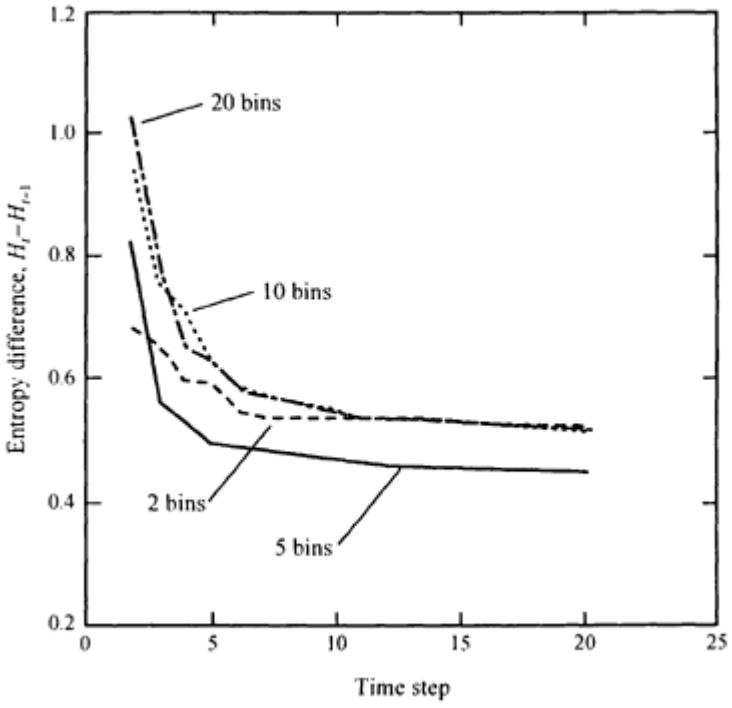


Figure 26.4
Change of entropy difference with time for different partitionings (numbers of bins). Data are for logistic equation with k = 3.7, using 1000000 values.

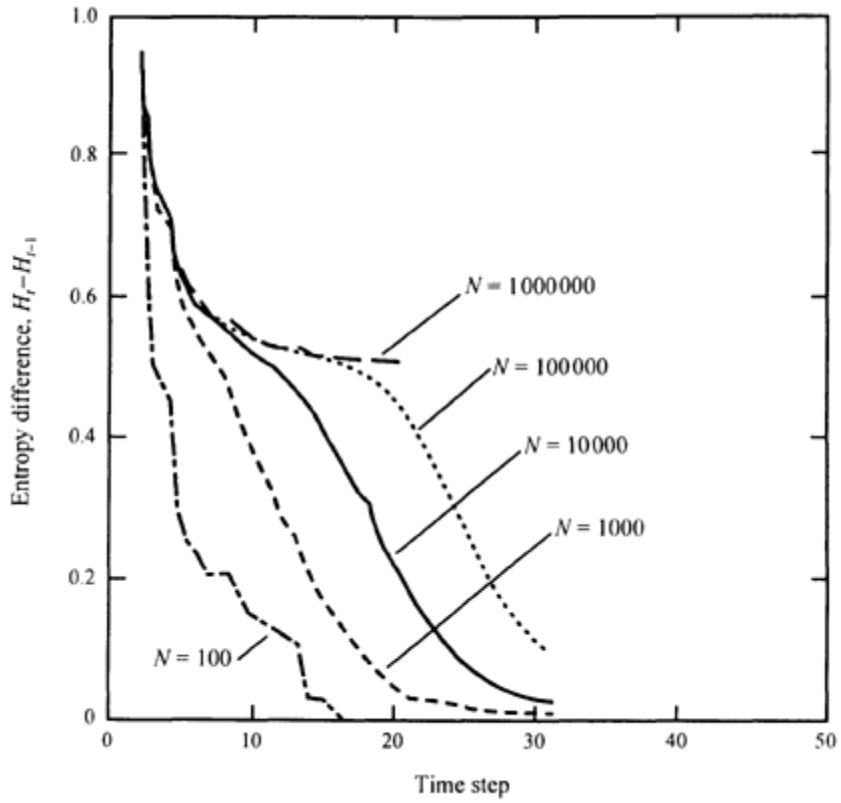


Figure 26.5
Effect of sample size on the relation between entropy difference H_t-H_{t-1} and time. Data are for logistic equation with k = 3.7, using 10 bins.

Figure 26.5 shows some typical effects of sample size in the K-S entropy step of plotting H_t-H_{t-1} against time. Data represent iterations of the logistic equation with k=3.7, using ten bins and sample sizes ranging from 100 to 1 million (about the most my computer can handle). The relation for a dataset of 1000000 values seems to become asymptotic toward a constant, positive entropy difference of about 0.47-0.48 (as was also shown in Fig. 26.4). The relation for *N=100000 seems to follow that same curve for about 15 time steps but then veers downward. Relations for datasets of 10000 or less offer no hope (in this example) of indicating the asymptotic entropy difference. In fact, we'd have a final entropy difference of zero (rather than the correct value, possibly 0.47-0.48) for all cases except where we have at least a million observations. And that's for a relatively "clean" (noiseless) dataset.*

• *Lag* Lag specifies the particular subset of data values used in estimating probabilities. Different subsets can lead to different sequence probabilities. For instance, using a lag of one, probabilities depend on the measured sequence x_t, x_{t+1}, x_{t+2}, and so on. Using a lag of two defines a different sequence of measurements, namely x_t, x_{t+2}, x_{t+4}, and so on. Sequence probabilities of getting the two different series of numbers very likely won't be the same, and therefore neither will the K-S entropy.

• *Embedding dimension (number of sequential measurements involved in the vector)* Sequence probabilities used in computing entropy (and hence the resulting value of K-S entropy) change as number of events (measuring occasions) increases. Most chaos calculations are in a context of pseudo (lagged) phase space, so they usually deal with lagged values of a single feature. Each additional lagged value attached to the vector is tantamount to another embedding dimension. That is, embedding dimension plays the same role as time or number of events. Each additional dimension or event brings an entire new set of phase space routes (Fig. 26.2) and sequence probabilities. Hence, K-S entropy can change with embedding dimension or time (Figs 26.3, 26.4, 26.5).

For a given number of possible states, the number of routes increases geometrically with number of lag space dimensions, in a manner analogous to Equation 6.6. In the present context, Equation 6.6 says that the *number of routes* (here probabilities)=$states^{embedding\ dimension}$.

That power-law relation plays a major role in the calculation of K-S entropy. Estimating the sequence probabilities for each of the various possible routes can require datasets numbering in the millions. Such large amounts of data are unrealistic for many of us. Consequently, researchers have pursued other ways to estimate K-S entropy. One proposal is to calculate and sum the positive Lyapunov exponents (based on the so-called Pesin theorem). Another is to do a mutual-information analysis, the subject of the next chapter. Both of those alternate approaches need further development, at least for applications to real-world data.

Identifying the Dynamical Regime

Let's now see how K-S entropy can distinguish between various types of data. We'll take the easy way and look only at noiseless data. In practice, of course, most data aren't noiseless. And therein lies the rub. One step in getting K-S entropy involves taking the limit as bin size goes to zero (Eq. 26.6). In practice, once bin size diminishes to the noise level (the error in the measurements), further decreases in bin size are meaningless. Our estimate of K-S entropy then becomes less accurate. The mutual-information approach discussed in the next chapter is an initial attempt to get around this problem.

The basic equation we'll use to estimate K-S entropy is Equation 26.9:

$$H_{KS} = \lim_{\varepsilon \to 0} \lim_{t \to \infty} (H_t - H_{t-1}).$$

H_t, is entropy at time t, computed as

$$\sum_{i=1}^{N_t} P_S \log (1/P_S),$$

and H_{t-1} is the entropy at time t-1. Our recipe therefore is as follows:

1. Choose bin size ε.
2. For a given time step, estimate P_s for each route and compute H_{t-1} as the sum of $P_s \log(1/P_s)$ for all routes.
3. Repeat step 2 for the next time step and compute the entropy for that time, H_t.
4. Subtract to get the quantity H_t-H_{t-1}.
5. Repeat steps 2-4 for higher successive times.
6. See what happens to H_t-H_{t-1} as time increases.
7. Repeat steps 1-6 for a smaller bin size, then a still smaller bin size, and so on, to see what happens as ε gets very small. (This step may not always be necessary.)

Fixed points

The logistic equation helps show the value of K-S entropy for various nonchaotic attractors. Let's start with ten compartments (step 1 from the above list). The logistic equation yields a fixed-point attractor when the control parameter k is less than 3. That is, for k<3 all iterations (once the transients die out) yield a constant fixed value. For instance, with k=2.8 the fixed point is at x*=0.643 (Ch. 10). Excluding transients, the time series consists of a string of values of 0.643 (or of whatever value the fixed point has). Every route, in other words, consists of endless values of the same number. Sequence probabilities therefore are always 1.0. Log(1/P) therefore is log(1), which is zero, so $P_s \log(1/P_s)$ is zero and entropies for all time steps are zero (steps 2-6). That takes care of the first limit in the K-S entropy equation.

Now for the second (final) limit, where bin size goes to zero (step 7). The single attractor value always falls in the same bin, for a given number of bins. The system is sure to go to that bin. All sequence probabilities therefore are always 1.0. In other words, using smaller and smaller bins has no effect. So, the K-S entropy—the limit as time goes to infinity and bin size goes to zero—is zero for a fixed-point attractor.

Period-two attractor

Now we'll increase k and enter the period-doubling zone. The simplest case is period two. As usual, we'll neglect transients. At k=3.4, the system alternates between the two fixed values of x*=0.452 (bin boundaries 0.4 to 0.5, for a 10-bin partition) and x*=0.842 (bin boundaries 0.8 to 0.9). The first limit to get is the limit as time goes to infinity, for constant bin size. We'll use ten bins (step 1). The first calculations are for time 1. The two possible routes for time 1 are into bin 5 (x* between 0.4 and 0.5) or into bin 9 (x* within 0.8 to 0.9). At any arbitrary starting time, there's a 50 per cent chance (an ordinary probability P of 0.5) that the system is in either box. The quantity $P \log_2(1/P)$ for each bin is 0.5 $\log_2$(1/0.5) or 0.500. Summing the two values gives H_{t-1}=0.500+0.500=1.0 (step 2).

Step 3 is to do the same computations over a duration of two time increments, thus getting H_t. Only two of our ten compartments (bins 5 and 9) are possible for our period-two attractor. Regardless of which bin the system goes to at time 1, it's sure (probability=1) to go to the other at time 2. In other words, the only routes possible over two events are of the sort AB and BA. The sequence probability for each of those routes is 0.5 (for time 1) × 1 (for time 2), or 0.5. The two values of 0.5 for *P*s $\log(1/P_s)$ sum to 1 (step 3). That is, entropy again is 1.0. Furthermore, all subsequent conditional probabilities are always 1.0 because, given one of the two attractor values, we know the value at the next time. Entropy therefore has a constant value of 1.0 for all times. Subtracting the entropy of any one time from that of the next time (step 4) gives an entropy difference of zero, at every time (steps 5 and 6). So, our first limit is zero.

The second limit (step 7) involves decreasing the bin size to very small. As with the fixed-point case, decreasing the bin size has no effect because the two attractor values always fall into just two known bins. K-S entropy therefore is zero for a period-two attractor.

Higher-periodicity attractors

Again let's start with bin sizes ε of 0.1 (step 1). A higher k of 3.52 produces a periodicity of four (with sequential x^* values of 0.879, 0.373, 0.823, and 0.512). That's a fixed succession. For instance, if an observation (say 0.373) is in bin 4 (x^* between 0.3 and 0.4), the system is certain to go next to bin 9 (x^* between 0.8 and 0.9), because the next attractor value in the sequence is 0.823. If in bin 6 (x^* between 0.5 and 0.6, such as for 0.512), the system is also certain to go next to bin 9, because the next attractor value in the sequence is 0.879. The probability for such certain paths is 1. If, however, we know only that an observation is in bin 9, we wouldn't know whether the actual value is 0.879 or 0.823. Hence, we have to assume a probability of 0.5 that the system goes next to bin 4 (*x^*=0.373) and also a probability of 0.5 that it goes instead to bin 6 (x^*=0.512). The branching pattern going to time 2 therefore looks like that shown in Figure 26.6.*

What are the sequence probabilities, entropies, and entropy differences for such data?

- At time 1, ordinary probabilities are 0.25, 0.25, and 0.50 that the system might be in bins 4, 6, or 9, respectively (Fig. 26.6). The three corresponding values of $P \log_2(1/P)$ each compute to 0.500. Summing these gives an entropy of 1.500 (step 2).

- At time 2, there are four possible routes. If the system is in bins 4 (0.373) or 6 (0.512) at time 1, the routes to time 2 are certain (namely, to bin 9). Sequence probabilities for each of those two routes are 0.25 × 1.0=0.25. $P_s \log_2(1/P_s)$ is 0.500 for each. The remaining possible bin at time 1 is bin 9 (possible values of 0.823 or 0.879). For each value, a route leads to time 2. The two possible routes from bin 9 each have a conditional probability of 0.5, so the associated values of P_s are each 0.50 × 0.50=0.25. $P_s \log_2(1/P_s)$ for each of those two routes is 0.500. All four values of $P_s \log_2(1/P_s)$ therefore are 0.500. The sum (the entropy) is 2.0 (step 3). Subtracting the entropy at time 1 (1.500) from that at time 2 (2.0) gives 2-1.5=0.5 (step 4).

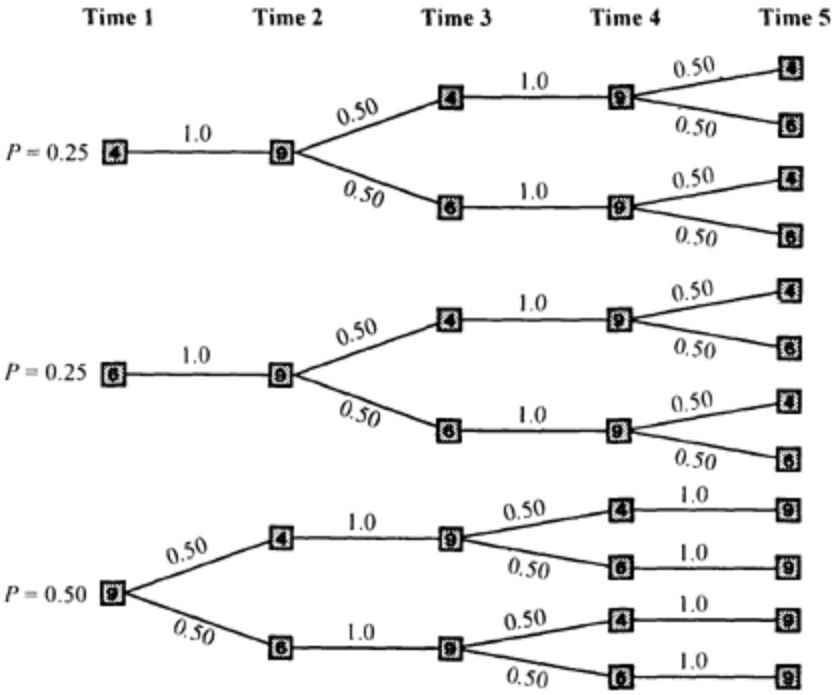


Figure 26.6
Branching patterns for five iterations of the logistic equation at period four (k = 3.52). Range of 0-1 was divided into 10 bins. Sequential attractor values are 0.373 (bin 4), 0.823 (bin 9), 0.512 (bin 6), and 0.879 (bin 9). Boxed numbers are bin numbers; numbers along individual routes are probabilities.

- Six possible routes lead to time 3 (Fig. 26.6). As always, the task is to determine sequence probabilities and associated values of $P_s \log_2(1/P_s)$ for each route, then sum the results to get the entropy. Entropy here turns out to be 2.0 (step 3). Subtracting the preceding entropy (2) gives 2-2=0 (step 4). In fact, entropy for each succeeding time step also turns out to be 2.0 (step 5). Entropy differences therefore are zero at each time (step 6). Also, decreasing the bin size (step 7) again has no effect because of the fixed attractor values.

As with periodicities of two and four, K-S entropy also is zero for higher periodicities.

Summarizing, K-S entropy is zero for logistic-equation nonchaotic attractors. Nonchaotic systems don't evolve with time. The attractor consists either of one value (a point attractor), a group of values regularly repeated with time (periodic attractor or limit cycle), or a torus on which the frequencies are either periodic or quasiperiodic. Once we identify the attractor, we can reliably predict further results with virtually zero uncertainty. Entropy stays constant with time. The K-S entropy of zero means that we don't gain any new information over time.

Chaotic regime

We've already done some sample computations for the chaotic regime (Figs 26.3, 26.4, 26.5). A chaotic system continually evolves with time. Compared to any chosen time, the system at a later time is in an unpredictably different state. That unpredictable evolution provides a steady supply of new information. In other words, the K-S entropy for chaotic systems is some positive constant. For example, Figures 26.3, 26.4, and 26.5 show that the logistic equation's chaotic regime at k=3.7 has a K-S entropy of about 0.47-0.48.

Even with unlimited amounts of noiseless data, extrapolating an asymptotic curve to estimate the limiting value is not easy. For instance, on Figures 26.3-5 it's no problem getting to within 0.1 of the approximate asymptotic value. However, we probably should question values in the literature that are reported to the nearest 0.01.

Random data

K-S entropy for random data can best be understood by looking at sample computations for the two limits in the K-S entropy equation (Eq. 26.9). (As usual, "random" here means that any determinism is essentially negligible. In addition, the treatment here assumes that the underlying probability distribution is uniform.) The first limit is time going to infinity. The second is bin size going to zero.

limit as time goes to infinity

Finding this first limit requires keeping the number of possible states constant. Let's use two states (step 1). For a uniform distribution, probabilities are equal of the system going to either state. At time 1, the ordinary probabilities are 0.5 for each state. $P \log_2(1/P)$ for each state then is $0.5 \log_2(1/0.5)$ or 0.500. Entropy H_{t-1} for time 1 is the sum of the two values of $P \log_2(1/P)$, or 0.500+0.500, or 1.0 (step 2).

Time 2 and later times require sequence probabilities (P_s) in the entropy equation. There are four possible sequences or routes to time 2. Regardless of which of the two states the system went to at time 1, it has an equal chance of going to either state at time 2. In other words, each of the four routes is equally likely. Each sequence probability therefore is 0.50×0.50 or 0.25. $P_s \log_2(1/P_s)$ for each of the four routes becomes $0.25 \log_2(1/0.25)$ or 0.500. Adding the four values of 0.500, entropy H_t=2.0 (step 3). Subtracting the entropy at time 1 (1.0) from that at time 2 (2.0) gives an entropy difference of 2.0-1.0 or 1 (step 4).

Step 5 is to move up one time step and make similar calculations. At time 3, there are eight possible routes, again all equally likely. Each sequence probability now is $0.50 \times 0.50 \times 0.50$=0.125. $P_s \log_2(1/P_s)$ is $0.125 \log_2(1/0.125)$ or 0.375. Adding up eight such values gives a new H_t of 3.0. Entropy difference now is entropy at time 3 minus entropy at time 2=3.0-2.0=1—exactly the same value as for the preceding time step.

Similar computations to include more successive observations yield the same result: at each time or iteration, H_t-H_{t-1} is 1.0. Thus, for two states the limit of H_t-H_{t-1} as time goes to infinity (step 6) is a constant, namely 1.0. The remaining job is to find out what happens to that constant as we increase the number of possible states (i.e. as bin size goes to zero) (step 7).

limit as bin size goes to zero

In the foregoing discussion, we used only two states and got H_t-H_{t-1}=1.0 for any and all times. Now let's divide our interval of 0 to 1 into three possible states instead of two. For each time, computing the entropy and subtracting that of the preceding time again produces a constant value of H_t-H_{t-1}. However, the constant now is 1.585 instead of 1.0. Using four possible states produces a constant difference (H_t-H_{t-1}) of 2.0 for all times. In fact, each constant is the log (here base 2) of N_s, where N_s is the number of possible states. ($\text{Log}_2 2$=1.0; $\log_2 3$=1.585; $\log_2 4$=2.0; etc.) Hence, for uniformly distributed data, H_t-H_{t-1}=logN_s.

Since H_t-H_{t-1}=log of number of states, H_t-H_{t-1} increases as long as the number of possible states increases, that is, as long as bin size ε decreases. The increase is at a steady, arithmetic rate as N_s increases geometrically. For example, as N_s progresses from 1 to 10 to 100, H_t-H_{t-1} goes from 0 to 3.32 to 6.64, etc. Thus, H_t-H_{t-1} doesn't become asymptotic to any constant. When the classes become infinitesimally narrow and approach zero in width, the number of possible states becomes infinitely large. H_t-H_{t-1} then approaches infinity. Hence, for random data, K-S entropy is infinite.

Let's review briefly the relation between H_t-H_{t-1} and time (number of iterations) for the various types of data. For pre-chaotic data, the relation is zero right from the start (time 1), so K-S entropy is zero. For chaotic data, the curve starts at a relatively high value of H_t-H_{t-1} and decays by lesser and lesser amounts with time, becoming asymptotic toward some positive constant. Thus, K-S entropy is positive. For random and uniformly distributed data, the relation increases regularly (logarithmically) with number of states, so K-S entropy is infinite. In practice, noise, insufficient data, and computer capability complicate attempts to distinguish the three classes. On the other hand, K-S entropy for each category is different enough that it can sometimes be helpful.

Summary

"Entropy" has many meanings or interpretations. They include the proportion of energy available for doing work, disorder, probability of an event, uniformity of a distribution, uncertainty, randomness or unpredictability, freedom of choice, diversity, surprise, information, amount of information needed to specify the state of a system, and accuracy of data. Probabilities (relative frequencies) are the basic data used in calculating entropy or information; high probability implies a small amount of information, and vice versa. When all states of a system are equally probable, entropy increases as number of choices (phase space compartments) increases. When the states aren't all equally probable, entropy is

$$\sum_{i=1}^{N_s} P_i \log (1/P_i),$$

where P_i is the probability of the ith compartment. In that formula, entropy varies with number of compartments, distribution of probabilities among the compartments, and with value of control parameter.

Entropy also has various other general properties (see also Krippendorff 1986: 16ff.). For example, its value is always zero or positive. Entropy is zero (a minimum) when any bin has a probability of one (absolute certainty). It's a maximum for a uniform distribution (equally probable outcomes). Its value for that equally probable outcome case increases with number of possible states. Entropy doesn't care what the system or process is—it only evaluates a set of measured frequencies or probabilities. That is, it's a number that characterizes an ensemble of probabilities. That means, also, that it makes no *a priori* assumptions about the nature of the frequency distribution (that means it's nonparametric). An entropy is an average (a weighted sum). Finally, its generality has led to applications in many fields. including engineering, physics, chemistry, biology, sociology, Earth sciences, literature, art, journalism, television, and others.

Information entropy is always positive and finite. By itself, it doesn't identify chaos. To identify chaos, we need entropy rate (the average entropy per unit time) at the limiting conditions of time increasing toward infinity and of box size diminishing toward zero. That special case of entropy is called Kolmogorov-Sinai (or K-S) entropy (H_{KS}). It requires sequence probabilities. In symbols, K-S entropy

$$\begin{aligned} H_{KS} &= \lim_{\varepsilon \to 0} \lim_{t \to \infty} (H_{\Delta t}/time) \\ &= \lim_{\varepsilon \to 0} \lim_{t \to \infty} \left[\sum_{i=1}^{N_s} P_S \log (1/P_S) \right] /time \\ &= \lim_{\varepsilon \to 0} \lim_{t \to \infty} (H_t - H_{t-1}). \end{aligned}$$

K-S entropy quantifies how chaotic a dynamical system is: H_{KS} is zero for a deterministic system that isn't in chaos; it's a positive constant for a chaotic system; and it's infinite for a random process, at least for uniformly distributed data.

Chapter 27
Mutual information and redundancy

The idea of "information" (Ch. 6) is becoming very important in chaos theory. As long as there is some relation between two variables, one contains information about the other. Take, for example, the Christmas season. Sales of toys are highest by far at that time of year. Looking at a list of the volume of sales throughout the year, we can immediately recognize the Christmas season, even if we're not given any associated months of the sales. Measuring just one variable (either month or sales) gives *some* information about the other. (Low sales indicate that it's not Christmas time, and vice versa.) Similarly, the amount of water in a river (Ch. 9) typically rises to a peak in the springtime, then recedes to an annual low in late summer (at least where I live). A few sequential measurements of just one of the variables (either time of year or streamflow) give us some information about the other.

Mutual information, defined explicitly later in the chapter, appears in many fields. Like entropy, it's a commonly studied property of dynamical systems. Potential applications in chaos theory include estimating the optimum embedding dimension for attractor reconstruction (Fraser 1989a), estimating the accuracy of data (Fraser 1989a), indicating whether K-S entropy is zero or positive (and hence in revealing chaos) (Fraser 1989a, Paluš 1993), finding whether a time series is periodic, chaotic, or random, estimating the optimum lag to use in attractor reconstruction (Fraser 1989a, Pineda & Sommerer 1993), estimating how far into the future we can make reliable predictions, and testing data for nonlinearity (Paluš 1993). We'll discuss those research directions briefly, near the end of the chapter. However, most of the chapter explains the fundamentals of mutual information. The early part of the chapter establishes our foundation or tools. (We feel much more comfortable using a tool when we know how it's put together and what it's made from.) As with the entropy chapter, we'll rely very heavily on the probability and information material of Chapter 6.

Mutual information has several ingredients. Those ingredients are specialized but straightforward variants of entropy. The entropy or information discussed in the last chapter is sometimes called **self-entropy** or **self-information**, because it's the entropy of one system or variable. In other words, it's the information one system reveals about itself. This chapter extends our treatment of entropy from one system to two or more systems or dimensions. Arbitrarily, we'll couple them.

The "systems" can be anything—babies' heights and weights, dice, and so on. Also, coupling or combining systems doesn't imply anything about when we measured the data for each. We may have measured the systems at the same time (babies' heights and weights), at different times, or sequentially (a time series of one physical feature).

Joint and Conditional Entropies

Joint and conditional entropies, as defined in this chapter and as used in information theory, apply to *two* systems or dimensions; for more than two dimensions (discussed later in the chapter), terminology is still developing. For simplicity, we'll usually deal here with two systems. The first, system X, has discrete observations x_i; the second, system Y, has values y_j.

Chapter 6 discussed joint probability—the probability of seeing particular values of two or more variables in some sort of combination. In fact, we rolled a pair of dice (one white and one green, representing separate systems) and built a joint probability distribution. The joint probability distribution gave the probability of the joint occurrence of each possible combination of values of the two systems (Fig. 6.6). Since entropy is based solely on probability, the idea of joint probability easily applies to a so-called **joint entropy**—the average amount of information obtained (or uncertainty reduced) by individual measurements of two or more systems.

Having built a word concept of joint entropy, let's follow our standard procedure and express it in symbols. There are two cases. In the first, knowing value *x* from system X has no effect on the entropy of system Y. (In other words, individual observations *x* and *y* aren't mutually related.) In the second, knowing *x* affects our estimate of the entropy of system Y (and so *x* and *y* have some kind of association). For each case, the symbols (and hence the form of the definition) can be either those of probabilities or of entropies.

Mutually unrelated systems

Let's begin by talking about the first case, in which *x* and *y* aren't mutually related. We can express joint entropy in terms of probabilities or in terms of entropies. First we'll do the probability form, then the entropy form.

Probability Form

In probability terms, the foundation for the symbol definition of joint entropy for two unassociated but coupled systems is Equation 6.17c. That is the general one-system equation derived in Chapter 6 for information (or entropy),

$$I_w = -\sum_{i=1}^{N_s} P_i \log P_i . \qquad \text{(6.17c, repeated)}$$

In that equation, I_w is information, N_s is the number of nonzero probabilities, and P_i is the ordinary probability of the *i*th bin. First we'll apply that general equation just to system X. Since it'll apply to just one system, we'll call it "self-entropy" instead of "information." We therefore replace symbol I_w with H_X (the self-entropy of system X) and the probability symbol P_i with $P(x_i)$ to indicate the probability of bin x_i. Equation 6.17c as rewritten for our system X then becomes

$$H_X = -\sum_{i=1}^{N_x} P(x_i) \log_2 P(x_i) \qquad (27.1)$$

in which the sum is over all bins that have at least one observation (and so have a nonzero probability).

Now for system Y. By analogy with Equation 27.1, we'll use $P(y_j)$ for the probability of bin y_j of system Y. The self-entropy H_Y for system Y then is

$$H_Y = -\sum_{j=1}^{N_y} P(y_j) \log_2 P(y_j) \qquad (27.2)$$

Now we want an equation for the joint entropy of the "combined" system of *x* and Y. Such an equation will have the same form as Equations 27.1 and 27.2. We only need to make two changes in Equations 27.1 and 27.2. First, the appropriate probabilities are no longer those of x_i or y_j alone. Instead, they are now joint probabilities-the chances associated with the joint occurrence of each possible combination of *x* and *y*. A statistician would say they are the joint probabilities that, for any given time or sequence, $x=x_i$ and $y=y_j$. The global symbol $P(x_i,y_j)$ represents the entire group of such joint probabilities. Thus, x_i and y_j are bin indices that identify any bin in the joint probability distribution.

The second change is that we now sum over all occupied bins in the joint probability distribution. Three of the more common ways of symbolizing that summation are

$$\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}, \quad \sum_{i}\sum_{j}, \text{ and } \sum_{ij}.$$

I'll use the first one.

With those two changes, we use the same format as Equations 27.1 and 27.2 to express a joint entropy in probability terms. The joint entropy of the two systems, $H_{X,Y}$, here is for mutually unrelated systems:

$$H_{X,Y} = -\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P(x_i,y_j) \log_2 P(x_i,y_j) \qquad (27.3)$$

Computing $H_{X,Y}$ in practice involves first preparing a joint-distribution table by empirically determining the joint probability of obtaining each possible combination of *x* and *y*, just as we did with the dice in Figure 6.6. We estimate that joint probability by looking at our basic data and counting the number of occurrences of each combination of classes of *x* and *y*, then dividing by the total number of all occurrences in the entire dataset. Next, Equation 27.3 says we need to calculate the quantity $P(x,y) \log_2 P(x,y)$ for every combination represented. Lastly, Equation 27.3 tells us to add up all those products.

The probabilities to use in computing joint entropy can be tricky. Although they are joint probabilities, they can be based on "simultaneous" associations or on sequential observations (Ch. 6). The most common situation in chaos theory is lag space. Observations x_i then become x_t, and observations y_j become x_{t+m}, where *m* is lag. In that case, joint probability $P(x_i,y_j)$ is the sequence variety of joint probability, $P(x_t, x_{t+m})$. That's the probability that one particular value of *x* follows another. It's comparable to successive-event routes AA, AB, BA, or BB, that we used as examples for two sequential time steps in Chapter 6. To compute the joint entropy, we have to estimate the sequence probability of each route or lag vector represented in the data. We do that by counting the number of occurrences of each particular route, then dividing each sum by the total number of occurrences of all observed two-event routes.

Entropy Form

An alternate form shows that joint entropy for two mutually unrelated systems is the sum of their self-entropies. To see that, we'll just algebraically manipulate the probability form (Eq. 27.3) such that certain probability expressions within the restructured equation represent entropies.

Suppose you and I each flip a coin once, at the same time. There are four possible outcomes: HH, HT, TH, and TT, where the first letter (H for heads, T for tails) stands for your result and the second letter my result. With four possible outcomes, we know intuitively that, any time we flip our coins, each of the four possible outcomes has a 1 in 4 chance of happening. In other words, each outcome (HH, etc.) has a joint probability of 0.25. Let's express that in a more mathematical way. Assume that any flip has a probability of 0.5 of being heads. The joint probability that both of us get heads (HH) therefore is $0.5 \times 0.5=0.25$ (the same as that of the other three possible outcomes). In other words, the joint probability of any specified outcome, or $P(x_i,y_j)$, equals the ordinary probability of the one value $P(x_i)$ (here, that would be 0.5 that you get heads) times the ordinary probability of the other, $P(y_j)$ (here, 0.5 that I also get heads). Generalizing that multiplicative rule and writing it in symbol form:

$$P(x_i,y_j) = P(x_i)P(y_j) \qquad (27.4)$$

Equation 27.4 lets us substitute the product $P(x_i)P(y_j)$ in place of $P(x_i,y_j)$ in Equation 27.3:

$$H_{X,Y} = -\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P(x_i)\ P(y_j) \log_2 [\ P(x_i)\ P(y_j)\ ]. \qquad (27.5)$$

There's a log term on the far right. It has the form log(ab), in which a is $P(x_i)$ and b is $P(y_j)$. Since log(ab)=loga+logb (Appendix), we can alter that log term and rewrite Equation 27.5 as

$$H_{X,Y} = -\sum_{i=1}^{N_x}\sum_{j=1}^{N_x} P(x_i)\ P(y_j)\ [\log_2 P(x_i) + \log_2 P(y_j)]. \qquad (27.6)$$

The equation in this form says, among other things, to multiply each of the two log terms by $P(x_i)P(y_j)$. That's analogous to ab (c+d), which is abc+abd. Making that change in Equation 27.6:

$$H_{X,Y} = -\sum_{i=1}^{N_x}\sum_{j=1}^{N_x} [P(x_i)\ P(y_j) \log_2 P(x_i) + P(x_i)\ P(y_j) \log_2 P(y_j)]. \qquad (27.7)$$

Everything to the right of the two summation signs is just the sum of two groups of products. Both summation signs apply to each of those two groups. We can show that by writing Equation 27.7 as follows:

$$_{X,Y} = -\sum_{i=1}^{N_x}\sum_{j=1}^{N_x} P(x_i)\ P(y_j) \log_2 P(x_i)$$

$$-\sum_{i=1}^{N_x}\sum_{j=1}^{N_x} P(x_i)\ P(y_j) \log_2 P(y_j)\ . \qquad (27.8)$$

The right-hand side of the equation now consists of two main parts, each involving

$$\sum_{i=1}^{N_x}\sum_{j=1}^{N_x}.$$

In that combined symbol,

$$\sum_{i=1}^{N_x}$$

applies to *x* values (as indicated by the subscript *i*), and

$$\sum_{j=1}^{N_x}$$

applies toy values (as indicated by the subscript *j*). Stating that same thing in symbols,

$$\sum_{i=1}^{N_x}\sum_{j=1}^{N_x} (x_i)\ (y_j) = \sum_{i=1}^{N_x} (x_i) \sum_{j=1}^{N_x} (y_j).$$

We therefore rewrite Equation 27.8 such that

$$\sum_{i=1}^{N_x}$$

more clearly applies to values involving *x* and

$$\sum_{j=1}^{N_x}$$

more clearly applies to values involving *y*:

$$H_{X,Y} = -\sum_{i=1}^{N_x} P(x_i) \log_2 P(x_i) \sum_{j=1}^{N_y} P(y_j) - \sum_{j=1}^{N_y} P(y_j) \log_2 P(y_j) \sum_{i=1}^{N_x} P(x_i) \quad (27.9)$$

The sum of the individual probabilities for any system is 1, so

$$\sum_{i=1}^{N_x} P(x_i) \text{ is 1 and } \sum_{j=1}^{N_y} P(y_j) \text{ is 1.}$$

Making those substitutions, Equation 27.9 reduces to

$$H_{X,Y} = -\sum_{i=1}^{N_x} P(x_i) \log_2 P(x_i) - \sum_{j=1}^{N_y} P(y_j) \log_2 P(y_j). \quad (27.10)$$

The first summed term is nothing more than the self-entropy H_X (Eq. 27.1). The second summed term is self-entropy H_Y (Eq. 27.2). Equation 27.10 therefore says simply that

$$H_{X,Y} = H_X + H_Y \quad (27.11)$$

And that finishes our task. We've got those probabilities we started with expressed as entropies. Thus, the joint entropy of two mutually unaffiliated systems or of a pair of unrelated variables is just the sum of their self-entropies. In terms of uncertainty, the uncertainty of a joint event *x*,*y* is the uncertainty of event *x* plus the uncertainty of event *y*, as long as *x* and *y* have no mutual association.

Mutually related systems

Now for the second of our two main cases, namely the case where *x* and *y* have some kind of mutual relation. We can again express joint entropy in terms of either probabilities or entropies.

Probability Form

Joint entropy as based on *x*,*y* pairs is a general concept that can apply to any two systems or variables, whether they're mutually affiliated or not. Equations 27.3 and 27.11 apply when they're totally disconnected. They might, on the other hand, have some kind of mutual association. For example, *x* might influence *y*. Even if the two processes aren't physically interdependent, the association means that measuring *x* gives some clue or information about the value of *y*. The probability associated with y_j therefore becomes a conditional probability—the probability of getting value y_j, given that value x_i has occurred (Ch. 6). In other words, when systems X and Y are mutually related we use a conditional probability for the second system (here Y), to reflect that affiliation.

Use of conditional probability means that $P(y_j)$ (the probability of getting the value y_j) for unrelated systems or variables (e.g. Eqs 27.2, 27.4) becomes $P(y_j|x_i)$ (the probability of getting y_j, given that x_i has occurred). So, Equation 27.4, namely $P(x_i,y_j)=P(x_i)P(y_j)$, changes to

$$P(x_i,y_j) = P(x_i)P(y_j|x_i) \quad (27.12)$$

As a result, the probability expression for joint entropy (Eq. 27.3) becomes

$$H_{X,Y} = -\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P(x_i, y_j|x_i) \log_2 P(x_i, y_j|x_i). \quad (27.13)$$

Entropy Form

Deriving the entropy form, as before, is just a matter of straightforward algebra. Rewriting the joint entropy of Equation 27.6 so that it applies to the mutually affiliated case:

$$H_{X,Y} = -\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P(x_i)\ P(y_j|x_i)\ [\log_2 P(x_i) + \log_2 P(y_j|x_i)]. \qquad (27.14)$$

The terms to the right of the summation signs in Equation 27.14 now consist of items having the form a(b+c). That's the same as ab+ac. Rewriting the equation that way while still applying the summations to everything:

$$H_{X,Y} = -\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P(x_i)\ P(y_j|x_i) \log_2 P(x_i)$$

$$-\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P(x_i)\ P(y_j|x_i) \log_2 P(y_j|x_i). \qquad (27.15)$$

The first pair of summation signs applies to the product of three items: $P(x_i)$ times $P(y_j|x_i)$ times $\log_2 P(x_i)$. Since the three items are multiplied, we can write them in any order. We'll write them as $P(x_i) \log_2 P(x_i) P(y_j|x_i)$. In that order, the first two probabilities involve *x* and the third involves *y*. With the product rearranged in that order, we now restate the entire first term so that each summation sign applies to its appropriate probabilities. The first summations to the right of the equals sign thereby become

$$-\sum_{i=1}^{N_x} P(x_i) \log_2 P(x_i) \sum_{j=1}^{N_y} P(y_j|x_i).$$

Each of the two summations in that expression refers to a familiar quantity. For instance, the first summation

$$-\sum_{i=1}^{N_x} P(x_i) \log_2 P(x_i),$$

is nothing more than our friend H_X, the self-entropy of system X (Eq. 27.1). The expression therefore reduces to

$$H_X \times \sum_{j=1}^{N_y} P(y_j|x_i).$$

The other summation in that expression is the sum of all the probabilities of *y* given *x*. Probabilities for any system sum to 1, so the latter sum is 1. The entire first term therefore is merely H_X times 1 or H_X. Inserting H_X in place of the first term in Equation 27.15 changes Equation 27.15 into

$$H_{X,Y} = H_X - \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P(x_i)\ P(y_j|x_i) \log_2 P(y_j|x_i). \qquad (27.16)$$

We can simplify that equation by rewriting the double summation. In keeping with their definitions, we'll associate the summation of *i*'s, symbolized as

$$\sum_{i=}^{N_x},$$

with the *x* probabilities and the summation of *j*'s

$$\sum_{i=1}^{N_x},$$

with the *y* probabilities. Rewriting the double summation that way gives

$$\sum_{i=1}^{N_x} P(x_i) \sum_{j=1}^{N_x} P(y_j|x_i) \log_2 P(y_j|x_i).$$

The first summation,

$$\sum_{i=1}^{N_x} P(x_i),$$

like all summations of probability distributions, is 1. The double summation therefore reduces to just

$$-\sum_{j=1}^{N_x} P(y_j|x_i) \log_2 P(y_j|x_i).$$

That term is a variation of the entropy of system Y as expressed in the last term of Equation 27.10. What's different about it now is that all *y* probabilities here are *conditional* probabilities, $P(y_j|x_i)$. For that reason, that revised entropy expression is a **conditional entropy**. A conditional entropy measures the average uncertainty in *y* from system Y, *given a measurement* of *x* from coupled system X. For two-dimensional lag space, conditional entropy measures the average uncertainty in x_{t+m}, given that we measured x_t at time *t*. (And again, the conditional probabilities in lag space are the probabilities that x_{t+m} will follow x_t.)

Let's use $H_{Y|X}$ as the symbol for the conditional entropy of Y,

$$-\sum_{j=1}^{N_x} P(y_j|x_i) \log_2 P(y_j|x_i).$$

Also, we'll retain $H_{X,Y}$ for the joint entropy of two systems (whether related or not). In terms of entropies, the joint entropy for two mutually associated systems (Eq. 27.16) then becomes

$$H_{X,Y} = H_X + H_{Y|X} \qquad (27.17)$$

The derivation is just as valid with the roles of the variables reversed. In that case, we symbolize the conditional entropy of *x* as $H_{X|Y}$. It represents the uncertainty in *x* as a result of knowing *y*. We then have

$$H_{Y,X} = H_Y + H_{X|Y}. \qquad (27.18)$$

Equations 27.17 and 27.18 are both analogous to Equation 27.11 ($H_{X,Y}=H_X+H_Y$), the disassociated case. Equations 27.17 and 27.18 simply say that the joint entropy of two mutually related systems is the self-entropy of one system plus the conditional entropy of the other. In terms of uncertainty: the uncertainty of the joint event *x*,*y* is the sum of the uncertainty of *x* plus the uncertainty of *y* when we know *x*. Hence, for both the mutually unrelated and related cases joint entropy is the sum of two self-entropies or uncertainties—the uncertainty in *x* and the uncertainty in *y*.

Recapitulation

Let's summarize our two main accomplishments so far in this chapter.

1. We've developed four expressions for joint entropy. Two of them (a probability form and an entropy form) are for systems that have no mutual relation. The other two are counterpart forms for systems that are related. The probability form of joint entropy for two unassociated systems is Equation 27.3. The entropy form is Equation 27.11 (the sum of the two self-entropies). The probability form of joint entropy for two mutually associated systems is Equation 27.13. The entropy form is either of 27.17 or 27.18, that is, the self-entropy of one system plus the conditional entropy of the other.

2. We've stipulated what *conditional entropy* (e.g. $H_{Y|X}$) is. Our discussion actually revealed three alternative ways of writing or defining it:

(a) In terms of probabilities, as the second of the two major terms in Equation 27.16:

$$H_{Y|X} = -\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P(x_i)\ P(y_j|x_i) \log_2 P(y_j|x_i). \qquad (27.19)$$

(b) In terms of probabilities, by replacing $P(x_i)P(y_j|x_i)$ in Equation 27.19 with its equal, $P(x_i,y_j)$, per Equation 27.12:

$$H_{Y|X} = -\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P(x_i,y_j) \log_2 P(y_j|x_i). \qquad (27.20)$$

(c) In terms of entropies, by rearranging Equation 27.17:

$$H_{Y|X} = H_{X,Y} - H_X \qquad (27.21)$$

This last equation says that the conditional entropy of system Y, given a measurement of system X, is the joint entropy of the two systems minus the self-entropy of system X. All three of the equations can also be written for $H_{X|Y}$, the conditional entropy of system X, given a measurement of *y*.

Mutual Information

Two systems that have no mutual relation have no information about one another. Conversely, if they do have some connection, each contains information about the other. For instance, a measurement of *x* helps estimate *y* (or provides information about *y*, or reduces the uncertainty in *y*). Suppose now that we want to determine the amount (if any) by which a measurement of *x* reduces our uncertainty about a value of *y*. Two "uncertainties of *y*" contribute. First, *y* by itself has an uncertainty, as measured by the self-entropy H_Y. Secondly, there's an uncertainty of *y* given a measurement of *x*, as measured by the conditional entropy $H_{Y|X}$.

Conditional entropy $H_{Y|X}$ is a number that represents an amount of information about *y*. In particular, the basic uncertainty H_Y is lessened or partially relieved by an amount equal to $H_{Y|X}$. In symbols, that statement says that the overall decrease in uncertainty is $H_Y - H_{Y|X}$. The name for that difference or reduced uncertainty is *mutual information*, $I_{Y;X}$:

$$I_{Y;X} = H_Y - H_{Y|X}. \qquad (27.22)$$

Using a semicolon in the subscript Y;X characterizes our symbol for *mutual* information. For *joint* probabilities or entropies, we used a comma, as in joint entropy $H_{X,Y}$. For *conditional* probabilities or entropies, we used a vertical slash, as in $H_{Y|X}$

Alternate definitions

The technical literature and chaos theory also express mutual information in two other ways. As before, one is in terms of probabilities, the other in terms of entropies. This time we'll take the easy one (the entropy form) first.

Entropy Form

The entropy form stems from Equation 27.22, $I_{Y;X}=H_Y-H_{Y|X}$. The last quantity on the right, conditional entropy $H_{Y|X}$, is defined by Equation 27.21 as $H_{X,Y}-H_X$. Substituting that difference in place of $H_{Y|X}$ in Equation 27.22:

$$\begin{aligned} I_{Y;X} &= H_Y-[H_{X,Y}-H_X] \\ &= H_Y+H_X-H_{X,Y}\,. \end{aligned} \qquad (27.23)$$

In this form, mutual information is the sum of the two self-entropies minus the joint entropy.

We used the mutually associated case to derive Equation 27.23. However, Equation 27.23 also applies to the mutually unassociated case. For that case, mutual information turns out to be zero. To see that, we use the equation for the unassociated case, Equation 27.11 ($H_{X,Y}=H_X+H_Y$). Using that definition, we substitute H_X+H_Y in place of $H_{X,Y}$ in Equation 27.23. That gives $I_{Y;X}=H_Y+H_X-(H_X+H_Y)$, or $I_{Y;X}=0$. So, the mutual information of two independent systems is zero. In other words, one system tells us nothing about the other.

Incidentally, rearranging Equation 27.23 provides an alternate expression for joint entropy, $H_{X,Y}$:

$$H_{X,Y} = H_Y + H_X-I_{Y;X}. \qquad (27.24)$$

Joint entropy for two systems or dimensions, whether mutually related or not, therefore is the sum of the two self-entropies minus the mutual information. (For two unrelated systems, mutual information $I_{Y;X}$ is zero. Eq. 27.24 then reduces to Eq. 27.11, $H_{X,Y}=H_Y+H_X$.)

Probability Form

The probability form of mutual information relieves us of having to compute H_Y, H_X, and $H_{X,Y}$ (or $H_{Y|X}$ per Eq. 27.22) individually. It's a more economical form, making mutual information easier and faster to calculate. The procedure in deriving it is easy. We'll do it with Equation 27.23, which is $I_{Y;X}=H_Y+H_X-H_{X,Y}$. First, we rewrite H_Y and H_X in terms of probabilities. We do so in such a way that, insofar as possible, those probabilities consist of the same kinds of probability-terms as $H_{X,Y}$. Then we simplify the resulting expression of $H_Y+H_X-H_{X,Y}$ (Eq. 27.23) by factoring out some common terms. (And so it's our trusty three-step procedure for developing symbol definitions. We've got the concept that mutual information is the sum of the two self-entropies minus the joint entropy. Next we express that in probability symbols. Finally, we simplify.)

Let's begin with H_Y (Eq. 27.2):

$$H_Y = -\sum_{j=1}^{N_y} P(y_j)\log_2 P(y_j). \qquad (27.2, \text{repeated})$$

The crux of that equation is $P(y_j)$. Equation 6.8 defines $P(y_j)$ as

$$\sum_{i=1}^{N_x} P(x_i,y_j).$$

Substituting that into the *first* $P(y_j)$ term in Equation 27.2:

$$H_Y = -\sum_{i=1}^{N_x}\sum_{j=1}^{N_y} P(x_i,y_j)\log_2 P(y_j). \qquad (27.25)$$

That gives an expression for H_Y that's longer and more cumbersome than Equation 27.2, but there's a reason for such madness. The reason is that H_Y now begins with

$$-\sum_{i=1}^{N_x}\sum_{j=1}^{N_y} P(x_i, y_j),$$

just like $H_{X,Y}$ (Eq. 27.3) does.

Next we do the same for H_X. That is, we take Equation 6.7, namely

$$P(x_i) = \sum_{j=1}^{N_y} P(x_i, y_j)$$

and substitute the right-hand side for the *first* $P(x_i)$ term in Equation 27.1:

$$_X = -\sum_{i=1}^{N_x} P(x_i) \log_2 P(x_i) \qquad \text{(27.1, repeated)}$$

$$= -\sum_{i=1}^{N_x}\sum_{j=1}^{N_y} P(x_i, y_j) \log_2 P(x_i) \quad . \qquad \text{(27.26)}$$

Rewriting H_Y and H_X in that fashion puts all three of H_Y, H_X, and $H_{X,Y}$ into the common form of

$$-\sum_{i=1}^{N_x}\sum_{j=1}^{N_y} P(x_i, y_j) \log_2 P(\mathrm{A}),$$

where A represents certain bins and varies for each of the three expressions (Eqs 27.25, 27.26, and 27.3, respectively). Now we substitute Equations 27.25 (H_Y), 27.26 (H_X), and 27.3 ($H_{X,Y}$) into Equation 27.23:

$I_{Y;X} = H_Y + H_X - H_{X,Y}$ (27.23, repeated)

= Equation 27.25+Equation 27.26 - Equation 27.3

$$= -\sum_{i=1}^{N_x}\sum_{j=1}^{N_y} P(x_i, y_j) \log_2 P(y_j)$$
$$+\left[-\sum_{i=1}^{N_x}\sum_{j=1}^{N_y} P(x_i, y_j) \log_2 P(x_i)\right]$$
$$-\left[-\sum_{i=1}^{N_x}\sum_{j=1}^{N_y} P(x_i, y_j) \log_2 P(x_i, y_j)\right].$$

That completes step two—expressing our concept in terms of symbols. The final step is to simplify. In the lengthy expression just created, the double summation

$$\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}$$

and the multiplication by $P(x_i,y_j)$ are operations that apply to all three entropies. Hence, we can factor out those common ingredients and place them in front. It's like writing -ab -ac+ad as a(-b -c+d). That shortens our equation to

$$I_{Y:X} = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P(x_i, y_j) \left[-\log_2 P(x_i) - \log_2 P(y_j) + \log_2 P(x_i, y_j) \right]$$

in which, for convenience, I have switched the order of the first two log terms. (And so we've done a lot of simplifying or condensing already!)

We can do more. One of the rules of logarithms is that -loga=log(1/a). Applying that rule to the first two terms within the brackets,

$$I_{Y:X} = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P(x_i, y_j) \left[\log_2 \frac{1}{P(x_i)} + \log_2 \frac{1}{P(y_j)} + \log_2 P(x_i, y_j) \right].$$

We can now finish the entire escapade by using another log rule to combine the log terms within the brackets. That rule of logarithms is that loga+logb+logc=log(abc). The brackets include the summation of three such log terms. Multiplying those three log terms gives $\log_2[P(x_i,y_j)]/[P(x_i)P(y_j)]$. Uniting the log terms in that fashion brings the entire expression for mutual information to

$$I_{Y:X} = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P(x_i, y_j) \log_2 \frac{P(x_i, y_j)}{P(x_i)\ P(y_j)} . \qquad (27.27)$$

That step completes our derivation. It puts mutual information into a condensed symbol form, using probabilities. The final result (Eq. 27.27) doesn't look like $H_Y+H_X-H_{X,Y}$ (Eq. 27.23), but it is. (That's what we started out with, several paragraphs ago.)

The log term in Equation 27.27, namely $\log_2[P(x_i,y_j)/P(x_i)P(y_j)]$, involves a ratio. The *i* and *j* subscripts tell us to consider each bin individually, in turn, in the joint probability distribution. The numerator of the ratio is the joint probability of a given bin, $P(x_i,y_j)$. The denominator is the product of the two marginal probabilities. That is, it's the product of $P(x_i)$ (the marginal probability of state x_i) times $P(y_j)$ (the marginal probability of state y_j). Also, Equation 27.27 says to multiply the log of that ratio by $P(x_i,y_j)$. We therefore end up using each joint probability $P(x_i,y_j)$ twice in the overall calculations pertaining to each bin—first as the numerator in the ratio and then again as the multiplier. Finally, once we have the product of $P(x_i,y_j)$ times $\log_2[P(x_i,y_j)/P(x_i)P(y_j)]$ for every observed bin in the joint probability distribution, Equation 27.27 says to add up all those products.

Mutual information is a global measure in that it uses probabilities that we have measured over the entire attractor. (We have at least measured those probabilities over our entire dataset, and we hope and assume those data represent the entire attractor.)

In lag space, variable x_i becomes x_t and y_j becomes x_{t+m}. The joint probabilities $P(x_i,y_j)$ of Equation 27.27 then become sequence probabilities, $P(x_t, x_{t+m})$. That's the estimated probability that value x_{t+m} follows value x_t in the time series. Equa tion 27.27 requires an estimate of that probability for every observed phase space route. Also, $I_{Y:X}$ as computed with Equation 27.27 has the by-now-familiar basic "*P*log*P*" form. As such, it's a weighted average value—a weighted average mutual information. (Abarbanel et al. 1993 and some other authors in fact call it "average mutual information." For brevity, however, I'll just call it mutual information.)

Figure 19.9 dealt with estimating optimum lag in attractor reconstruction by the method of the first minimum in mutual information. Either Equation 27.27 or 27.23 is the equation to use for that purpose.

Interpretations

Figure 27.1 helps clarify the general concepts of self-entropy, conditional entropy, joint entropy, and mutual information for two systems. Circles in the figure represent self-entropies H_X and H_Y. The first case (part (a) of the figure) is that in which x and y aren't related. The joint entropy is the sum of areas H_X and H_Y (per Eq. 27.11). Neither of those entropies contains information about the other. Mutual information therefore is zero (the smallest possible value).

As an aside, let's take a short paragraph to see how a mutual information of zero can occur in terms of probabilities. Equation 27.27 says that zero mutual information can only come about when $\log_2[P(x_i,y_j)/P(x_i)P(y_j)]=0$ for every bin in the joint probability distribution. That log term is 0 for any one bin when the numerator of the log term equals the denominator, that is, when $P(x_i,y_j)=P(x_i)P(y_j)$ for that bin. The reason is that when the two are equal, their ratio is 1, and the log of 1 is 0. In other words, if the joint probability $P(x_i,y_j)$ equals the product of the marginal probabilities $P(x_i)P(y_j)$ for every bin, the two systems are statistically independent. Mutual information then is zero.

Now suppose that the two systems or variables have some sort of loose association. That means that measuring one variable gives some information about (perhaps even a rough approximation of) the other. Figure 27.1b shows the same two self-entropies as Figure 27.1a. The amount of overlap of the two self-entropies indicates the degree of affiliation between systems. That overlap zone represents the mutual information, $I_{Y;X}$. The finite value of mutual information (or existence of an overlap zone) means that measuring just one variable (i.e. without measuring the other) reduces our uncertainty in the other. For instance, whereas our uncertainty in y was a maximum (H_Y) in the unrelated case (Fig. 27.1a), it decreases to the conditional entropy $H_{Y|X}$ (the uncertainty in y after knowing x) for the mutually related case. The magnitude of that lesser uncertainty ($H_{Y|X}$) is the area of the hachured overlap zone in Figure 27.1b. The amount of reduction in uncertainty is the size of the overlap zone. In symbols, that's H_Y-$H_{Y|X}$, and it's the mutual information (Eq. 27.22). The same ideas also apply to x.

Joint entropy in Figure 27.b is the area within the outer perimeter of the combined systems. Any of Equations 27.17, 27.18, or 27.24 gives its magnitude.

The largest possible value of mutual information (not shown diagrammatically in Fig. 27.1) occurs when x completely determines y and vice versa. For that case, as a few trials on a joint distribution grid show, two results are certain. First, the sums of the marginal probabilities of the two systems are the same, so the self-entropies of the two systems are the same (Eqs 27.1 or 27.2). Secondly, the sum of the joint probabilities

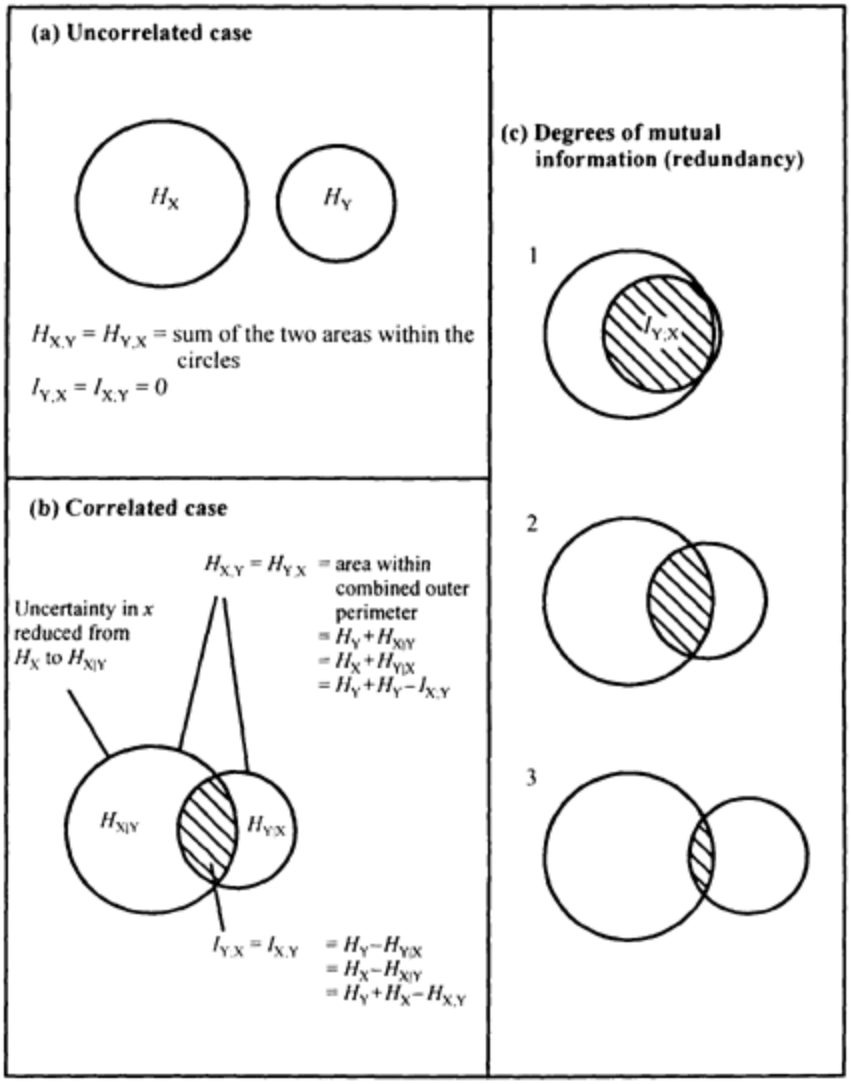


Figure 27.1
Sketches showing relations between self-entropy, joint entropy, conditional entropy, and mutual information (redundancy), for two systems, X and Y. (a) Uncorrelated case. (b) Correlated case (in which knowing a value y reduces the uncertainty in x, and vice versa). Modified and expanded from Cover & Thomas (1991: Fig. 2.2). (c) Differing amounts of mutual information or redundancy. Part 1 has a high value of mutual information (high correlation between x and y). Part 3 has a low value of mutual information (weak correlation between x and y).

$$\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P(x_i y_j),$$

equals the sum of the marginal probabilities for either system, for example

$$\sum_{i=1}^{N_x} P(x_i).$$

Hence, the joint entropy per Equation 27.3 equals the self-entropy of either system. For instance, $H_{X,Y}=H_X$. When that occurs, we can substitute H_X in place of $H_{X,Y}$ in the general relation $I_{Y;X}=H_Y+H_X-H_{X,Y}$ (Eq. 27.23). That leaves just $I_{X;Y}=H_Y$. (Also, $H_{X,Y}=H_Y$, so $I_{Y;X}=H_X$.) So when one system uniquely determines the other, mutual information is just the self-entropy of one system.

Overall, then, mutual information ranges from a minimum of zero (no relation between the two systems) to a maximum of the self-entropy of either system (which occurs at the limiting condition of one system uniquely determining the other).

With the above background, let's look at three ways to interpret mutual information:

- One interpretation is in terms of uncertainty. Mutual information is the amount by which a measurement of one variable reduces the uncertainty in another (Fig. 27.1 b). In other words, it's the reduction in the uncertainty of one random variable as a result of knowing the other (Cover & Thomas 1991: ch. 2).

- A second interpretation is in terms of information. Three ways to restate the general idea are as follows. Mutual information is:

  - A quantitative measure (in bits) of the average amount of common information between two variables. The overlap zones in Figure 27.1c, discussed below, demonstrate this concept. This interpretation shows that we can write our symbol for mutual information either as $I_{Y;X}$ or $I_{X;Y}$, as in the legends of Figures 27.1 a,b.

  - A measure of the average amount of information contained in one system, process, or variable about another. That is, it's the average amount of information *x* gives about *y*. If mutual information is relatively minor, *x* doesn't contain much information about *y*. In fact, a mutual information of zero means we can't find out a thing about *y* by measuring *x*. In other words, *x* and *y* are independent (Fig. 27.1a). Conversely, where mutual information is large, *x* contains much information about *y*.

  - The information contained in a process or variable (e.g. H_Y) minus the information contained in that variable when we've measured another variable (e.g. $H_{Y|X}$) (per Eq. 27.22).

- From a third point of view, mutual information measures the degree of association between two processes or random variables. For instance, a high value for mutual information (large reduction in uncertainty) implies that *x* and *y* have a close mutual affiliation (Fig. 27.1c, part 1). Conversely, a low value for mutual information indicates that *x* and *y* have only a weak affiliation (Fig. 27.1c, part 3). In fact, a value of zero for mutual information means that *x* and *y* are independent of one another (Fig. 27.1a). (The notion of the degree of association between two variables again shows why we can symbolize mutual information alternatively as $I_{Y;X}$ or $I_{X;Y}$.)

All of the above interpretations can be applied to lag space. In the context of lag space, mutual information is:

- The amount (in bits) by which a measurement of x_t reduces the uncertainty of x_{t+m}

- The amount of information that one measurement gives about a later measurement of the same variable (Fraser & Swinney 1986). Here Equation 27.27 gives the average mutual information between observations at time *t* and those at time *t+m*. In other words, it's the average amount of information about x_{t+m} that we get by making observation x_t.

- A measure of the degree to which knowing *x* at time *t* specifies *x* at time *t+m*. In other words, it measures predictability—how well we can predict *x* at time *t+m*, given that we measured *x* at time *t*.

Redundancy

Within chaos theory, **redundancy** (explained below) is nothing more than the extension of mutual information to three or more dimensions. Even then, some chaologists don't use the term "redundancy" that way and prefer, instead, to use "mutual information" for any number of dimensions. As usual, terminology is still developing. In any case, we'll develop our tools in the rest of this chapter.

First, here's how the idea of "redundancy" applies. Suppose *x* and *y* have some kind of direct (straight-line) relation and we know the equation of the line. Measuring one variable then tells us the other exactly. In that case, measuring the second variable as well as the first is a waste of time. It merely duplicates information we already got by measuring the first variable. Measuring the second variable in that case is unnecessary, superfluous, or *redundant*. Thus, when measurements of one system provide complete information about the other (i.e. reduce the uncertainty in the other to zero), then mutual information is a maximum. At the same time, all measurements of the second system are wholly redundant. That is, both mutual information and *redundancy* are a maximum.

Now let's go to the other extreme, namely where there's no relation at all between the two variables. Measuring *x* for unrelated systems (Fig. 27.1a) provides absolutely no information at all about *y*. Mutual information then is zero, and there's nothing redundant (zero redundancy) in measuring the second variable as well as the first. So, as before, we can think of mutual information and redundancy as the same thing.

For intermediate conditions, *x* provides some but not all information about *y* (and vice versa). Mutual information then is at some intermediate value, and measurements of the second variable might be described as partly redundant (Fig. 27.1b).

To recapitulate, what we've done in the last three paragraphs is lay a general foundation between mutual information and the idea of redundancy. Two systems that have a perfect one-to-one relationship have maximum mutual information. All measurements of the second system in that case are completely redundant (have maximum redundancy). For partially affiliated systems, mutual information is less. Measurements of the second system then are partly redundant. Finally, with unrelated systems, mutual information is zero. Measurements of the second system therefore aren't the least bit redundant (redundancy also is zero).

Figure 27.1c summarizes those concepts graphically. In part 1 of the figure, two systems or variables are closely related. Measuring one variable enables us to estimate the other quite well, without even having to measure that other variable. Measuring the second variable in addition to measuring the first is largely redundant. The relatively large area taken up by the mutual information on the sketch (i.e. the large value of mutual information) reflects that high degree of redundancy. In part 2 of the figure, the interrelation between systems is less. Measurements of the second variable then are still redundant to some extent, but not as much as in part 1. The lesser area represented by the mutual information indicates that lesser redundancy. Finally, in part 3, measuring one variable gives only a vague idea of the other. In that case, there's very little redundancy associated with measuring the second variable in addition to the first. Mutual information also is small.

The point here is that, for the two-system case, mutual information measures (and equals) redundancy. Any value for mutual information also is a quantitative measure of redundancy. Redundancy thereby joins company with entropy, information, and mutual information as a numerical measure of an ensemble of probabilities.

Derivations of redundancy

Original Definitions

Original definitions of redundancy came from information theory, long before chaos came along. Those definitions aren't exactly the same as the one used in chaos theory. In other words, the idea of redundancy as mutual information isn't the only concept of "redundancy" in the literature. However, most schools of thought, within both information theory (e.g. Shannon & Weaver 1949: 56) and chaos theory (e.g. Gershenfeld & Weigend 1993: 48), define redundancy in ways that are closely related to the "chaotic" ones. For background, let's take a quick look at how information theory usually defines redundancy. There are two ways. Each can be derived from the other.

The first definition is:

> *redundancy = the difference between some reference entropy and an actual or observed entropy.* (27.28)

Depending on the circumstances or problem, we can choose different definitions for "actual entropy" and "reference entropy." For instance, for one system or variable, redundancy might be the difference between the maximum possible entropy (the reference entropy) and the actual or self-entropy, say H_X. Maximum possible entropy occurs when the probability of every bin is the same; in that case it equals $\log N_s$, where N_s is the number of possible states or bins. Redundancy by the above definition then would be the difference between the reference entropy (here $\log N_s$) and self-entropy (here H_X), or $(\log N_s)$-H_X.

The second definition is a standardized version of the first definition. The standardization is done by dividing the basic difference by the reference entropy. That puts all redundancies within the range of 0 to 1. In the example just used, standardized redundancy becomes definition one divided by the reference entropy, or $(\log N_s - H_X)/\log N_s$. That ratio, according to the rules of basic algebra, is $(\log N_s/\log N_s)$-$(H_X/\log N_s)$, or 1-$(H_X/\log N_s)$. Thus, the second definition of redundancy as used in information theory is:

> *redundancy* = 1 -(*actual entropy*/*reference entropy*).

The ratio of the actual entropy to the reference entropy is a type of **relative entropy** (the ratio of one entropy to another).

Chaos Theory Definition in Terms of Entropies

Now we advance to chaos theory definitions of redundancy. In that context, as mentioned above, "mutual information" and "redundancy" are essentially the same thing. However, following Fraser (1986), I'll use "mutual information" only for *two* (and only two) variables, systems, or dimensions. "Redundancy," on the other hand, will refer to *three or more* variables, systems, or dimensions. That is, redundancy will be the generalization of mutual information from two dimensions to three or more dimensions.

Further, we'll usually apply "redundancy" to lag space. Let's now generalize our equation for mutual information (Eq. 27.23) so that it expresses redundancy (that is, it applies to three or more dimensions) and applies to lagged values of one feature (x). In this section we'll do it in terms of entropies, in the next in terms of probabilities.

Mutual information is the starting point. By definition, it involves two dimensions. It's just the sum of the two self-entropies minus the joint entropy (Eq. 27.23):

$$I_{X;Y} = H_X + H_Y - H_{X,Y}. \qquad \text{(27.23, repeated)}$$

For lag space, system *x* becomes X_t and system Y becomes X_{t+m}. To avoid the inconvenience of having a "sub-subscript," we'll write $H(X_t)$ for H_{Xt} (the self-entropy of system X_t) and make similar changes for the other symbols. With those modifications, Equation 27.23 as applied to lag space becomes

$$I(X_t;X_{t+m}) = H(X_t) + H(X_{t+m})-H(X_t,X_{t+m}).$$

In getting our new definition (redundancy), we follow our standard three-step procedure for defining a new quantity. The first step is to conceptualize. Here the concept is that redundancy (the multidimensional version of mutual information) is the sum of three or more self-entropies minus the joint entropy.

Step two is to express that idea in symbols. The modified version of Equation 27.23 that we just wrote is for just two dimensions (X_t, X_{t+m}). Now we want to generalize that relation to *D* dimensions, where *D* is three or more. Embedding dimension *D* represents *D* axes in lag space or *D* lagged measurements of *x*. The last member of the group of entropies then corresponds to $H(X_{t+(D-1)m})$, as explained in the vector chapter. For instance, the symbol for mutual information, $I(X_t;X_{t+m})$, (spoken as "*I* of X sub *t* and X sub *t* plus *m*") when generalized to redundancy *R* in that way, becomes $R(X_t,X_{t+m}, \ldots ,X_{t+(D-1)m})$. Similarly, the entire symbol definition for redundancy, based on our modified version of Equation 27.23, becomes:

$$R(X_t,X_{t+m}, \ldots ,X_{t+(D-1)m}) = H(X_t)+H(X_{t+m})+\ldots+H(X_{t+(D-1)m}) - H(X_t,X_{t+m}, \ldots , X_{t+(D-1)m}).$$

So, the sum of the self-entropies is symbolized by $H(X_t)+H(X_{t+m})+\ldots+H(X_{t+(D-1)m})$. Similarly, the joint entropy that we're subtracting from that sum is $H(X_t,X_{t+m}, \ldots X_{t+(D-1)m})$

Finally, step three of our standard procedure is to simplify and condense. We can do that by letting

$$\sum_{i=1}^{N} H(x_i)$$

represent the sum of the self-entropies. That is,

$$\sum_{i=1}^{N} H(x_i) = H(X_t)+H(X_{t+m})+\ldots+H(X_{t+(D-1)m}).$$

In that global summation symbol, *N* is the total number of self-entropies, $H(X_i)$ is the group of self-entropies, and $i=t, t+m, \ldots , t+(D-1)m$. That step condenses our symbol definition for redundancy in terms of entropies to:

$$R(X_t, X_{t+m}, \ldots, X_{t+(J-1)m}) = \left[\sum_{i=1}^{N} H(x_i) \right] - H(X_t, X_{t+m}, \ldots, X_{t+(J-1)m}). \quad (27.29)$$

Chaos Theory Definition in Terms of Probabilities

Another common chaos theory definition of redundancy is in terms of probabilities. Here we'll again follow our three-step procedure but only as a first stage, namely to get a definition applicable to two systems or variables. Then we'll generalize it to lag space and the multidimensional case.

For step one (the concept), we invoke the above idea that redundancy is a reference entropy minus an actual entropy (Eq. 27.28). Step two is to express that in probability symbols. For that purpose, we use a reference entropy based on marginal probabilities (Ch. 6):

$$-\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P(x_i, y_j) \log_2 [P(x_i)P(y_j)].$$

The actual or observed entropy is the joint entropy,

$$-\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P(x_i, y_j) \log_2 P(x_i, y_j). \quad \text{(Eq. 27.3, repeated)}$$

Redundancy *R* then is

$$R = \textit{reference entropy} \text{ minus } \textit{actual entropy}$$

$$= -\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P(x_i, y_j) \log_2 [P(x_i)P(y_j)] - \left[-\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P(x_i, y_j) \log_2 P(x_i, y_j) \right].$$

The third and final step is to simplify. The actual entropy (the expression on the far right) in that equation is a negative quantity that's subtracted from something. That makes it positive, so we rewrite the preceding equation by putting the actual entropy first:

$$R = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P(x_i, y_j) \log_2 P(x_i, y_j) - \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P(x_i, y_j) \log_2 [P(x_i)P(y_j)].$$

Common features of the two entropies on the right-hand side of the equation are

$$\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P(x_i, y_j).$$

In other words, the expression is comparable to ab-ac. That's a(b-c). Thus:

$$R = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P(x_i, y_j) \{ \log_2 P(x_i, y_j) - \log_2 [P(x_i)P(y_j)] \}.$$

Since loga-logb=log(a/b),

$$R = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} P(x_i, y_j) \log_2 \frac{P(x_i, y_j)}{P(x_i) P(y_j)}. \quad \text{(27.27, repeated)}$$

That expression for "redundancy" is a bit of a misnomer because it applies only to two systems, as indicated by the *x*'s and *y*'s. In fact, I've labeled it "Equation 27.27" because it's the same as Equation 27.27 for mutual information. Incidentally, the log term is another relative entropy, in this case the ratio of the entropy of the joint probability distribution to that of the "product" probability distribution. The products are those of the marginal probabilities (cf. Cover & Thomas 1991: 18, eq. 2.28).

Let's now generalize the simplified equation we just got to *lag space* and three or more dimensions. Lag space means we're using time-sequential values of one variable. That in turn means we need sequential probabilities. For lag space, we reinterpret the symbols in the usual way: system X becomes X_t, system Y becomes X_{t+m}, and the final system in our group is $X_{t+(D-1)m}$; for actual observations, x_i becomes x_t and y_j becomes x_{t+m}; we indicate data from additional "systems" or "variables" by additional lags, up to $X_{t+(D-1)m}$; and our former summation over all possible values of *i* and *j* translates to summing over all represented phase space routes, N_r. Incorporating those symbols lets us generalize Equation 27.27 to *D* dimensions as follows:

$$R(X_t, X_{t+m}, \ldots, X_{t+(D-1)m}) \qquad (27.30)$$

$$\sum_{i=1}^{N_r} P(x_t, x_{t+m}, \ldots, x_{t+(D-1)m}) \log_2 \frac{P(x_t, x_{t+m}, \ldots, x_{t+(D-1)m})}{P(x_t)\ P(x_{t+m}) \ldots P(x_{t+(D-1)m})} .$$

That equation is our goal—redundancy for lag space, in terms of probabilities.

The joint probability $P(x_t, x_{t+m}, \ldots, x_{t+(D-1)m})$ that occurs twice in Equation 27.30 is deceptive. It's a global symbol that can encompass a *very large* number of probabilities. Each variable or dimension (x_t, x_{t+m}, etc.) represents an additional "event" or measurement in the time sequence. Number of events in any specified route or sequence greatly increases the number of sequence probabilities we need to estimate, according to Equation 6.6: $N_r=N_{sn}$, where N_r is number of phase space routes, N_s is number of possible states, and *n* is number of events. We need a sequence probability for every route, or at least for those routes that occur at least once. Therefore, redundancy for more than two or three events or embedding dimensions usually requires estimating many thousands of joint (sequence) probabilities. That, in turn, means the dataset (time series) must be of colossal length.

Incremental redundancy and its interpretations

The definitions given above form the basis for the interpretations of redundancy that I'll mention here. Those definitions generally stem from or follow Fraser's (1989a) ideas. However, some concepts and definitions aren't yet standardized because the field is relatively young. For example, Gershenfeld & Weigend (1993: 48) define and interpret redundancy differently from Fraser.

Redundancy as defined in Equation 27.30 quantifies the average amount of common information among several systems or variables (Paluš 1993). In the context of Figure 27.1, for instance, we'd be looking at the overlap zone of not just two spheres but three or more spheres. If the variables are lagged measurements of one time series (e.g. x_t, x_{t+m}, x_{t+2}), redundancy is the number of bits that are redundant in a vector (Fraser & Swinney 1986), that is, redundant in the vector's components as a group.

Chaos theory sometimes focuses on the amount of change in the redundancy of a vector when we increase the vector's dimension by one (e.g. from two dimensions to three), at a constant lag. That quantity is just the difference between the two successive redundancies. That's the redundancy at the higher dimension minus that at the lower. (And so it's really just another version of first-differencing, discussed in earlier chapters.) Some authors symbolize it as R_{D+1}-R_D, others as R_D-R_{D-1}. Fraser (1989a) coined the term **marginal redundancy** for each such difference. I'm going to call it **incremental redundancy**. (Like "fractal dimension," terminology for mutual information and redundancy can be inconsistent from one author to another.)

One way to interpret incremental redundancy when dealing with lag space is in terms of information. Incremental redundancy quantifies the average amount of information that several successive measurements of *x* give about the next *x* (Paluš 1993: 390). It's the average amount of information that *n* sequential observations contain about observation *n*+1. (In these interpretations, each measurement gets a phase space dimension, so there are *D* dimensions providing information about dimension or observation number *D*+1.)

Another lag space view of incremental redundancy is in terms of prediction. Here incremental redundancy is a quantitative measure of the average number of bits that several sequential measurements of *x* can predict about the next *x*. That is, it estimates the number of bits that the earlier components of a vector can forecast about the last component. Hence, it reflects predictability—how well the last member of a measurement sequence can be predicted from its predecessors. The minimum value of incremental redundancy (zero) means that the data are useless in regard to predictions; we might just as well throw a dart at the wall or guess any number within the range of possible values. In contrast, a high value means close association and reliable predictions.

Typical values

Because redundancy depends strictly on probabilities, anything that affects those probabilities affects the value of redundancy. Five such influential factors are the number of possible states (bin size), number of sequential events (embedding dimension), lag, distribution of probabilities, and size of dataset. The last two—distribution of probabilities and size of dataset—are fixed once we've measured the time series. Hence, in analyzing a given time series for redundancy, we can regulate only the first three—number of possible states (the original partitioning), embedding dimension, and lag. Varying each of those three factors yields a different set of probabilities and hence a new value for redundancy. Therefore, each value of redundancy that we calculate pertains only to a particular combination of bin size, lag, and embedding dimension.

Suppose we choose a particular bin width or partitioning. Then redundancy varies only with lag and embedding dimension. The usual procedure is to choose a lag and compute redundancy for each of several successive embedding dimensions. In keeping with the philosophy of mutual information and redundancy, we need at least two measurements (two embedding dimensions) to compute a meaningful redundancy. (Redundancy for an embedding dimension of one is zero.) The first meaningful embedding dimension to use is two. At lag one, that means x_t and x_{t+1}. So, we'd first go methodically through the entire dataset to estimate the probabilities. Then we'd compute redundancy (actually mutual information, since *D=2). After that, we compute redundancy at that same lag (one) for three embedding dimensions (*x_t, x_{t+1} and x_{t+2}). Then for lag one at four embedding dimensions, five embedding dimensions, and so on. Data limitations may prevent computations of redundancy for more than about five or ten embedding dimensions, but it's worth doing as many as possible.

Once we've computed redundancy for successive embedding dimensions at lag one, the next step is to go to the next lag (lag two) and again compute redundancy for various successive embedding dimensions. Then repeat for lags three, four, and so on. Typically, we advance lag by increments of one until lag reaches about 10 or 15, then by increments of 5 thereafter. Again, data limitations may prevent calculations at lags of more than 30 or 40.

To show what some typical values look like, Table 27.1 lists redundancies and incremental redundancies based on the *x* values of 9000 iterations of the Hénon equations (Eq. 13.11). Parameter values used in the equations were a=1.4 and b=0.3. Those values produce a chaotic attractor. The table includes computations for four embedding dimensions at lags 1-15, then continues by lags of 5 through lag 50. At a given lag, incremental redundancy is the amount by which the redundancy changes with each successively higher embedding dimension. That is, it's the redundancy at a particular embedding dimension minus that at the preceding embedding dimension. For instance, at lag one, incremental redundancy for an embedding dimension of 4 is the redundancy at *D*=4 (having a value of 21.04, listed in col. 5 of Table 27.1) minus the redundancy at *D*=3 (value 14.31, in col. 4). That difference, R_{D+1}-R_D, is 6.73 (col. 8). The calculations are very crude; for instance, negative incremental redundancies, as columns 7 and 8 show for some greater lags, aren't possible theoretically.

Table 27.1 Redundancies and incremental redundancies computed from 9000 iterations of the Hénon equations (lagged values of *x* only; *y* values not involved); redundancy calculations provided by Andrew Fraser.

1	2	3	4	5	6	7	8
		Redundancies			Incremental redundancies		
		Embedding dimension			Embedding dimension		
Lag	1	2	3	4	2	3	4
1	0.00	6.43	14.31	21.04	6.43	7.88	6.73
2	0.00	5.76	12.27	16.89	5.76	6.51	4.62
3	0.00	4.80	9.97	12.14	4.80	5.17	2.17
4	0.00	4.17	7.39	8.26	4.17	3.22	0.87
5	0.00	3.03	5.35	5.42	3.03	2.32	0.07
6	0.00	2.38	3.76	3.60	2.38	1.38	-0.16
7	0.00	1.83	2.61	2.32	1.83	0.78	-0.29
8	0.00	1.43	1.93	1.67	1.43	0.50	-0.26
9	0.00	1.03	1.24	1.01	1.03	0.21	-0.23
10	0.00	0.65	0.72	0.62	0.65	0.07	-0.10
11	0.00	0.52	0.59	0.65	0.52	0.07	0.06
12	0.00	0.44	0.47	0.44	0.44	0.03	-0.03
13	0.00	0.28	0.31	0.40	0.28	0.03	0.09
14	0.00	0.16	0.25	0.37	0.16	0.09	0.12
15	0.00	0.09	0.11	0.16	0.09	0.02	0.05
20	0.00	0.00	0.04	0.02	0.00	0.04	-0.02
25	0.00	0.00	0.02	0.07	0.00	0.02	0.05
30	0.00	0.01	0.00	0.06	0.01	-0.01	0.06
35	0.00	0.01	0.03	0.09	0.01	0.02	0.06
40	0.00	0.00	0.00	0.06	0.00	0.00	0.06
45	0.00	0.00	0.03	0.11	0.00	0.03	0.08
50	0.00	0.00	0.01	0.03	0.00	0.01	0.02

Applications

Chaologists have suggested several applications of redundancy and incremental redundancy. Fraser's (1989a) work is the basis for the following idealized descriptions. The potential applications we'll discuss are:

- estimating optimal embedding dimension
- assessing the accuracy of the basic data
- estimating Kolmogorov-Sinai entropy
- distinguishing between periodic, chaotic and random data
- estimating optimal lag
- assessing predictability.

Each application carries certain qualifications that we needn't go into here. Also, all applications are exploratory and need further development to be easily and reliably usable (cf. Grassberger et al. 1991, Martinerie et al. 1992, Pineda & Sommerer 1993).

Embedding Dimensions

Two successive measurements of *random* data have no mutual relation. One value carries no information about the next. Their mutual-information values or redundancies are zero. The difference between two such redundancies (i.e. the incremental redundancy) therefore also is zero. This holds true regardless of embedding dimension (D) or lag. So, for any and all embedding dimensions with random data, any measurement of x is useless for predicting the latest x within a series (the latest component of a vector). (I'll sometimes speak of incremental redundancy as the difference between two successive redundancies, with symbol ΔR. At other times, I'll speak of it as the average number of bits that several sequential measurements of x can predict about the next x, that is, about the last component of a series of measurements.)

Data (chaotic or not) based on an underlying rule, on the other hand, have some association (apparent or not) between successive measurements of x. Such measurements therefore carry some mutual information or redundancy about each other. For these kinds of data, redundancy for any series of measurements (any embedding dimension) increases as we add another measurement (another embedding dimension), for a fixed lag. For example, three successive values (x_t, x_{t+m}, x_{t+2m}) are more redundant than two, four are more redundant than three, and so on. In other words, computed values of redundancy R increase with increase in embedding dimension, for such data (at fixed lag). Incremental redundancy in those cases is positive (>0), at least in theory.

One or two measurements of x (i.e. low embedding dimensions) may not give us a very good prediction about the last component of our string. In other words, incremental redundancies at small embedding dimensions tend to be small, for a fixed lag. Predictability of the last member of a series of measurements (the incremental redundancy) increases rather rapidly at first, as we incorporate more measurements or components into the vector, that is, as we increase the embedding dimension (Fig. 27.2). Further increases in the number of components bring a lesser and lesser gain in predictability of that last member. Eventually the incremental redundancy becomes approximately constant. A constant incremental redundancy means that further increases in embedding dimension don't improve the ability of the sequence of measurements to predict the last member of the string. Hence, there isn't any advantage to using an even larger embedding dimension. (The overall pattern is similar to that between the correlation dimension and embedding dimension, as we saw in Fig. 24.6.)

The best embedding dimension to use for attractor reconstruction with this approach is the one at which incremental redundancy no longer increases significantly, at that lag and measuring accuracy. One way to estimate that embedding dimension is from a graph of incremental redundancy as a function of embedding dimension, with lag as a third variable. (A simplified version of such a graph is Fig. 27.2, where we analyzed only one lag.)

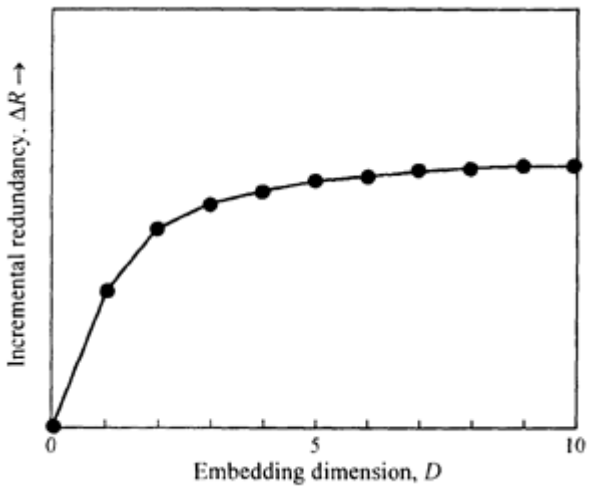


Figure 27.2
Hypothetical change of incremental redundancy
with embedding dimension, at a fixed lag.

A second type of graph for estimating an optimum embedding dimension is a plot of incremental redundancy as a function of lag, with embedding dimension as a third variable. Figure 27.3 shows hypothetical examples. Figure 27.4 (from Paluš 1993) shows a similar plot for real data—laboratory measurements of fluctuations in a far-infrared laser. On such graphs, relations for successively higher embedding dimensions become closer and closer to one another. When they become very close to one another, incremental redundancy at a given lag stays about the same (nearly constant) for still larger embedding dimensions (Fig. 27.2). In other words, there's little advantage in going to a larger embedding dimension. The optimum embedding dimension as estimated from graphs such as Figures 27.3 and 27.4 therefore is the smallest one for which the plotted relations get relatively close to one another. That's a subjective decision on our part.

Accuracy

Figures 27.3 and 27.4 show that the relations of incremental redundancy versus lag not only become closer as embedding dimension increases; they also become straighter. In fact, at small lags the lines converge toward some limiting perfectly straight "asymptotic accumulation line" as *D* goes to infinity. That asymptotic accumulation line, shown as a dashed line on Figure 27.3, is important.

Straight lines have the form y=c+bx, where y is the quantity plotted on the ordinate, x is on the abscissa, c is the intercept, and b is the slope. For our case, y is incremental redundancy ΔR and x is delay or lag, m. The straight line's equation therefore is ΔR=c+bm. Ordinarily, we would fit a straight line by some rigorous analysis of the data to find the parameters b and c. Our data unfortunately are noisy, less plentiful than we might prefer, and probably won't let us examine conditions of D nearing infinity. Therefore, we can't compute b and c directly. As a result, we have to fit the straight asymptotic accumulation line by eye. To do that, we first compute the relation for ΔR as a function of lag for larger and larger embedding dimensions D. Then, when those relations get very close to one another on the graph, we make a best guess as to where their limit at very large D (the straight asymptotic accumulation line) will plot, and we draw it in by eye. (For chaotic data, as in Fig. 27.3b, those straight sections only last over small lags. The asymptotic accumulation lines for chaotic data therefore apply only to those small lags.) Finally, we measure the values of slope b and intercept c directly from the graph. For instance, to get the intercept we extrapolate our straight line back to where it intersects the vertical axis (m=0) and read the value of ΔR.

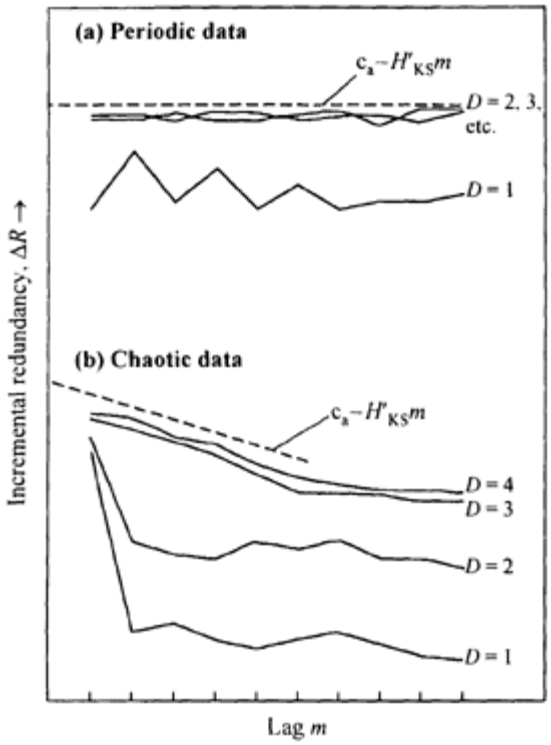


Figure 27.3
Idealized sketch of incremental redundancy as a function of lag.

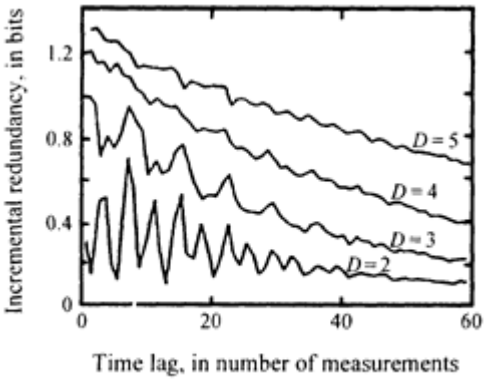


Figure 27.4
Incremental redundancy as a function of lag for measurements of fluctuations in far -infrared lasers (adapted from Palu 1993, with permission from Addison-Wesley-Longman Publishing Co.).

Let's look further at the intercept, c. In our present context, it's the value of incremental redundancy at a lag of zero, at an infinitely large embedding dimension. In other words, it reflects the number of bits of information that several measurements of x predict about the last of a group of such measurements, at that lag. Naturally, we wouldn't be interested in a lag of zero, but we might be interested in a lag very close to zero. Suppose we make several measurements very close together in time and average them. Then the only difference between that average and the next measurement (also taken almost immediately) will be attributable to noise. That noise represents the *accuracy* of that latest measurement. So, the incremental redundancy at a lag of zero (or very nearly zero), that is, the value of the intercept c, in a sense reflects the accuracy of our last measurement. (This concept of accuracy works in practice only for data that aren't very accurate.)

To help us keep that potential role of the intercept in mind, I'll relabel it c_a, in which subscript "a" stands for accuracy. The equation for the asymptotic accumulation line then becomes $\Delta R = c_a + bm$.

Kolmogorov-Sinai (K-S) Entropy

The standard definition of K-S entropy, written in terms of sequence probabilities (Eq. 26.6), doesn't work well with noisy data. Fraser (1989a) attempted to avoid that defect by developing an alternate definition based on incremental redundancies. The plot of incremental redundancy versus lag (e.g. Fig. 27.3) exemplifies his method. We'll label K-S entropy estimated in this way as H'_{KS}.

An esoteric mathematical theory of dynamical systems (not discussed here), when applied to a plot of incremental redundancies at successive lags, leads to defining the slope of the asymptotic accumulation line on the plot as the *negative* of K-S entropy, or $-H'_{KS}$. In other words, that quantity replaces slope b in the latest version of our straight-line equation. Thus, the equation of the asymptotic accumulation line (Fig. 27.3) is:

$$\Delta R = c_a - H'_{KS} m. \qquad (27.31)$$

Estimating K-S entropy with this approach is just a matter of measuring the slope of the straight asymptotic accumulation line and then reversing the sign. Slope as always is an ordinate distance divided by an abscissa distance. Here the ordinate distance Δy is $\Delta R_{m+1} - \Delta R_m$, and the abscissa distance Δx is $(m+1)-m$. where ΔR_{m+1} and ΔR_m are incremental redundancies at lags $m+1$ and m, respectively. Again, all of this discussion of the asymptotic accumulation line applies only in the limit where embedding dimension D approaches infinity. The symbols $D \to \infty$ refer to those conditions. Hence, the slope of the straight line is:

$$\text{slope} = \lim_{D \to \infty} (\Delta y / \Delta x)$$

or

$$\lim_{D \to \infty} \left[\frac{(\Delta R_{m+1} - \Delta R_m)}{(m+1) - m} \right].$$

Our estimate of K-S entropy (H'_{KS}) is the negative of that slope. So

$$H'_{KS} = \lim_{D\to\infty} -\frac{\Delta y}{\Delta x}$$

$$= \lim_{D\to\infty} -\left[\frac{\Delta R_{m+1} - \Delta R_m}{(m+1) - m}\right]. \qquad (27.32)$$

This method at present isn't particularly precise. However, it can be useful for identifying zero versus nonzero K-S entropy, as discussed in the next two paragraphs.

Nature of Time Series

Although K-S entropy as estimated from a plot of incremental redundancy versus lag isn't accurate, certain tendencies at high embedding dimensions on that plot can characterize or help distinguish between periodic, chaotic, and random data. Periodic data have a K-S entropy of zero, as calculated with Equation 27.32. When H'_{KS} is zero, Equation 27.31 ($\Delta R = c_a - H'_{KS} m$) reduces to just $\Delta R = c_a$ at all lags. In other words, the relation between incremental redundancy and lag (Fig. 27.3a) is a horizontal straight line at a positive and constant ordinate value of c_a. Usually, the overall trend with lag is only generally horizontal and can include periodic spikes (not shown here). An asymptotic accumulation line for such data also is roughly horizontal.

For chaotic data, K-S entropy H'_{KS} is positive. Equation 27.31 then is $\Delta R = c_a - (+H'_{KS})m = c_a - H'_{KS} m$. That equation says that, on the plot of ΔR versus lag, we get a straight line sloping downward (Fig. 27.3b).

Finally, as discussed earlier, incremental redundancies for random data are virtually zero, regardless of lag.

Lag

I mentioned above and in Chapter 19 that mutual information (Eqs 27.23 or 27.27) might help to indicate an optimum lag. (Mutual information by our definition means two dimensions, such as x_t and x_{t+m}.) That approach involved plotting the relation for mutual information versus lag, at a chosen embedding dimension, and taking the "best" lag to be the one corresponding to the first minimum in the plotted relation (e.g. Fig. 19.9).

Redundancy, including incremental redundancy, also can possibly help us estimate an optimum lag, for a given dataset and embedding dimension, D. With this approach, the optimum lag is the one for which the string of D measurements provides the maximum amount of useful information. "Useful information" here is a special expression involving redundancy but with an assumed small-scale noise factor removed. The method is as follows. First, some definitions (Fraser 1989a):

- the quantity $H'_{KS} m$=*the information lost to small scales in the time between any two successive measurements*
- the quantity $(D\text{-}1)H'_{KS} m$=the information lost to small scales over the entire string
- the quantity $Dc_a - R$=the total amount of information known over the entire string.

Using those estimates, an expression for the maximum amount of useful information is just a matter of algebraic manipulation, as follows. We define "useful information" as the total amount known minus that lost to small scales, where both amounts are taken over the string of D values. That useful information is definition 3 above minus definition 2, or ($Dc_a - R$) minus (D-1) $H'_{KS} m$. It's a function of lag m. We'll symbolize that function as $f_1(m)$. Stating that function as equal to the useful information (i.e. neglecting any proportionality constants):

$$f_1(m) = Dc_a - R - (D\text{-}1)H'_{KS} m.$$

Moving Dc_a to the left-hand side:

$f_1(m) - Dc_a = -R - (D-1)H'_{KS}m.$

Dividing everything by D-1:

$$\frac{f_1(m) - Dc_a}{D-1} = \frac{-R}{D-1} - H'_{KS}m.$$

Now we'll call the left-hand side $f_2(m)$. It's our new version of "useful information." The optimum value of lag is the lag that, for the chosen embedding dimension, gives the largest value for $f_2(m)$, that is, the largest value for the quantity $[-R/(D-1)]-H'_{KS}m$. For any $D \geq 2$, the value of m that maximizes $f_1(m)$ also maximizes $f_2(m)$, since D and c_a are fixed. That's true even if we add a constant to $f_2(m)$. For reasons that will soon become clear, we'll therefore add the constant c_a to $f_2(m)$, thereby obtaining what we'll call $f_3(m)$. So

$$\begin{aligned} f_3 m &\equiv c_a + f_2(m) \\ &= c_a + \left[\frac{-R}{D-1} - H'_{KS}m\right] \\ &= c_a - H'_{KS}m - \frac{R}{D-1}. \end{aligned} \qquad (27.33)$$

Equation 27.33 is our final expression for "useful information." We take the optimum lag to be the one for which Equation 27.33 yields the largest value.

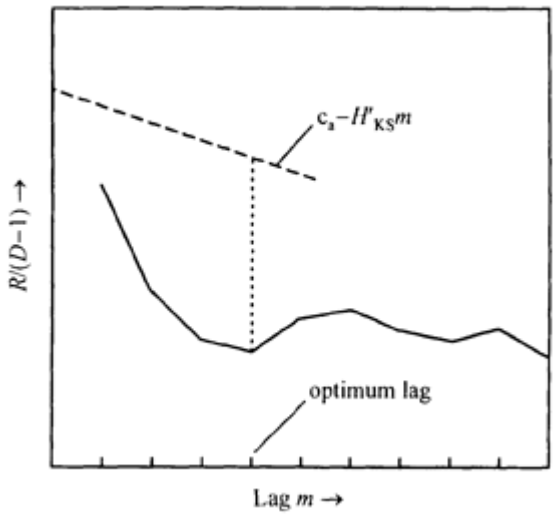


Figure 27.5
Sketch showing method of estimating
optimum lag, for one embedding dimension.

Writing "useful information" in the form of Equation 27.33 puts it in familiar quantities or relations. There are actually two such relations. Both are functions of lag m, as the left-hand side of Equation 27.33 indicates. One relation consists of lag as a function of the first two terms, $c_a - H'_{KS}m$. That's the straight asymptotic accumulation line (incremental redundancy at the limit of infinitely large D, per Eq. 27.31). The other relation is the remaining term, $R/(D-1)$, as a function of lag. As we'll see in a moment, $R/(D-1)$ is the average redundancy. Equation 27.33 says to subtract that average redundancy from the value of the accumulation line, at a given lag, in order to get a value for useful information.

Figure 27.5 is a hypothetical plot of $R/(D\text{-}1)$ against lag, for a given embedding dimension (the one we have chosen by whatever means). The same graph also includes the asymptotic accumulation line (incremental redundancy ΔR as D approaches infinity). Why can we plot that relation for ΔR on a graph on which the ordinate scale is $R/(D\text{-}1)$? Because in the limit of D approaching infinity the two are equal. That is, in the limit of D approaching infinity, $\Delta R = R/(D\text{-}1)$. Table 27.2 shows why that's true.

The first column in the table is embedding dimension D. The second column lists a series of incremental redundancies that asymptotically approach a limit (here 5.0) with increase in embedding dimension (in accordance with Fig. 27.2). Column 3 lists the associated total redundancies R. Finally, the last column is $R/(D\text{-}1)$.

Incremental redundancies, as in Table 27.2, are like the vertical distances from one floor to the next in a building. Summing those separate distances (those incremental redundancies) to any floor gives the elevation or total height of that floor (analogous to redundancy) above the street. In an alternate way, we'd get the distance from any one floor to the next (incremental redundancy) by subtracting their redundancies, that is, by subtracting the elevation of the upper floor from the elevation of the one below it.

Table 27.2 Hypothetical redundancies and associated values.

D	ΔR	R	$R/(D\text{-}1)$
1	1.0	-	-
2	3.0	1.0	1.0
3	4.0	4.0	2.0
4	4.5	8.0	2.67
5	4.75	12.5	3.125
6	4.875	17.25	3.45
7	4.9375	22.125	3.6875
8	4.96875	27.0625	3.866
9	4.984375	32.03125	4.004
10	-	37.015625	4.113

Besides the asymptotic increase of incremental redundancies, there are two other important features in the table. One is that the values of $R/(D\text{-}1)$ also approach a limit of 5.0 (although much more slowly than the incremental redundancies do). So, in the limit of D approaching infinity, both ΔR and $R/(D\text{-}1)$ will have the same value (here 5.0). In other words, in the limit of D approaching infinity, $\Delta R = R/(D\text{-}1)$. That's why we can plot the asymptotic accumulation line on the graph of $R/(D\text{-}1)$ versus m for the special condition of D nearing infinity. The second feature is that, to get an average value for any R, there are D-1 values that contribute. (There isn't any redundancy for a dimension of 1, that is, when we have only one measurement by itself.) So, the last column, $R/(D\text{-}1)$, lists average total redundancies.

As mentioned, the best lag is taken to be the one for which Equation 27.33 yields the largest value. That's the lag at which we get the largest amount of useful information, according to Equation 27.33. That value is biggest when $R/(D\text{-}1)$, subtracted from $c_a\text{-}H'_{KS}m$, gives the largest number. Graphically, that's the lag for which we have the greatest vertical distance between the asymptotic accumulation line (or $c_a\text{-}H'_{KS}m$) and $R/(D\text{-}1)$ (the other plotted relation on Fig. 27.5).

Predictability

Incremental redundancies quantify how well several sequential measurements of *x* predict the next *x*, at that lag and embedding dimension. For instance, say incremental redundancy at the optimum embedding dimension stays roughly constant with lag (as happens with deterministic, nonchaotic data, as in Fig. 27.3a). That means predictability about the next *x* remains constant at some positive value, regardless of the number of measurements included in the vector. In contrast, a downward-sloping relation (such as for chaos, as in Fig. 27.3b) means that our ability to predict the next *x* from *D* measurements decreases with lag.

A typical goal in regard to predictability is to find the smallest embedding dimension that provides optimum predictive power, for the given accuracy of the measurements and lag. That dimension is indicated by the proximity of the relations for successive embedding dimensions on the graph of incremental redundancy versus lag (e.g. Fig. 27.3), as mentioned earlier. When those relations get close to the asymptotic accumulation line, predictive ability has nearly reached the noise scale. Further increases in embedding dimension then are fruitless; no more predictive power can be gained.

The same types of graphs also indicate approximate limits of predictability in terms of lag. If the general trend of the ΔR-versus-m relation is roughly horizontal (e.g. Fig. 27.3a), then lag doesn't affect predictive power. If, instead, incremental redundancy decreases with lag (Fig. 27.3b at small lags) at a large embedding dimension, then we're losing predictive power as we increase lag. That decay of ability to predict typifies chaos, although it's not unique to chaotic systems (Ellner 1991). Predictability at that resolution becomes essentially zero when the relation falls to the vicinity of the abscissa (an incremental redundancy near zero).

In addition to the possible uses of mutual information and redundancy discussed above, other applications are being looked at as well. For instance, Paluš (1993) proposes a way to use redundancies to test for nonlinearity in a time series.

Summary

Mutual information extends the idea of entropy to two systems. Entropy thereby becomes a joint entropy—the average amount of uncertainty reduced (or information acquired) by measurements of two or more systems. Joint entropy for two mutually unrelated systems is the sum of their self-entropies. Joint entropy for two mutually related systems, in contrast, is the self-entropy of one system plus the conditional entropy of the other. Mutual information is the reduction in the uncertainty of a value of one system as a result of knowing a value of the other system. In terms of entropies, it's the sum of the two self-entropies minus the joint entropy. That's often expressed in one condensed equation, in terms of probabilities. Other interpretations of mutual information are that it's a measure of:

- the average amount of information contained in one system about another
- the degree of association between two processes or variables
- the amount of information that one measurement gives about a later measurement of the same variable.

Redundancy as used in chaos theory extends the idea of entropy to three or more systems. It's the mutual information of three or more systems or dimensions. As such, it's still the sum of the self-entropies minus the joint entropy. As with mutual information, it's often defined by a condensed equation in terms of probabilities. It quantifies the average amount of information common to several systems or variables. Incremental redundancy (sometimes called marginal redundancy) is the amount of increase in redundancy as a vector's embedding dimension is increased from *D* to *D*+1. Incremental redundancy also is the average amount of information that several successive measurements of *x* give about the next *x*. Alternatively, it's a quantitative measure of the average number of bits that several sequential measurements of *x* can predict about the next *x*.

EPILOGUE

Phase space portraits, correlation dimensions, Lyapunov exponents, K-S entropy, and other measures, are beautiful and potentially useful tools. However, they still need some development to be trustworthy in practice. For example, Caputo & Atten (1987:1311) comment that there's no reliable way to estimate Lyapunov exponents. Eubank & Farmer (1990: 171) say that "algorithms for computing dimension, Lyapunov exponents, and metric entropy are notoriously unreliable. They produce a number. But . . . it is very difficult to know a priori how much to trust the number . . . Many incorrect conclusions have been reached by naive application of these methods." Paul Rapp, lecturing at a 1992 us national conference on chaos, put it as follows: "These methods are not robust against misapplication. They fail in a particularly pernicious way. Rather than simply failing to produce a result, they can produce [plausible but totally] spurious results. Even when applied rigorously, care must be exercised when interpreting results."

The main reasons for those notorious practical difficulties are:

- Small and/or unrepresentative datasets.

- *Noise* The typical indicator of chaos was developed through numerical experiments with virtually unlimited amounts of noiseless data. In other words, the methods work best (and indeed depend) on clean, accurate data with a large number (say, thousands or millions) of observations. We can generate such voluminous and virtually noise-free datasets in computer experiments and sometimes in the laboratory. However, they're mighty scarce in the real world.

- *Number of important variables* The methods were developed with (and therefore work best on) low-dimensional systems. Identifying chaos in high-dimensional systems is much more difficult than in low-dimensional systems (Glass & Mackey 1988: 42).

- *Questionable fulfillment of basic assumptions* (e.g. deterministic systems; transients no longer relevant) Such assumptions may not hold for a real-world system (Glass & Kaplan 1993).

For all of these reasons, the very important practical aspect of identifying chaos in real data is still in relative infancy.

Because of the problems just mentioned and others, no single criterion by itself usually is enough or reliable for determining chaos (Wolf et al. 1985, Glass & Mackey 1988, Prichard & Price 1992, Provenzale et al. 1992). Instead, try several different tests or approaches. Most of these involve identifying and characterizing the chaotic attractor. Another advantage to doing several tests is that you learn more about the system, even if it turns out not to be chaotic or if you can't determine whether it's chaotic.

On another note, one important message I hope you've absorbed from this book is not to be intimidated by the mysterious, impressive-sounding names given to analytical tools in science and mathematics. Analyses or concepts with imposing names can be quite simple and straightforward. *D*-dimensional phase space, invariant probability distribution of attractor, autocorrelation function, K-S entropy, Lyapunov exponents, correlation dimension, mutual information, redundancy, and the many other associated ideas really aren't all that difficult once we pick them apart and see what they consist of.

We've come a long way, you and I. Take a second and think back on what you knew about chaos when you started this book . . . but only a second. Now that you're up to this rung of the ladder, it's on to the next. It's the climb that's fun. Hope to see you along the way.

APPENDIX
SELECTED LAWS OF POWERS, ROOTS AND LOGARITHMS

0^c =	0 (c being positive)
0^{-c}=	*∞ (c being negative), since $0^{-c}= 1/0^c=1/0=\infty$*
1^c =	*1 (regardless of whether c is positive or negative)*
x^0 =	1 (if *x* doesn't equal 0)
x^1 =	*x*
x^{-c}=	$1/x^c=(1/x)^c$
x^{-1} =	$1/x$
$(xy)^c$=	x^cy^c
$x^c(x^d)$ =	$x^{(c+d)}$
x^c/x^d=	$x^{(c-d)}$
$(x/y)^c$ =	x^c/y^c
$(x^c)^d$ =	x^{cd}
-(log*x*) =	$\log(x^{-1})=\log(1/x)$
log*x*=	*-[log(1/x)]*
$\log(x^c)$ =	clog*x*
log(*xy*) =	log*x*+log*y*
y =	cx^d transformed into logs: log*y*=logc+d (log*x*)
$\log_a x$ =	$\log_b x/\log_b a$

Hence, if a=2 and b=e: $\log_2 x=\log_e x/\log_e 2=\log_e x/0.69315=1.443 \log_e x$.

GLOSSARY

A

accuracy (a) A numerical *measure* of how close an approximation is to the truth; (b) correctness, in the sense of lack of bias.

adaptive histogramming Any histogramming procedure in which bin width is allowed to vary (usually according to local densities in the data).

adaptive kernel density estimator A *probability*-estimation technique consisting of the *kernel density estimator* but with the variation that bin width changes according to local density of data points.

affine Produced by a *transformation* in which the *variables* or *axes* haven't all been scaled by the same factor.

algorithm A recipe, plan, sequence of steps, *set* of instructions, list of rules, or set of mathematical equations used to solve a problem, usually with the aid of a computer. Some of the many possible forms of an algorithm include a loosely phrased group of text statements, a diagram called a flowchart, or a computer program.

almost-periodic orbit An *orbit* that comes closer and closer to repeating itself with time.

amplitude (a) The extreme range of a fluctuating quantity, such as a pendulum, tidal *cycle*, etc.; (b) the maximum difference between the value of a *periodic* oscillation and its *mean*. In "next-amplitude" plots, an amplitude is simply a *variable*'s local maximum or peak value of each oscillation, within a sequence of oscillations.

aperiodic (a) Not repeating with time, that is, lacking *periodicity* or *quasiperiodicity* (tantamount to having an infinite *period*); (b) occurring as a *transient* or pulse only once over infinite time.

arithmetic mean Same as *mean* (the sum of all data values divided by the number of such values).

attractor The *phase space* point or *set* of points representing the various possible *steady-state* conditions of a *system*; in other words, an equilibrium *state* or group of states to which a *dynamical system converges*.

attractor reconstruction See *reconstruction of attractor*.

autocorrelation Correlation of a *variable* at one time with itself at another time. Also called *serial correlation*.

autocorrelation coefficient A dimensionless numerical indicator (calculated as *autocovariance* divided by *variance*) of the extent to which the measurements of a *time series* are mutually related.

autocorrelation function The *spectrum* or entire series *of autocorrelation coefficients* for *a time series*.

autocorrelation time The time required for the *autocorrelation function* to drop to 1/e (= 1/2.718 = 0.37).

autocovariance A *measure* of the degree to which the observations of a *time series* are related to themselves, numerically calculated as:

$$\frac{1}{N}\sum_{t=1}^{N-m}(x_t-\bar{x})\,(x_{t+m}-\bar{x}),$$

autonomous (a) Self-governing, propagating from within; (b) independent of time.

axis (a) One of a group of mutually perpendicular reference lines passing through the *origin* of a graph; (b) a line about which a body or group of bodies can rotate.

axis vector *Principal axis*.

B

band-pass filter A *filter* that passes only those frequencies that fall within a desired range or band.

basic wave *Fundamental wave*.

basin of attraction The group of all possible *phase space* points that can evolve onto a given *attractor*.

basis A *set* of nonparallel and linearly independent *vectors*.

bifurcation (a) In general, a branching into parts or into connected segments (usually from one segment to two); (b) any abrupt change in the qualitative form of an *attractor* or in a *system*'s *steady-state* behavior, as one or more *parameters* are changed.

bifurcation-rate scaling law An equation that estimates the critical *parameter*-value at which a particular *bifurcation* occurs, within a series of bifurcations in *period-doubling*.

bifurcation diagram A graph showing all possible solutions (excluding *transients*) to an equation that relates a *variable* to a *control parameter*. Such a graph usually is drawn specifically to show *bifurcations* (*period-doublings*) of the variable.

bifurcation point A critical *parameter* value at which two or more branches of *system*-behavior emerge.

binary Having two equally likely, *mutually exclusive* possibilities.

binary digit Either of the two *digits*, conventionally 0 and 1, used in a *binary system* of reading numbers or of naming numbers; as such, it's the smallest amount of *information* that can be stored in a computer.

binary system A *system* that is set up or operates on *binary* principles.

bit (a) Contraction of "*binary digit*;" (b) either of the digits 0 or 1; (c) the unit of *information* when logs are taken to the base 2, that is, a *measure* of *information* equal to the decision content of one of two *mutually exclusive* and equally likely values or states (sometimes also called a shannon).

box-counting dimension *Similarity dimension* as obtained from a grid of boxes.

box dimension Same as *box-counting dimension*.

broadband spectrum A *frequency-domain* (*Fourier-analysis*) plot revealing no outstanding *periodicity*.

butterfly effect *Sensitive dependence on initial conditions*. The "butterfly" name stems from the theoretical possibility that a very slight change in the *state* of a *system* (such as a butterfly flapping its wings) might create slightly different "initial" conditions and thereby influence the long-term resulting pattern (the weather), even in some other part of the world.

byte (a) The amount of memory space needed to store one character on a computer (normally eight *bits*); (b) a string of *binary* elements operated on or treated as a unit.

C

Cantor set A *fractal* obtained by dividing a line into equal subparts, deleting one or more of those parts, and repeating that process indefinitely.

capacity A *dimension* D_c defined only at the limit where the scaling object approaches length zero; written in equation form as:
$D_c = \lim_{\varepsilon \to 0} (\log N / \log [1/\varepsilon])$.

cellular automaton A mathematical construction consisting of a *system* of entities, called cells, whose temporal evolution is governed by a *set* of rules and whose behavior over time becomes, or at least may appear, highly complex.

chain rule A rule that tells how to differentiate a *function* of a function, that is, how to find the *derivative* of a composite function. The name stems from the fact that the various functions fit together like a chain. For example, if *y* is a function of *t*, and *t* is a function of *x*, then the *chain rule* says that $(dy/dx)=(dy/dt)(dt/dx)$.

chaologist A person who studies *chaos*.

chaology The study of *chaos*.

chaos (a) Sustained and *random*-like long-term evolution that satisfies certain special mathematical criteria and that happens in *deterministic, nonlinear, dynamical systems:* (b) largely unpredictable long-term evolution occurring in a deterministic, nonlinear dynamical *system* because of *sensitivity to initial conditions.*

chaos theory The principles and operations underlying *chaos.*

chaotic attractor An *attractor* that shows *sensitivity to initial conditions* (*exponential divergence* of neighboring trajectories) and that, therefore, occurs only in the chaotic domain.

characteristic exponent *Lyapunov exponent*.

circle map Any of a class of *nonlinear difference equations* that *map* a point on a circle to another point on the circle (Middleton 1995).

classical decomposition The breaking down of a *time series* into its component constituents of *trend*, *seasonality*, *cyclicity*, and *noise*.

coarse graining Categorizing or *partitioning,* such as when reducing a *set* of observations of a *continuous* variable to a stream of *discrete* symbols.

coefficient (a) A constant number or symbol prefixed as a multiplier to a *variable* or unknown quantity (and hence a *constant* as opposed to a variable); (b) a dimensionless, numerical *measure* of a *set* of data, for example, an *autocorrelation coefficient*.

complex number Any number of the form a+*j*b, where a and b are real numbers and *j* is an *imaginary number* (the square root of -1).

complexity A type of *dynamical* behavior in which many independent agents continually interact in novel ways, spontaneously organizing and reorganizing themselves into larger and more complicated patterns over time.

conditional entropy The uncertainty in a measurement of *variable y* from *system* Y, given a value of variable *x* from coupled system X.

conditional probability The likelihood that a particular *state* or event will happen, given that one or more other specified states or events have taken place previously.

conservative Not losing energy (keeping instead a constant area or volume in *phase space*), with time.

constant A quantity that doesn't vary under specified conditions.

continuous Defined at all values of the given (independent) *variable*.

continuous random variable A *random variable* that can take on any value over a prescribed *continuous* range.

control parameter A controllable *constant* or quantity, different values of which can produce different cases or outcomes of an experiment or equation.

converge To approach a finite limiting value.

coordinate (noun) (a) One of a *set* of numbers that locate a point in space; (b) the *axis* of a graph, as in "coordinate axes" or "coordinate *system*."

coordinate vector A *vector* whose starting point is at the *origin* (0,0) of the graph. The word "coordinate" for such vectors is often dropped for convenience.

correlation dimension An exponent in a *power law* (i.e. slope of straight line) on a logarithmic plot of *correlation sum* (as the *dependent variable*) versus the radius ε of an encompassing circle, sphere, or *hypersphere*.

correlation exponent *Correlation dimension*.

correlation integral *Correlation sum*.

correlation sum A normalized total number of pairs of points within a circle or sphere of radius ε, obtained by dividing that total sum by the total number of points on the *attractor*.

correlogram A plot *of autocorrelation coefficient* (on the ordinate) versus *lag*, on arithmetic scales.

crisis An abrupt, discontinuous change in a *chaotic attractor* as a *parameter* is varied, characterized by either destruction of the chaotic attractor or its expansion to a much larger *interval* of *x*.

cubic polynomial A *polynomial* in which the highest *power* to which any *variable* is raised is 3.

cubic spline A smoothly varying, cubic-*polynomial* curve fitted between two data points, often used to interpolate "data" values between the two measured points.

cycle A series of events or observations that occur in a fixed sequence and return to the original *state* and which then repeat themselves in a regular pattern.

cyclicity Same as *periodicity*, that is, the tendency for any pattern to repeat itself over fixed *intervals* of time.

D

data-adaptive Said of methods of analysis in which the operations are applied to classes or subgroups that are defined differently from one to another, according to specified peculiarities of the data (e.g. number of observations in a class).

decomposition (a) The numerical expression of a quantity in terms of its simpler components; (b) see *classical decomposition*.

degrees of freedom The number of independent quantities (*variables* or *parameters*) or pieces of *information* that must be specified to define the *state* of a *system* completely. The field of statistics also has several other less common meanings.

delay-coordinate method Same as *time-delay method*.

delay method Same as *time-delay method*.

delay time (a) A delay, in time units, usually accompanied by an *integer parameter* to account for the units; (b) in some authors' usage, same as *lag* time.

delay vector Same as *lag vector*.

density Same as *probability density function*.

density estimation The estimation of probabilities.

density function Same as *probability density function*.

dependent variable The output *variable* of a *function*, as determined by the values of the *independent variables*.

derivative (a) Generally, a value that derives from, comes from, or is traceable to a particular source, such as a point on a curve; (b) in mathematics, the slope (rate of change) of the line that is tangent to a point on a curve.

deterministic (a) Completely and exactly specified (at least to within measuring *accuracy*) by one or more mathematical equations and a given initial condition; (b) said of a *system* whose past and future are completely determined by its present *state* and any new input.

deterministic fractal A *fractal* that looks exactly the same at all levels of detail.

detrend To transform trended data in such a way that they have an approximately constant *mean* and *variance* with time, so that the *trend* is removed.

diffeomorphic Smooth and *invertible*.

diffeomorphism A differentiable *mapping* that has a differentiable inverse.

difference equation A recurrence equation based on changes that occur at *discrete* times and solved by *iteration*. An example is the *logistic equation*. In mathematical jargon, a difference equation is an equation in which a difference operator is applied to a *dependent variable* one or more times. (A difference operator is a twofold mathematical operation in which (1) a small increment is added to the *independent variable*, and then (2) the original value of the independent variable is subtracted.)

difference plot A graph on which the *coordinates* of each plotted point are successive differences. Each successive difference is the difference between an observation x_t and some later observation (on the abscissa) and the difference between x_{t+m} and a similarly lagged observation (on the ordinate).

differencing The filtering of a *time series* by subtracting each value, in turn, from a subsequent (lagged) value. See also *first-differencing*, *second-differencing*, and *difference plot*.

differential equation An equation expressing a relationship between a *function* and one or more of its *derivatives* and that therefore is based on changes that occur continuously.

digit A code character, such as a number from 0 to 9 or a letter of the alphabet.

digital In numerical form, as in (a) calculation by numerical methods or by *discrete* units or (b) representation of data by numerical digits or discrete units.

digitize (a) To approximate by discrete samplings; (b) to put data into digital (numerical) notation (as for use in a digital computer).

dimension Generally, a magnitude measured in a particular direction, as on the axes of a graph. In *chaos theory*, "dimension" is used in any of several variations of that general meaning, such as (a) each *axis* of a *set* of mutually perpendicular axes in Euclidean space; (b) the number of *coordinates* needed to locate a point in space; (c) the maximum number of *variables* of a *system*; (d) any of various quantitative, topological measures of an object's *complexity*, and other variations.

discrete Defined or occurring only at specified values.

discrete Fourier transform (DFT) A mathematical operation that transforms a series of *discrete*, equally spaced observations measured over a finite range of time into a discrete, *frequency-domain spectrum*. Also called a finite *Fourier transform*.

discrete random variable A *random variable* that only takes on certain specified outcomes or values, with no possible outcomes or values in between.

discretize To extract equally spaced (in time) *discrete* values from a continuous *time series.*

dissipative system A *system* that loses energy with time. Evidence of energy loss with time includes an irreversible evolution toward an asymptotic or limiting condition over time and a decrease of *phase space* area or volume, with time.

distance formula The simple equation, based on the *Pythagorean theorem*, for finding the straight-line distance c between two points in N_d-dimensional space:

$$c = ([x_2-x_1]^2 + [y_2-y_1]^2 + [z_2-z_1]^2 + \ldots + [w_2-w_1]^2)^{0.5}$$

where w is the N_dth *variable* of the group of N_d variables.

distribution function A mathematical operation that gives the proportion of members in a sample or *population* having values less than or equal to any given value. See also *frequency distribution* and *probability distribution*.

dot product A *scalar* quantity for two *vectors* that have a common *origin*, obtained either by (a) multiplying their *x coordinates* (x_1x_2), their *y* coordinates (y_1y_2), etc. for any other *dimensions* involved, and then summing those coordinate products ($x_1x_2+y_1y_2$+etc.), or (b) multiplying the vectors' lengths by the cosine of their included angle.

dynamical Changing with time.

dynamical system (a) Anything that moves or that evolves in time; (b) any process or *model* in which each successive *state* is a *function* of the preceding state.

dynamical-systems theory (a) The study of phenomena that vary with time; (b) a language that describes the behavior of moving or evolving systems, especially as affected by external *control parameters*.

dynamics (a) That branch of physics (mechanics) that deals with forces and their relation to the motion and sometimes the equilibrium of bodies; (b) the pattern of change or growth of an object or phenomenon.

E

embedding The preparation of a *pseudo phase space* graph to reconstruct a *system*'s dynamics (*attractor*), using successively lagged values of a single *variable.*

embedding dimension The total number of separate *time series* (consisting of the original series and subgroups obtained by lagging that series) used in *a pseudo phase space* plot or in a more rigorous mathematical analysis.

emergent Said of phenomena or *systems* in which new, increasingly complex levels of order appear over time.

entropy (a) A *measure* of unavailable energy (thermodynamics), degree of disorder or disorganization, *probability* (in inverse proportion), uncertainty, randomness, variety of choice, surprise, or *information*; (b) a quantity,

$$\sum_{i=1}^{N} P_i \log (1/P_i)$$

computed for a *discrete random variable* whose *i*th outcome has *probability P*.

equation of motion An equation in which time is the *independent variable*.

equilibrium point Same as *fixed point*.

ergodic (a) The property whereby statistical measures of an ensemble don't change with time and, in addition, all statistics are *invariant* from one *time series* to another within the ensemble; (b) said of a *system* for which spatial or ensemble averages are equal to time averages (meaning that time averages are independent of starting time and that most points visit every region of *phase space* with about equal *probability*); (c) said of a *trajectory* if it comes back arbitrarily close to itself after some time; (d) the property whereby averages computed from a data sample *converge* over time to ensemble averages (i.e. statistics of all initial states ultimately lead to the same *set* of statistics). Eubank & Farmer (1990) mention additional usages of the word and *state* that "there is no universally accepted definition of the word 'ergodic'."

ergodic hypothesis (a) The *dynamical* theory that says that, in the limit as the number of observations goes to infinity, a time average equals a space average (i.e. the theory says that the *point* that represents the *state* of the *system* spends, in each *phase space* compartment, an amount of time proportional to the volume of that compartment); (b) the *dynamical* theory that says that, for a system in statistical equilibrium, all accessible states are equally probable, so that the system passes rapidly through all of them.

ergodic theory (a) The mathematical study of the long-term average behavior of *dynamical systems*; (b) the study of *measure-preserving transformations*; (c) *ergodic hypothesis*.

error The difference between a quantity and an estimate of the quantity.

Euclidean dimension The common or standard notion of "*dimension*" whereby a point has dimension zero, a line dimension one, a surface dimension two, and a volume dimension three.

exponential (a) In a general sense, relating to powers (exponents); (b) referring to a rate of change that's proportional to a *constant* raised to a *power*, where the power includes the *independent variable* (see *exponential divergence* and *exponential equation*); (c) referring to a specific mathematical series known as an "exponential series."

exponential divergence Temporal separation of two adjacent trajectories according to an exponential law, that is, by a straight-line relation between the log of separation distance (as the *dependent variable*) and time.

exponential equation (exponential function) An equation relating a *dependent variable* to some *constant* raised to a *power*, where the power includes the *independent* (given) *variable*. Examples are $y=c^x$ and $y=ac^{bx}$, where x and y are variables and a, b, and c are constants. An exponential equation plots as a straight line on semilog paper, with the log *scale* being used for y (or y/a) and the arithmetic scale for x.

F

false nearest neighbor A point in lagged space that is close to another point only because the *embedding dimension* is too low.

fast Fourier transform (FFT) Any member of a family of computer *algorithms* for calculating the various real and imaginary parts of the *discrete Fourier transform* (DFT) efficiently and quickly in about $N\log_2 N$ operations.

feedback That part of the output that returns to serve as input again, in a temporal process.

Feigenbaum constant Same as *Feigenbaum number*.

Feigenbaum number A universal *constant* (4.6692 . . .), discovered in the mid-1970s by Mitchell Feigenbaum, that represents the rate (in the limit where number of periods, *n*, becomes infinite) at which new periods appear during the *period-doubling* route to *chaos*. The Feigenbaum number (also known as the *Feigenbaum constant*) is defined as:

$$\lim_{n\to\infty} (k_n - k_{n-1}) / (k_{n+1} - k_n) = 4.6692\ldots$$

fidelity A *measure* of how closely linked one measurement is to its predecessor.

fiducial Referring to something used as a standard of reference for measurement or calculation. Examples: "fiducial point," *fiducial trajectory*.

fiducial trajectory A *trajectory* used as a reference trajectory from which to compute orbital gaps and the *Lyapunov exponent*.

filter (linear filter) Any mathematical technique or operator that systematically changes an input series, such as a *time series* or *power spectrum*, into a new (output) series that has certain desired qualities. Some purposes of *filters* are to (a) eliminate *periodicity* or *trend*; (b) reduce or eliminate *noise*; (c) suppress high or low frequencies; and (d) remove *autocorrelation*.

final state *Point attractor*.

first derivative The initial *derivative* of a *function*.

first-difference plot A graph on which the *coordinates* of each plotted point are first differenced data, in the form of the difference between x_{t+m} and x_t on the abscissa and the difference between x_{t+2m} and x_{t+m} on the ordinate, in which *m* is *lag*.

first-difference transformation A *transformation* performed by *first-differencing* the data.

first-differencing Calculation of the difference between a measured value (x_t) and a lagged measurement (x_{t+m}), for all values in a *time series*. This is the most common form of *differencing*.

first harmonic A wave that has the same *frequency* as the *fundamental wave*, in *Fourier analysis*.

first principal axis The *principal axis* that is stretched the most (or reduced the least) in the *phase space* evolution of an arbitrarily designated element.

first-return map A (pseudo) *phase space model* of *Poincaré-section* data, giving the value of some *variable* as related to its preceding value at that section.

fixed point (a) In *discrete* processes, a single *phase space* point that is its own *iterate* (a point for which $x_{t+1}=x_t$); (b) in *continuous* processes, a *constant*, time-independent solution to a *differential equation*. Also called *equilibrium point* or *steady state*. See also *stable fixed point* and *unstable fixed point*.

fixed-point attractor *Point attractor*.

flip bifurcation *Period-doubling*.

flow (a) A *set* of *differential equations*; (b) a *phase space trajectory* or bundle of trajectories obtained by solving a set of differential equations.

folding A topologist's explanation of how a particular range of output values results from (a) two different ranges of input values during *iteration*; or (b) continued *phase space* evolution of a chaotic *trajectory* as it reaches the limiting value of the *variable*(s) and then rebounds or deflects back onto the *attractor*.

forced oscillator A device to which extra energy is periodically added, by some external means.

fork-width scaling law An equation that estimates the width between the two parts of any particular *bifurcation* (fork), within a series of bifurcations in *period-doubling*.

Fourier analysis A mathematical technique for uniquely describing a *time series* in terms of the *frequency-domain* characteristics of its *periodic* constituents.

Fourier coefficient (a) A numerical value reflecting the strength of a particular constituent wave relative to that of other constituent waves; (b) one of the *coefficients* needed to express a *function* formally in terms of its *Fourier series*.

Fourier cosine series An equation for getting wave height for a particular *harmonic* by adding the cosines of angles associated with constituent *wavelengths*, in *Fourier analysis*.

Fourier integral A mathematical expression that extends the *Fourier series* to the more general situation of an infinitely long *period* by decomposing a continuous *time series* into *sinusoidal* constituents at all *frequencies* and merging the *variances* into a *continuous* distribution of variances.

Fourier series An equation that describes a *periodic time series* in terms of the cosines and/or sines of constituent *harmonics* and their associated *coefficients*.

Fourier sine series An equation for getting wave height for a particular *harmonic* by adding the sines of angles associated with constituent *wavelengths*, in *Fourier analysis*.

Fourier transform A mathematical *frequency-domain* characterization of a *time series*, consisting of constituent *amplitudes* and *phases* at each *frequency*.

fractal (a) A pattern that repeats the same design and detail or definition over a broad range of *scale*; (b) a *set* of points whose *dimension* is not a whole number (Lorenz 1993).

fractal dimension A generic term for any *dimension* (e.g. *similarity dimension*, *capacity*, *Hausdorff-Besicovich dimension*, *correlation dimension*, etc.) that can take on a non-*integer* value.

fractional dimension *Fractal dimension*.

frequency (a) In physics, the number of repeating *wavelengths*, *periods*, or *cycles* in a unit of time; (b) in statistics, the number of observations or individuals in a class; (c) also in statistics, a *probability* or proportion (*relative frequency*).

frequency analysis Same as *Fourier analysis*.

frequency distribution A list of class *intervals* or values and their associated number of observations (*frequencies*). The frequencies or number of observations in each class are often normalized to range from 0 to 1 and are often assumed equal to probabilities.

frequency domain The representation of *time series* data in terms of their *frequencies* (or of some other wave characteristic) and respective *variances*.

frequency locking *Frequency* adjustment by an oscillator in response to some *periodic* stimulus.

frequency spectrum A plot of the distribution of calculated *powers* as a function of *frequency*, as obtained in a *Fourier analysis*. Also known as *power*-density *spectrum*, *power spectrum*, or spectral density.

function (a) An output *variable* or *dependent variable* whose value is uniquely determined by one or more input (*independent*) *variables*; (b) an equation or relation between two groups A and B such that at least one member of group A is matched with one member of group B (a "single-valued" *function*) or is matched with two or more members of group B (a "multivalued" function). A single-valued function, for instance, is $y=3x$; for any value of x, there is only one value of y, and vice versa. A multivalued function is $y=x^2$; for $y=4$, x can be+2 or -2, that is, one value of y is matched with two values of x. Many authors use "function" to mean single-valued function.

fundamental frequency The *frequency* of the selected *basic wave* in a *Fourier analysis*.

fundamental wave A wave (usually the longest available, or the length of the *time series*) chosen as a standard or reference wave in *Fourier analysis*.

fundamental wavelength The *wavelength* of the selected *basic wave* in a *Fourier analysis*.

G

geometric progression A sequence of terms whose successive members differ by a constant ratio or multiplier. (Also known as a *geometric series*.)

geometric series A series of numbers in which the ratio of each member to its predecessor is the same throughout the sequence. Also known as a *geometric progression*. Example: 1, 2, 4, 8, 16, 32, etc., in which the ratio of any number to its predecessor is 2.

golden mean The unique value obtained by sectioning (a) a straight line such that the ratio of the shorter segment to the longer segment equals the ratio of the longer segment to the total length, or (b) the sides of a rectangle such that the ratio of their difference to the smaller equals the ratio of the shorter to the longer.

Gram-Schmidt orthogonalization Same as *orthogonalization*.

H

Hamiltonian system A *system* with no friction.

harmonic (a) As an adjective, expressible in terms of sine and cosine *functions*; (b) as a noun, any component of a *periodic* quantity having a *frequency* that's an *integer* multiple of a *given fundamental frequency*. In a *Fourier analysis*, for example, the *first harmonic* is the *fundamental wave*; the *second harmonic* is the constituent wave having a frequency twice that of the fundamental wave; the third harmonic is the constituent wave that has a *frequency* three times that of the fundamental wave; etc.

harmonic analysis The *frequency-domain* description of a *time series*, especially of a *periodic function*, by summing sine and cosine functions (its constituent harmonics).

harmonic number The number corresponding to a particular *harmonic*. For example, the harmonic number of the *second harmonic* is two, that of the third harmonic is three, etc.

Hausdorff dimension That critical *dimension* D_H at which the computed value of the *measure* $M\varepsilon$ changes abruptly from zero to infinity (or from infinity to zero), using the relation $M\varepsilon = N\varepsilon^{D_H}\varepsilon^{D-D_H}$.

Hausdorff-Besicovich dimension Same as *Hausdorff dimension*.

Heaviside function A simple mathematical *function* or number that is zero if a specified expression is negative (less than zero) and 1 if it is positive (zero or greater).

Hénon map A *phase* space plot of iterates of the equations $x_{t+1}=y_t+1-ax_{t2}$, and $y_{t+1}=bx_t$ where a and b are *constants*.

high-pass filter A *filter* that lets high frequencies pass and blocks low frequencies.

histogram A bar diagram showing the *frequency distribution* of a *variable*.

homoclinic orbit An *orbit* that is asymptotic to a *fixed point* or *periodic* orbit and that emanates from the same point or orbit (Lorenz 1993).

homoclinic point *The fixed point* from which a *homoclinic orbit* emanates and which it subsequently approaches (Lorenz 1993).

Hopf bifurcation An abrupt increase (by one) in the number of fundamental frequencies of a *system*, caused by the increase of a *control parameter* past a critical value. Commonly applied in the quasiperiodic route to *chaos*. Named for the work done on the subject by mathematician Eberhard Hopf in 1942.

hypercube (a) An imaginary *phase space* zone or subspace of four or more dimensions and characterizable by a length; (b) the multidimensional analog of a cube.

hyperspace Space of more than three dimensions.

hypersphere (a) An imaginary *phase space* zone or subspace of four or more dimensions and characterizable by a length (e.g. radius); (b) the multidimensional analog of a sphere.

I

identity line (or identity map) A 45° straight line on an arithmetic-*scale* two-*coordinate* graph, representing the relation $y=x$.

imaginary number Any number consisting of a real number times the square root of minus one.

incremental redundancy The amount of change in the *redundancy* of a *vector* when we increase its *dimension* by one, at a constant *lag*.

independent variable (a) An input number to a *function*; (b) a *variable* unaffected by the value taken on by other variables; (c) a variable that an experimenter deliberately manipulates, to find its effect on some other quantity (the latter being the dependent variable).

information A numerical *measure* of (a) knowledge or content of any statement, (b) how much is learned when the contents of a message are revealed or (c) the uncertainty in the outcome of an experiment to be done.

information dimension The slope of a straight line on a semilog plot of information, here defined as I_ε (*dependent variable*, on arithmetic *scale*) versus $1/\varepsilon$ (log scale), where ε is characteristic size of measuring device and

$$I_\varepsilon = \sum_{i=1}^{N_\varepsilon} P_i \log_2 (1/P_i).$$

information entropy A *measure* devised by Shannon (Shannon & Weaver 1949; called by him simply "entropy") for the amount of *information* in a message, and identical in equation form to *thermodynamic entropy*.

information theory The formal, standard treatment of *information* as a mathematically definable and measurable quantity.

initial conditions Values of variables at the beginning of any specified time *period*.

inner product Same as *dot product*.

integer A whole number with no decimal or fractional part.

integral (a) A sum obtained by adding a group of separate elements; (b) the result of mathematically integrating a *function* or an equation.

intermittency A complex *steady-state* behavior (often a route to *chaos*) consisting of orderly *periodic* motion (regular oscillations, with no *period-doubling*) interrupted by occasional bursts of *chaos* or *noise* at uneven *intervals*.

interpolation The process of estimating one or more intermediate values between two known values.

interval (a) The length of time between successive events; (b) a *set* of real numbers that fall between two end-member real numbers.

invariant (a) Independent of the particular *coordinates*, that is. unaffected by a change in coordinates or by a particular *transformation*, such as a change from an original *phase space* to a time-delay reconstruction; (b) remaining forever in (never escaping from) a particular region of *phase space*; (c) unchanged by the *system*'s *dynamics* over time.

invariant manifold A collection of *phase space* trajectories, none of which ever leave the collection.

invariant measure (a) A measurable property that's unchanged by transformations; (b) a measurable feature that doesn't change with time, that is, doesn't change under the action of the *dynamics*; (c) a *probability-distribution function* describing the long-time likelihood of finding a *system* in a particular zone of *phase space*.

invariant probability distribution The *frequency distribution* that is approached as time goes to infinity. Also known as (an *attractor*'s) *probability distribution*, *probability density*, *probability density distribution*, *invariant measure*, and other combinations of these terms.

invertible (a) Having an inverse; (b) having a unique successor or predecessor, or in other words, capable of being solved uniquely either forwards or backwards in time. Example: capable of indicating either x_t or x_{t+m}, given the other.

isotropic Independent of direction.

iterate (a) As a verb, to repeat an operation over and over, often with the aim of coming closer and closer to a desired result; (b) as a noun, a value calculated by any mathematical process of successive approximation.

iteration (a) Any process of successive approximation; (b) repeated application of a mathematical procedure, with the outcome of one solution fed back in as input for the next; (c) each successive step of an iterative process.

J

joint entropy The average amount of *information* obtained (or average amount of uncertainty reduced) by individual measurements of two or more systems.

joint probability The *probability* that two specified events (usually independent events) will happen together. "Together" doesn't necessarily mean "simultaneously."

joint probability distribution A list, table, or *function* for two or more *random variables*, giving the *probability* of the joint occurrence of each possible combination of values of those variables.

K

K-S entropy *Kolmogorov-Sinai entropy*.

kernel density estimator A *probability*-estimation technique in which a fixed, local *probability distribution* (kernel) is applied to the neighborhood centered on each datum point.

Koch snowflake A geometric pattern formed from a straight line by applying a particular, constant, repeated generating procedure. The most common generating procedure is to replace the middle third of the line with an equilateral triangle (also known as "*von Koch snowflake*").

Kolmogorov-Sinai (K-S) entropy Entropy (based on sequence probabilities) per unit time in the limits where time goes to infinity and bin size goes to zero. Also known as source entropy, entropy of the source, *measure-theoretic* entropy, *metric-invariant* entropy, and *metric entropy*.

L

lag The basic time *interval* or amount of offset between any two values being compared, within a *time series*.

lag-m autocorrelation coefficient *Autocorrelation coefficient*.

lag-m serial correlation coefficient *Autocorrelation coefficient*.

lag space A special but very common type of *pseudo phase space* in which the axes or dimensions represent successive values of the same feature (x) separated by a constant time *interval*.

lag vector A plotted point as defined by successively lagged values of some *variable* in reconstructed (pseudo) *phase space*.

lagged phase space See *lag space*.

limit cycle A self-sustaining phase space loop that represents *periodic* motion. Hence, it's a *periodic attractor*.

line spectrum *Periodogram*.

linear (a) Pertaining to lines, usually straight ones; (b) having no *variable* raised to a *power* other than one (see *linear equation*).

linear equation An equation having a straight line for its graph, that is, an equation in which the *variables* are raised to the first *power* only (also called a *polynomial* equation of the first degree). Examples of linear equations are $ax+b=0$, $y=c+bx$, and $ax+by+cz=0$, where x, y, and z are variables and a, b, and c are parameters.

linear filter See *filter*.

linear function A mathematical relationship in which the *variables* appear only in the first degree, multiplied by *constants*, and combined only by addition and subtraction.

linear interpolation *Interpolation* based on straight-line relations between observations.

linear system A *system* in which all observations of a given *variable* plot as a straight line (arithmetic scales) against observations of a second (or lagged) variable. In *lag space*, for example, all observations then are said to be "linearly related" to observations at a later time.

local Lyapunov exponent The *exponential* rate of *trajectory* convergence or *divergence* in a local region of an *attractor*.

logarithm An exponent that is the *power* to which a base number (usually 10, 2, or e) is raised to give another number.

logarithmic equation An equation of the form $a^y=cx^d$, in which a, c, and d are *constants*. Such an equation plots as a straight line on semilog paper (with x on the log *scale* and y on the arithmetic scale).

logistic (a) In mathematics and statistics, referring to a special type of so-called growth curve (an expression that specifies how the size of a *population* varies with time) (however, that curve isn't the "logistic equation" of *chaos theory*; see *logistic equation*); (b) in a general sense, skilled in computation; (c) in military usage: referring to the provision of personnel and material.

logistic equation (a) Historically and generally, the relation $x_{t+1}=k/(1+e^{a+bx})$ where the *constant* b is less than zero; (b) in *chaos theory*, the relation $x_{t+1}=kx_t(1-x_t)$. Both equations are iterative types that are popular in biology as models for *population* growth.

low-pass filter A *filter* that lets low frequencies pass and blocks high frequencies.

Lyapunov characteristic exponent *Lyapunov exponent*.

Lyapunov characteristic number *Lyapunov number*.

Lyapunov exponent The average of many local *exponential* rates of convergence or divergence of adjacent trajectories, expressed in *logarithms* and measured over the entire *attractor*. As such, it reflects the average rate of expansion or contraction of neighboring trajectories with time.

Lyapunov number The number whose *logarithm*, to a given base, equals the *Lyapunov exponent*.

M

manifold (a) In general, an object consisting of many diverse elements; (b) in mathematics, a class with subclasses (e.g. a plane is a two-dimensional manifold of points because it is the class of all its points); (c) any smooth geometric object (point, curve, surface, volume, or multidimensional counterpart); (d) in the nonchaotic domain, an *attractor*; (e) the basic space of a *dynamical system*.

map (a) A *function*, mathematical *model*, or rule specifying how a *dynamical system* evolves; (b) a synonym for correspondence, function, or *transformation*; (c) the mathematical process of taking one point to another. In the latter sense, a map tells how *x* will go, by a *discrete* step, to a new value of *x*. More than one *variable* can be involved. Common forms of a map are an iterative equation and a graph; in graphical form, a map shows a historical sequence of values.

mapping (a) A series of *iterations* of a *map*; (b) a *dynamical system* whose *variables* are defined only at *discrete* times; (c) a *function*, correspondence, or *transformation*.

marginal distribution The complete distribution (summing to 1.0) of probabilities for any chosen *variable* within a joint distribution.

marginal probability The *probability* of getting a particular value of one *variable* within a joint distribution, regardless of the value of the other variable.

marginal probability distribution Same as *marginal distribution*.

marginal redundancy Same as *incremental redundancy*.

marginally stable Tending to keep a *perturbation* at about its original magnitude, over time.

mathematical fractal An object having the property that any suitably magnified part looks exactly like the whole.

mean Arithmetic average (sum of values divided by the number of values).

measure (noun) (a) The size or quantity of something expressed in terms of a standard unit; (b) a *scalar* associated with a *vector* and indicating its magnitude and sense but not its orientation; (c) the *probability* of finding a value of a *variable* within a particular domain.

measure-preserving transformation A one-to-one *transformation* made such that the measure of every *set* and of its inverse image agree.

measure-theoretic entropy Same as *Kolmogorov-Sinai* (*K-S*) *entropy*.

metric (noun) (a) In general, a standard of measurement (e.g. "there is no metric for joy"); (b) in mathematics, a way of specifying values of a *variable* or positions of a point (e.g. "a Euclidean metric"); (c) also in mathematics, a differential expression of distance in a generalized *vector* space; (d) (adjective) referring to measurement or to the meter (100 centimeters).

metric entropy Same as *Kolmogorov-Sinai* (*K-S*) *entropy*.

metrical Quantitative.

microstate A *phase space* compartment, possible outcome, or solution.

model A simplified representation of a real phenomenon, in other words, a stripped-down or uncomplicated description or version of a real-world process. Models can be classified into physical (*scale*), mathematical, analog, and conceptual models.

monotonic Always increasing or always decreasing, that is, pertaining to a *continuous* line along which the slope keeps the same sign at all points.

moving average A transformed *time series* in which each value in the original time series is replaced by the *arithmetic mean* of that value and a specified, constant number of adjacent values that come before and after it.

multifractal Having different fractal scalings (*dimension* values) at different times or places, on an *attractor*.

multitaper spectral analysis A type of *Fourier analysis* that optimally combines *information* from *orthogonal* tapered estimates to minimize leakage and bias in the spectral estimate at each *frequency*.

mutual information (a) The amount by which a measurement of one *variable* reduces the uncertainty in another; (b) the quantity of *information* one *system* or variable contains about another; (c) a *measure* of the degree to which two processes or *random variables* are mutually related; (d) the amount of surprise or predictability associated with a new measurement.

mutually exclusive A statistical term meaning that only one of various possible outcomes can occur at a time.

N

nat The unit of measurement for the separation (or convergence) rate of two neighboring trajectories when logs are taken to base 2.

natural fractal An object having the property that any suitably magnified part looks approximately like the whole, the differences being minor, negligible and ascribable to chance.

natural measure A *measure* or observation that is a convergent time-average value.

natural probability measure *Invariant measure*.

nearest neighbor A *pseudo phase space* datum point that plots close to another point, for a particular *embedding dimension*.

next-amplitude map A one-dimensional *first-return map* that uses only the high values of successive oscillations (the successive maxima) of the *time series*.

next-amplitude plot *Next-amplitude map*.

noise (a) Any unwanted disturbance superimposed on useful data and tending to obscure their *information* content; (b) unexplainable variability or fluctuation in a quantity with time; (c) in a general sense, anything that impedes communication. Such disturbance or variability may be *random* fluctuations, reading *errors*, analytical errors, sampling errors, and other factors.

nonautonomous Time-dependent.

noninvertible Not capable of being uniquely solved backwards in time, for example, not able to indicate x_t from a given value of x_{t+m}.

nonlinear Not having a straight-line relationship, that is, referring to a response that isn't directly (or inversely) proportional to a given *variable*.

nonlinear dynamics The study of motion that doesn't follow a straight-line relation, that is, the study of nonlinear movement or evolution. As such, nonlinear dynamics is a broad field that includes *chaos theory* and many mathematical tools used in analyzing complex temporal phenomena.

nonlinear equation An equation in *variables x* and *y* which cannot be put in the form y=c+bx, where b and c are *coefficients*.

nonlinear system A *system* in which the observations of a given *variable* do not plot as a straight line (on arithmetic scales) against observations of a second (or lagged) variable.

nonmonotonic Pertaining to a *continuous* line along which the slope changes sign.

nonparametric Not involving assumptions about specific values of parameters or about the form of a distribution. ("Parameter" in this case usually refers to the statistical definition, namely numerical characteristics of a *population*.)

nonperiodic Same as *aperiodic*.

nonstationary Said of a *time series* for which (a) a *moving average* isn't approximately *constant* with time or (b) the *mean* and *variance* aren't approximately constant with time.

non-uniformity factor (NUF) The *standard deviation* of the local rates of convergence or divergence of neighboring trajectories, sampled over the entire *attractor*.

norm The magnitude of a *vector*.

normal distribution A special type of symmetrical (bell-shaped) and *continuous frequency distribution*, the graph or curve of which is given by a particular general equation (not reproduced here).

normalization The process of adjusting or converting one or more values to some standard *scale*. The standard scale for a group of values usually is from 0 to 1. The conversion then consists of dividing each value of the original dataset by some maximum reference quantity, such as the greatest value in the dataset or a theoretical maximum value. A *vector* is normalized by dividing it by its magnitude, yielding a so-called *unit vector*. *Probability distributions* are normalized by changing the *variable* so that the distribution has a *mean* of 0 and a *variance* of 1.

null hypothesis A hypothesis that supposes no significant difference between a statistic for one group and the same statistic for another.

O

observable (noun) A physical quantity that can be measured.

one-dimensional map An equation (in the form of either a written statement or a graphical plot) that gives the value of a *variable* as a *function* of its value at one or more earlier times. The typical expression is x_{t+1} as some function of x_t.

orbit The path through space taken by a moving body or point. Examples: (a) a *trajectory* that completes a circuit (as, "the Moon orbits the Earth"); (b) a trajectory or chronological sequence of states as represented in *phase space*.

ordinary probability A nonstandard term used in this book to mean the likelihood of getting a specified *state* at a particular time.

origin A reference point in ordinary space or *phase space*. Most often, it's the point at which all *variables* have a value of zero.

orthogonal (a) Perpendicular or normal (having to do with right angles); (b) unrelated or independent; (c) said of elements having the property that the product of any pair of them is zero.

orthogonalization A procedure for realigning two or more nonorthogonal vectors into a *set* of an equal number of mutually *orthogonal* vectors, all of which have the same *origin*.

orthonormal Said of axes or vectors that are mutually perpendicular and of normalized (unit) length.

orthonormalization The process of reducing mutually *orthogonal* vectors to unit length.

P

parabola The curve or path of a moving point that remains equidistant from a *fixed point* (the focus) and from a fixed straight line (the directrix).

parallelogram law A graphical method for adding two vectors, whereby their starting points are placed together to form two sides of a parallelogram, the two opposite sides then are drawn in, and the sum or *resultant* is given by the diagonal drawn from the two starting points to the opposite corner.

parameter (a) In physics, a controllable quantity kept *constant* in an experiment as a *measure* of some influential or driving environmental condition; (b) in mathematics, an arbitrary constant in a mathematical expression and that can be changed to provide different cases of the phenomenon represented; (c) in mathematics, a special *variable* in terms of which two or more other variables can be written; (d) in statistics, a numerical characteristic of a *population* (such as, for example, the *arithmetic mean*).

parametric Involving parameters.

parametric equations In math, a *set* of equations that express quantities in terms of the same set of *independent variables* (called parameters).

partition (a) As a noun, the collection of possible outcomes of an experiment; (b) as a noun, a compartment (bin, cell, etc.) or group of compartments within *phase space*; (c) as a verb, to subdivide a phase space or the range of values of a *variable* into a *set* of *discrete*, connected subintervals.

partitioning The process of dividing a dataset into classes or groups such that any one observation is a member of only one class.

percentage-of-moving-average method Same as *ratio-to-moving-average method*.

period The amount of time needed for a *system* to return to its original *state*, that is, the time required for a regularly repeating phenomenon to repeat itself once.

period-doubling A process whereby increases in a *control parameter* in certain iterated equations produce trajectories made up of a successively doubled number of *attractor* values (e.g. 2, 4, 8, 16, etc. *attractor* values), eventually leading to *chaos*.

periodic Regularly repeating. The repetition can be exact (in a pure mathematical sense) or approximate (as with virtually all measured data).

periodic attractor An *attractor* consisting of a finite *set* of points that form a closed loop or *cycle* in *phase space*.

periodic toroidal attractor A *torus* on which the composite *trajectory*'s motion repeats itself regularly, i.e., on an exact or *integer* basis.

periodic points Points (values) that are members of a *cycle* (and hence that recur periodically).

periodicity (a) The number of measured observations or *iterations* to a *cycle* or *period*; (b) the quality of recurring at a definite *interval* (in other words, repetition of a given pattern over fixed intervals).

periodogram A graph showing the relative strengths (usually variances) of constituent waves at their *discrete* frequencies. Discrete *wave periods* can be used in place of frequencies and are probably the origin of the name.

perturbation (a) A difference between two neighboring observations, at a given time: (b) an intentional displacement of an observation, at any given time; (c) a deliberate change (usually slight) in one or more parameters of an equation.

perturbation vector *Principal axis*.

phase (a) The stage that a *dynamical system* is in at any particular time; (b) the fraction of a *cycle* through which a wave has passed at any instant.

phase angle The starting angle or reference point within a wave *cycle*, from which a process begins.

phase diagram Same as *phase portrait*.

phase locking The beating in harmony or resonating together of many individual oscillators.

phase portrait A *phase space* plot of a *system*'s various possible conditions, as shown by one or more trajectories.

phase space An abstract mathematical space in which *coordinates* represent the *variables* needed to specify the *phase* (or *state*) of a *dynamical system* at any time.

phase space portrait Same as *phase portrait*.

phase space reconstruction See *reconstruction of phase space*.

pitchfork bifurcation The spawning of two *equilibrium points* from one *equilibrium point*, as in *period-doubling*.

Poincaré map Same as *return map*.

Poincaré return map Same as *return map*.

Poincaré section A slice or cross section in *phase space*, cutting through an *attractor* transverse to the *flow*. The section shows dots that represent the *trajectory*'s intersections with the plane or slice.

point A *state* of a *system* (values of its *variables* at a given time).

point attractor A single *fixed point* in *phase space*, representing the final or equilibrium *state* of a *system* that comes to rest with the passage of time.

polynomial An algebraic expression involving two or more summed terms, each term consisting of a constant multiplier and one or more *variables* raised to nonnegative *integer* powers. (In some mathematics texts, a polynomial can have one term.)

population Any group of items that is well defined (has a common characteristic).

power (a) The number of times, as indicated by an exponent, in which a given number (say, x) is used when being repeatedly multiplied by itself (e.g. $x^3=x\cdot x\cdot x$, and we say x is raised to the third power); (b) a term used in *Fourier analysis* as a synonym for *variance*.

power equation An equation of the form $y=ax^c$, where x and y are *variables* and a and c are *constants*. A power equation plots as a straight line on log paper. See also *power law*.

power law Any equation in which the *dependent variable* varies in proportion to a specified *power* of the *independent variable*. Synonymous with *power equation*.

power spectrum The ensemble of variances (powers) calculated in a *Fourier analysis* and plotted against their respective wave characteristic (*frequency*, *wavelength*, *period*, or *harmonic number*).

precision (a) In general, the *accuracy* with which a calculation is made; (b) reproducibility, as with repeated samplings.

principal axis Any of a body's two or three axes of structural symmetry that were mutually perpendicular before deformation.

probability (a) Limiting *relative frequency*; (b) the ratio of the number of times an event occurs to the number of trials; (c) the encoding of all that we know about the likelihood that a particular event will happen, with the encoding expressed as a number between zero (no chance that the event will take place) and 1 (certainty that the event will happen); (d) the formal study of chance occurrences.

probability density Same as *probability density function*.

probability density curve Same as *probability density function*.

probability density distribution Same as *probability density function*.

probability density function The limiting *relative frequency density* as sample size becomes infinitely large and bin width goes to zero, for a *continuous random variable*.

probability distribution A list of the possible outcomes of an experiment and their associated probabilities, that is, a table or *function* that assigns to each possible value of a *random variable* the *probability* of that value's occurrence. Also known (possibly depending on whether the *variable* is discrete or *continuous*) as *distribution function*, statistical distribution, *probability density function*, *density function*, or *frequency* function.

projection The image of a geometric object or *vector* superimposed on some other vector. As such, the projection is a new vector and is called the projection of the first vector onto the second.

pseudo phase space An imaginary graphical space in which the first *coordinate* represents a physical feature and the other coordinates represent lagged values of that feature.

Pythagorean theorem The theorem that, in a right triangle, the square of the length of the hypotenuse equals the sum of the squared lengths of the other two sides.

Q

quadratic Of the second degree, that is, involving one or more unknowns (*variables*) that are squared but none that are raised to a higher *power*.

quadratic map An iterated equation (and sometimes also the *phase space* plot of those iterates) in which x_{t+1} is given as some *function of* x_{t2}.

quasiperiodic toroidal attractor A *torus* on which the composite *trajectory*'s motion almost but not exactly repeats itself regularly.

quasiperiodicity (a) A *dynamical* state characterized by the superposition of two or more *periodic* motions (characteristic frequencies); (b) a route to *chaos* caused by the simultaneous occurrence of two periodicities whose frequencies are out of *phase* (not commensurate) with one another.

R

random (a) Based strictly on a chance mechanism (the luck of the draw), with negligible *deterministic* effects (the definition used in this book); (b) disorganized or haphazard; (c) providing every member an equal chance of selection; (d) unlikely repeatability of any observation; (e) unpredictability of individual events to within any reasonable degree of certainty; (f) difficult to compute.

random process A process based on *random* selection, as in a process in which new observations depend strictly on chance, or in which every possible observation has an equal chance of selection.

random variable (a) A *variable* whose value cannot be foretold, except statistically; (b) a variable that takes on different values at different random events (experiments or trials).

ratio-to-moving-average method A *classical decomposition* technique for isolating the value of [*seasonality* times *random* deviations] by dividing the value of the *variable* by the *moving-average* value, at each sequential time.

reconstruction dimension The *embedding dimension* in which an *attractor* is *reconstructed*.

reconstruction of attractor The graphical or analytical recreation of the *topology* or es sence of a multidimensional *attractor* by analyzing lagged values of one selected *variable*.

reconstruction of phase space A pseudo (lagged) *phase space* plot in two or three dimensions, made with the hope of seeing an *attractor* (if there is one). Also known as *phase space reconstruction*, *state space reconstruction*, *phase portrait reconstruction*, *trajectory* reconstruction, and similar expressions.

recursive Referring to any process or *function* in which each new value is generated from the preceding value. (Hence, iterative.) Examples: (a) the *logistic equation* of *chaos theory*; (b) forecasting techniques in which each successive forecast incorporates the preceding forecast.

reductionism The notion that the world is an assemblage of parts.

redundancy The multidimensional generalization of *mutual information*.

regression A mathematical *model* stating how a *dependent variable* varies with one or more *independent variables*.

relative entropy A ratio of two entropies (typically an actual *entropy* to a reference entropy), reflecting the magnitude of one relative to the other.

relative frequency The ratio of number of occurrences of an outcome to total number of possible occurrences.

relative frequency density The ratio of *relative frequency* to class width.

relative frequency distribution A list of relative frequencies, for a particular process or *system*.

renormalization A mathematical scaling technique consisting of rescaling a physical *variable* and transforming a *control parameter* such that the properties of an equation at one *scale* can be related to those at another scale, and the properties at the limiting scale (infinity) can be determined.

reorthonormalization A procedure for again realigning all vectors to be mutually perpendicular and then making each *vector* of unit length.

residual The difference between an observed value and the value predicted by a *model* fitted to the general dataset.

resolution (a) The act of breaking something up into its constituent elements; (b) the act of determining or rendering distinguishable the individual parts of something; (c) the size of the biggest element in a *partitioning*, such as the length of a box or radius of a sphere.

resultant A *vector* that is the sum of two or more other vectors.

return map A rule, *function*, or *model* that tells where the *trajectory* will next cross a particular one-dimensional *Poincaré section*, given the location of a previous crossing at that section. "Locations" along the line are described in terms of distance (L) along the line from some arbitrary *origin* on the line. Commonly, a *first-return* (or first-order-return) *map* is a model of L_{n+1} as a function of L_n, for the Poincaré section; a *second-return* (or *second-order-return*) *map* is a model of L_{n+2} versus L_n; and so on.

Richardson plot A logarithmic graph of a coastline's measured length (on the ordinate) versus the associated ruler length.

robust Resistant to, or steady under, *perturbation*.

S

saddle (saddle point) An *unstable steady-state phase space* condition (*fixed point*, *limit cycle*, etc.), the *flow* near which is *stable* in some directions (*converges* toward the fixed point) and unstable in others (moves away from the fixed point).

scalar A number representing a magnitude only, as might be indicated on a simple *scale*.

scalar product Same as *dot product*.

scalar time series An ordinary *time series*, in which each successive measurement is recorded along with its time or order of measurement (and, hence, synonymous with the general meaning of *time series*).

scale (a) A sequence of collinear marks, usually representing equal steps or increases, used as a reference in measuring; (b) the proportion that a *model* bears to whatever it represents (e.g. a scale of one centimeter to a kilometer); (c) to reduce or enlarge according to a fixed ratio or proportion; (d) to determine a quantity at some order of magnitude by using data or relationships known to be valid at other orders of magnitude.

scale invariance The property whereby an object's size cannot be determined or estimated without additional *information* because the item looks like any of a family of objects of different size but similar appearance.

scale length Length of tool or device used (usually by successive increments) to estimate an object's size.

scaling ratio A proportion (and hence a value less than 1.0), used in scaling, and equal to that fraction of the original object which the new (smaller) object represents as measured in the direction of any *topological dimension*.

scaling region A straight-line relation (the slope of which is the *correlation dimension*) on a plot of log of *correlation sum* versus log of measuring radius.

Schmidt orthogonalization Same as *orthogonalization*.

seasonality Same as *periodicity* (the tendency for any pattern to repeat itself over fixed *intervals* of time). In this book, the fixed intervals of time for seasonality are arbitrarily taken as one year or less, whereas periodicity is used as a generic or general term and can have any fixed interval.

second derivative The *derivative* of the *first derivative* of a *function*.

second-differencing The computation of differences of differenced data.

second harmonic A wave that has twice the *frequency* as the *fundamental wave*, in *Fourier analysis*.

second-order return map *Second-return map*.

second-return map A *pseudo phase space* plot of *Poincaré-section* data on which each observation at the section is plotted against its second return to that section.

self-affine That property whereby an object is reproduced by a *transformation* involving different scaling factors for different dimensions or *variables*.

self-entropy (a) The *entropy* for one *system* or *variable*; (b) the average amount of *information* gained by making a measurement of(or provided by) a *random process* about itself.

self-information Same as *self-entropy*.

self-organization The spontaneous, self-generated occurrence of some kind of pattern or structure from an (usually) orderless *dynamical system*.

self-organized criticality The property whereby a *dynamical system* naturally evolves toward a critical *state* (a condition where the *system* undergoes a sudden change). Various *perturbations* of the system at that critical state provoke different responses that follow a *power law*.

self-similarity The property whereby any part of an object, suitably magnified, looks mostly like the whole.

self-similarity dimension *Similarity dimension*.

sensitive dependence on initial conditions (a) The quality whereby two slightly different values of an input *variable* evolve to two vastly different trajectories; (b) the quality whereby two initially nearby trajectories diverge exponentially with time.

sensitivity to initial conditions *Sensitive dependence on initial conditions*.

separatrix A boundary between regions of *phase space*, such as a boundary between two basins of attraction in a *dissipative system*.

sequence probability A nonstandard term used in this book to mean the *joint probability* of certain successive events in time and representing the *probability* that a given sequence of events will take place.

serial correlation Same as *autocorrelation*.

set A group of points or values that have a common characteristic or rule of formation.

similarity dimension A *dimension D* defined as $D=\log N/\log(1/r)$, used in transforming figures into similar figures of different sizes.

sine wave A wave corresponding to the equation $y=\sin x$, or more generally, $y=A\sin(x+\phi)$, where A is *wave amplitude* and ϕ is *phase angle*.

singular spectral analysis An alternative to *Fourier analysis* that uses empirical *basis functions* formed by a principal component analysis of the embedded *phase space* data matrix instead of the sine/cosine *basis* functions.

singular system analysis A *phase space reconstruction* technique in which *orthonormal* reconstruction axes near each *point* x_t are a maximal set of linearly independent vectors that are derived from the local distribution of points near x_t by *singular value decomposition*.

singular value decomposition *Singular system analysis*.

sink A *phase space* point toward which all nearby points *flow*.

sinusoid (a) The sine curve ($y=\sin x$); (b) any curve derivable from the sine curve by multiplying by a *constant* or by adding a constant; (c) any curve shaped like a sine curve but with different *amplitude*, *period*, or intercepts with the *axis*.

sinusoidal (a) Relating to, and shaped like, a sine curve; (b) describable by a sine or cosine equation.

smooth Said of a *frequency distribution* that is *continuous* or, in mathematical terms, is differentiable at every point.

smoothing (a) In general, reduction or removal of discontinuities or fluctuations in data; (b) more specifically, application of a *low-pass filter*, that is, any mathematical technique that combines two or more observations to reduce variability and to highlight general trends.

smoothing parameter Bin width, usually as required in methods for estimating *probability* densities.

soliton Solitary wave.

source A *phase space* point away from which all nearby *points flow*.

space-filling curve A curve passing through every point in a space of two or more dimensions.

spectral analysis The calculation and interpretation of a *spectrum*, especially in terms of the *frequency-domain* characteristics of a *time series*. Generally used synonymously with *Fourier analysis*.

spectrum A collection of all the frequencies, *wavelengths*, or similar components of a complex process, showing the distribution of those components according to their magnitudes.

spline (a) A flexible strip used in drafting to form—and to help draw—a curve between two *fixed points*; (b) a smoothly varying curve between two data points, often fitted by a *cubic polynomial* and used to interpolate "data" values between the two measured points.

stable Tending to dampen *perturbations* or initial differences, over time.

stable fixed point A type of *attractor* in the form of a *phase space* point that attracts all neighboring points, that is, a point to which all iterates beginning from some other point *converge*.

stable manifold The set of points that asymptotically approach a given point, over time.

standard deviation A descriptive statistic that reflects the spread of a group of values about their *mean*, usually computed as:

$$\left[\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2\right]^{0.5}.$$

As such, it's the square root of the average squared deviation, or the square root of the *variance*.

standard phase space A term used in this book for the usual *phase space*, the *coordinates* for which represent the measured and different physical features that are to be plotted on the associated graph or *phase portrait*.

standardization A *transformation* done by subtracting the *mean* from each observation of a dataset and then dividing each such difference by the *standard deviation*. Such a transformation, for example, removes *seasonality*. It converts data of different units to a common or standard *scale*, namely a scale for which the mean is zero and the *standard deviation* is one. The new scale measures the transformed *variable* in terms of the deviation from its mean, in units of standard deviation, and is dimensionless.

state (a) The condition of a *system* or values of its *variables*, at any one time; (b) an arbitrarily defined subrange (usually with specified numerical boundaries) that one or more variables of a system can be in at one particular time.

state space Same as *phase space*.

state space reconstruction See *reconstruction of phase space*.

state space vector Same as *lag vector*.

stationary Time-*invariant*, that is, (a) lacking a *trend*; or (b) (more rigidly), keeping a constant *mean* and *variance* with time; or (c) (more formally still), said of a randomlike process whose statistical properties are independent of time. ("Independence of time" is often a matter of the particular time scale being used.)

statistical self-similarity The property whereby any suitably magnified part of an object looks approximately like the whole, the differences being minor, negligible and ascribable to chance.

steady state (a) A condition that doesn't change with time; (b) the *state* toward which the *system*'s behavior becomes asymptotic as time goes to infinity. The associated equation gives a constant solution.

stochastic (a) Having no regularities in behavior; (b) characterizable only by statistical properties, thus involving randomness and *probability* as opposed to being mainly *deterministic*; (c) *random*; (d) developed according to a probabilistic *model*.

strange attractor (a) Same as *chaotic attractor*; (b) an *attractor* having such geometrical features as *fractal dimension*, *Cantor-set* structure, and so on.

stretching (a) A topologist's interpretation of either (1) the amplification of a certain range of input values to a larger range of output values during *iteration* or (2) the *phase space exponential* divergence of two nearby chaotic trajectories; (b) a *transformation* of the form $x'=ax$ and $y'=ay$, where $a>1$.

stroboscopic map A lag-time *phase space model* in which data are taken at equal time *intervals*.

subharmonic bifurcation *Period-doubling*.

successive-maxima plot Same as *next-amplitude map*.

surface of section *Poincaré section*.

surrogate data Artificially generated data that mimic certain selected features of an observed *time series* but that are otherwise *stochastic*.

system (a) An arrangement of interacting parts, units, components, *variables*, etc. into a whole; (b) a group, series, or sequence of elements, often in the form of a chronological dataset.

T

tangent bifurcation That special situation on a *one-dimensional map* whereby the *function* is tangent to the *identity line*.

theoretic (a) Restricted to theory; (b) lacking verification.

thermodynamic entropy A *measure* conceived by Clausius for describing the unavailable energy of a closed system such as a heat-producing engine, and subsequently modified for application to other thermodynamic systems.

time-delay method A lag-time analytical technique that uses data of just one measured physical feature (x) (regardless of any other features that may have been measured), whereby x_t is compared to one or more lagged subseries, often with the aim of reconstructing an *attractor*.

time domain The representation of *time series* data in their raw or unaltered form.

time series (a) A chronological list or plot of the values of any *variable* or variables and their time or order of measurement; (b) in a narrow and more rare sense, a *set* of values that vary randomly with time.

topological dimension An *integer* that reflects the *complexity* of a geometric continuum, equal to 1+the *Euclidean dimension* of the simplest geometric object that can subdivide that continuum. The topological dimension usually has the same value as the Euclidean dimension.

topology (a) That branch of geometry that deals with the properties of figures that are unaltered by imposed deformations or cumulative transformations; (b) the branch of mathematics that studies the qualitative properties of spaces, as opposed to the more delicate and refined geometric or analytic properties.

toroidal attractor *Torus*.

torus (a) A three-dimensional ring- or doughnut-shaped surface or solid generated by rotating a circle about any *axis* that doesn't intersect that circle; (b) an *attractor* that represents two or more *limit cycles*. Also called a toroid.

trajectory (a) A path taken by a moving body or point (and hence an *orbit*); (b) a sequence of measured values or list of successive states of a *dynamical system*; (c) a solution to a *differential equation*; (d) graphically, a line on a *phase space* plot, connecting points in chronological order.

transformation (a) A change in the numerical description or *scale* of measurement of a *variable* (in other words, a *filter*); (b) a change in position or direction of the axes of a coordinate *system*, without altering their relative angles; (c) a *mapping* between two spaces.

transient An early, atypical observation or temporary behavior in a *system* when first activated that dies out with time.

translation A shifting (*transformation*) of the axes of a coordinate *system* while keeping the same orientation of those axes. For example, say a is the amount that the *x axis* is shifted and *b* is the amount that the *y* axis is shifted; then the *origin* of new axis *x'* is at $x+a$ and the origin of new axis *y*' is at $y+b$; any point plotted on the new *coordinates* is located at $x'=x-a$ and $y'=y-b$.

trend (a) A systematic change, prevailing tendency, or general drift (in *chaos*, usually a prevailing direction of plotted points over some *period* of time); (b) a nonconstant *mean*.

triangle law A graphical method of adding two vectors whereby the starting point of the second *vector* is placed at the terminal point of the first vector, the sum or *resultant* being given by the new vector drawn from the starting point of the first vector to the terminal point of the second.

truncate (a) To shorten a number by keeping only the first few (significant) *digits* and discarding all others; (b) to approximate an infinite series by a finite number of terms; (c) to exclude sample values that are greater (or less) than a specified constant value.

two-dimensional map A pair of equations, in each of which x_t and y_t together are used to yield x_{t+m} (with the first equation) and y_{t+m} (with the second equation).

U

unit vector A *vector* having a magnitude of 1. It's usually obtained by dividing a vector by its length (magnitude).

universal Typical of entire classes of systems.

universality The property whereby many seemingly unrelated systems or equations behave alike in a particular respect, so that they can be grouped into a class by their generic behavior.

unstable Tending to amplify *perturbations* or initial differences, over time.

unstable fixed point A *fixed point* from which successive iterates move farther and farther away.

unstable manifold The *set* of points that exponentially diverge from a given point, over time.

unstable orbit (unstable trajectory) A *trajectory* for which, arbitrarily close to any input value, there's another possible input which gives rise to a vastly different trajectory.

V

variable A characteristic or property that can have different numerical values.

variable-partition histogramming Same as *adaptive histogramming*.

variance A *measure* of the spread or dispersion of a group of values about their *mean*, specifically the average squared deviation from the mean, or
$\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2$.

vector A directed straight line, that is, a straight line representing a quantity that has both magnitude and direction, drawn from its starting point (point of *origin*) to its terminal point.

vector array The list of paired values that make up a *vector time series*.

vector time series A listing or series of measurements of a *variable* in which each value is associated not with its time or order of measurement explicitly but rather with its value at some constant *lag* time later.

von Koch snowflake *Koch snowflake*.

W

wave amplitude A *measure* of the maximum height of a wave, taken either as the vertical height from trough to crest or as half the vertical height from trough to crest.

wave frequency See *frequency*.

wavelength The distance from any point on a wave to the equivalent point on the next wave.

wave period The time needed for one full *wavelength* or *cycle*.

weight A factor by which some quantity is multiplied and that reflects that quantity's relative importance.

weighted sum A sum obtained by adding weighted quantities.

weighting The multiplication of each item of data by some number that reflects the item's relative importance.

white noise Data that are mutually unrelated. For instance, an observation made at one time has no relation to observations made at earlier times. Such *noise* might, for example, be generated from independent, identically distributed observations of a *variable*. The label "white" comes from the analogy with white light and means that all possible *periodic* oscillations or frequencies are present with equal *power* or *variance*.

window A selected subrange or *interval* of values.

REFERENCES

Abarbanel, H. D. I. 1992. Local and global Lyapunov exponents on a strange attractor. In *Nonlinear modeling and forecasting*, M. Casdagli & S. Eubank (eds), 229-47. Redwood City, California: Addison-Wesley.

Abarbanel, H. D. I., R. Brown, J. B. Kladtke 1990. Prediction in chaotic nonlinear systems: methods for time series with broadband Fourier spectra. *Physical Review A* **41**(4), 1782-807.

Abarbanel, H. D. I., R. Brown, M. B. Kennel 1991a. Variation of Lyapunov exponents on a strange attractor. *Journal of Nonlinear Science* **1**, 175-99.

———. 1991b. Lyapunov exponents in chaotic systems: their importance and their evaluation using observed data. *International Journal of Modern Physics B* **5**(9), 1347-75.

———. 1992. Local Lyapunov exponents computed from observed data. *Journal of Nonlinear Science* **2**, 343-65.

Abarbanel, H. D. I., R. Brown, J. J. Sidorowich, L. S. Tsimring 1993. The analysis of observed chaotic data in physical systems. *Reviews of Modern Physics* **65**(4), 1331-92.

Abarbanel, H. D. I & U. Lall 1996. Nonlinear dynamics of the Great Salt Lake: system identification and prediction. *Climate Dynamics* **12**, 287-97.

Abraham, R. H. & C. D. Shaw 1982. *Dynamics-the geometry of behavior*, part 1: *Periodic behavior*. Santa Cruz, California: Aerial Press.

———. 1983. *Dynamics-the geometry of behavior*, part 2: *Chaotic behavior*. Santa Cruz, California: Aerial Press.

Abraham, N. B., A. M. Albano, N. B. Tufillaro 1989. Complexity and chaos. In *Measures of complexity and chaos*, N. B. Abraham, A. B. Albano, A. Passamante, P. E. Rapp (eds). 1-27. New York: Plenum.

Albano, A. M., A. I. Mees, G. C. de Guzman, P. E. Rapp 1987. Data requirements for reliable estimation of correlation dimensions. In *Chaos in biological systems*, H. Degn, A. V. Holden, L. F. Isen (eds), 207-220. New York: Plenum.

Aleksic, Z. 1991. Estimating the embedding dimension. *Physica D* **52**, 362-8.

Amato, I. 1992. Chaos breaks out at NIH, but order may come of it. *Science* **256**(26 June), 1763-4.

Anonymous 1994. Balancing broomsticks. *The Economist* **331**(7869), 85-6.

Atmanspacher, H. & H. Scheingraber 1987. A fundamental link between system theory and statistical mechanics. *Foundations of Physics* **17**(9), 939-63.

Atmanspacher, H., H. Scheingraber, W. Voges 1988. Global scaling properties of a chaotic attractor reconstructed from experimental data. *Physical Review A* **37**(4), 1314-22.

Barcellos, A. 1984. The fractal geometry of Mandelbrot. *College Mathematics Journal* **15**, 98-114.

Bergé, P., Y. Pomeau, C. Vidal 1984. *Order within chaos*. New York: John Wiley.

Berryman, A. A. 1991. Chaos in ecology and resource management: what causes it and how to avoid it. In *Chaos and insect ecology*, J. A. Logan & F. P. Hain (eds), 23-38. Information Series 91-3, Virginia Agricultural Experiment Station and Virginia Polytechnic Institute and State University.

Berryman, A. A. & J. A. Millstein 1989. Are ecological systems chaotic—and if not, why not. *Trends in Ecology and Evolution* **4**(1), 26-8.

Bloomfield, P. 1976. *Fourier analysis of time series: an introduction*. New York: John Wiley.

Borowitz, S. & L. A. Bornstein 1968. *A contemporary view of elementary physics*. New York: McGraw-Hill.

Bracewell, R. N. 1978. *The Fourier transform and its applications*, 2nd edn. New York: McGraw-Hill.

Briggs, J. & F. D. Peat 1989. *The turbulent mirror*. New York: Harper & Row.

Briggs, K. 1990. An improved method for estimating Lyapunov exponents of chaotic time series. *Physics Letters A* **151**(1-2), 27-32.

Brigham, E. 0. 1988. *The fast Fourier transform and its applications*. Englewood Cliffs, New Jersey: Prentice-Hall.

Broomhead, D. S. & G. P. King 1986. Extracting qualitative dynamics from experimental data. *Physica D* **20**(2/3), 217-36.

Broomhead, D. S., R. Jones, G. P. King 1987. Topological dimension and local coordinates from time series data. *Journal of Physics A* **20**, L563-L569.

Bryant, P. 1992. Computation of Lyapunov exponents from experimental data. In *Proceedings of the first experimental chaos conference*, S. Vohra, M. Spano, M. Shlesinger, L. Pecora, W. Ditto (eds), 11-23. Singapore: World Scientific.

Bryant, P., R. Brown, H. D. I. Abarbanel 1990. Lyapunov exponents from observed time series. *Physical Review Letters* **65**(13), 1523-6.

Burrough, P. A. 1984. The application of fractal ideas to geophysical phenomena. *The Institute of Mathematics and its Applications* **20**(March/April), 36-42.

Çambel, A. B. 1993. *Applied chaos theory: a paradigm for complexity*. Boston: Academic Press.

Campbell, D. 1989. Introduction to nonlinear phenomena. In *Lectures in the sciences of complexity*, D. L. Stein (ed.), 3-105. Redwood City, California: Addison-Wesley.

Campbell, D.,J. Crutchfield, D. Farmer, E. Jen 1985. Experimental mathematics: the role of computation in nonlinear science. *Communications of the Association for Computing Machinery* **28**(4), 374-84.

Caputo, J. G. & P. Atten 1987. Metric entropy: an experimental means for characterizing and quantifying chaos. *Physical Review A* **35**(3), 1311-16.

Casdagli, M. 1989. Nonlinear prediction of chaotic time series. *Physica D* **35**, 335-56.

Caswell, W. E. & J. A. Yorke 1986. Invisible errors in dimension calculations: geometric and systematic effects. See Mayer-Kress (1986: 123-36).

Cipra, B. 1993. If you can't see it, don't believe it. *Science* **259**(January 1), 26-7.

Collet, P. & J-P. Eckmann 1980. *Iterated maps on the interval as dynamical systems*. Basel: Birkhäuser.

Cover, T. M. & J. A. Thomas 1991. *Elements of information theory*. New York: John Wiley.

Cremers, J. & A. Hübler 1987. Construction of differential equations from experimental data *Zeitschrift für Naturforschung*, Series A **42**(8), 797-802 .

Crutchfield, J. P., J. D. Farmer, N. H. Packard, R. S. Shaw 1986. Chaos. *Scientific American* **255**(6), 46-57.

Crutchfield, J. P. & B. S. McNamara 1987. Equations of motion from a data series. *Complex Systems* **1**, 417-52.

Crutchfield, J. P. & K. Young 1989. Inferring statistical complexity. *Physical Review Letters* **63**(2), 105-108.

Cvitanovic, P. (ed.) 1984. *Universality in chaos.* Bristol: Adam Hilger.

Davis, J. C. 1986. *Statistics and data analysis in geology,* 2nd edn. New York: John Wiley.

Denbigh, K. G. & J. S. Denbigh 1985. *Entropy in relation to incomplete knowledge*. Cambridge: Cambridge University Press.

Devaney, R. L. 1988. Fractal patterns arising in chaotic dynamical systems. See Peitgen & Saupe (1988: 137-67).

Ditto, W. L. & L. M. Pecora 1993. Mastering chaos. *Scientific American* **269**(August), 78-84.

Dvorák, I. & J. Klaschka 1990. Modification of the Grassberger-Procaccia algorithm for estimating the correlation exponent of chaotic systems with high embedding dimension. *Physics Letters A* **145**(5), 225-31.

Eckmann, J-P. & D. Ruelle 1985. Ergodic theory of chaos and strange attractors. *Reviews of Modern Physics* **57**(3) (part 1), 617-56.

———. 1992. Fundamental limitations for estimating dimensions and Lyapunov exponents in dynamical systems. *Physica D* **56**, 185-7.

Eckmann, J-P., S. Oliffson-Kamphorst, D. Ruelle, S. Ciliberto 1986. Lyapunov exponents from time series. *Physical Review A* **34**(6), 4971-9.

Elliott, D. F. & K. R. Rao 1982. *Fast transforms: algorithms, analyses, applications*. Orlando, Florida: Academic Press.

Ellner, S. 1991. Detecting low-dimensional chaos in population dynamics data: a critical review. In *Chaos and insect ecology*, J. A. Logan & F. P. Hain (eds), 63-90. Information Series 91-3, Virginia Agricultural Experiment Station and Virginia Polytechnic Institute and State University.

Essex, C. & M. A. H. Nerenberg 1990. Fractal dimension: limit capacity or Hausdorff dimension? *American Journal of Physics* **58**(10), 986-8.

———. 1991. Comments on "deterministic chaos: the science and the fiction" by D. Ruelle. *Proceedings of the Royal Society of London A* **435**, 287-92.

Eubank, S. & J. D. Farmer 1990. An introduction to chaos and randomness. In *1989 lectures in complex systems*, E. Jen (ed.), 75-190. Redwood City, California: Addison-Wesley.

Farmer, J. D. 1982. Information dimension and the probabilistic structure of chaos. *Zeitschrif für Naturforschung*, Series A 37, 1304-325.

Farmer, J. D., E. Ott, J. A. Yorke 1983. The dimension of chaotic attractors. *Physica D* **7**, 153-80.

Farmer, J. D. & J. J. Sidorowich 1987. Predicting chaotic time series. *Physical Review Letters* **59**(8), 845-8.

———. 1988a. Exploiting chaos to predict the future and reduce noise. In *Evolution, learning, and cognition*, Y. C. Lee (ed.), 277-330. Singapore: World Scientific.

———. 1988b. Predicting chaotic dynamics. In *Dynamic patterns in complex systems*, J. A. S. Kelso, A. J. Mandell, M. F. Shlesinger (eds), 265-92. Singapore: World Scientific.

Feder, J. 1988. *Fractals*. New York: Plenum.

Feigenbaum, M. J. 1978. Quantitative universality for a class of nonlinear transformations. *Journal of Statistical Physics* **19**(1), 25-52.

———. 1979. The universal metric properties of nonlinear transformations. *Journal of Statistical Physics* **21**(6), 669-706.

———. 1980. Universal behavior in nonlinear systems. *Los Alamos Science* **1**, 4-27.

———. 1993. The transition to chaos. In *Chaos: the new science*, J. Holte (ed.), 45-53. Lanham, Maryland: University Press of America.

Firth, M. 1977. *Forecasting methods in business and management*. London: Edward Arnold.

Fisher, A. 1985. Chaos: the ultimate asymmetry. *Mosaic* **16**(1), 24-33.

Fokas, A. S. 1991. Some remarks on integrability. *Nonlinear Science Today* **1**(3), 6-11.

Ford, J. 1983. How random is a coin toss? *Physics Today* **36**(4), 40-47.

———. 1989. What is chaos, that we should be mindful of it? In *The new physics*, P. Davies (ed.), 348-72. Cambridge: Cambridge University Press.

Fowler, T. & D. Roach 1991. Dimensionality analysis of objects and series data. In *Nonlinear dynamics, chaos and fractals*, G. V. Middleton (ed.), 59-81. Short Course Notes 9, Geological Association of Canada, Toronto.

Fraedrich, K. 1986. Estimating the dimensions of weather and climate attractors. *Journal of the Atmospheric Sciences* **43**(5), 419-32.

Fraser, A. M. 1986. Using mutual information to estimate metric entropy. See Mayer-Kress (1986: 82-91).

———. 1989a. Information and entropy in strange attractors. *Institute of Electrical and Electronics Engineers (IEEE) Transactions on Information Theory* **35**(2), 245-62.

———. 1989b. Reconstructing attractors from scalar time series: a comparison of singular system and redundancy criteria. *Physica D* **34,** 391-04.

Fraser, A. M. & H. L. Swinney 1986. Independent coordinates for strange attractors from mutual information. *Physical Review A* **33**(2), 1134-40.

Froehling, H., J. P. Crutchfield, D. Farmer, N. H. Packard, R. Shaw 1981. On determining the dimension of chaotic flows. *Physica D* **3**, 605-617.

Fujisaka, H. 1983. Statistical dynamics generated by fluctuations of local Lyapunov exponents. *Progress of Theoretical Physics* **70**(5), 1264-75.

Gatlin, L. L. 1972. *Information theory and the living system*. New York: Columbia University Press.

Gershenfeld, N. 1988. An experimentalist's introduction to the observation of dynamical systems. In *Directions in chaos* (vol. 2), B-L. Hao (ed.), 310-84. Singapore: World Scientific.

———. 1993. Information in dynamics. In *Workshop on physics and computation, PhysComp '92*, D. Matzke (ed.), 276-80. Los Alamitos, California: Institute of Electrical and Electronics Engineers (IEEE) Computer Society Press.

Gershenfeld, N. A. & A. S. Weigend 1993. The future of time series: learning and understanding. See Weigend & Gershenfeld (1993: 1-70).

Glass, L. & M. C. Mackey 1988. *From clocks to chaos: the rhythms of life*. Princeton, New Jersey: Princeton University Press.

Glass, L. & D. T. Kaplan 1993. Complex dynamics in physiology and medicine. See Weigend & Gershenfeld (1993: 513-28).

Gleick, J. 1987. *Chaos: making a new science*. New York: Penguin.

Godfray, H. C. J. & S. P. Blythe 1990. Complex dynamics in multispecies communities. *Philosophical Transactions of the Royal Society of London A* **330,** 221-33.

Goldman, S. 1953. *Information theory*. New York: Prentice-Hall.

Goodchild, M. F. & D. M. Mark 1987. The fractal nature of geographic phenomena. *Annals of the Association of American Geographers* **77**(2), 265-78.

Grassberger, P. & I. Procaccia 1983a. Characterization of strange attractors. *Physical Review Letters* **50**(5), 346-9.

———. 1983b. Measuring the strangeness of strange attractors. *Physica D* **9**, 189-208.

———. 1984. Dimensions and entropies of strange attractors from a fluctuating dynamics approach. *Physica D* **13**, 34-54.

Grassberger, P., T. Schreiber, C. Schaffrath 1991. Nonlinear time sequence analysis. *International Journal of Bifurcation and Chaos* **1**(3), 521-47.

Grebogi, C., E. Ott, J. A. Yorke 1982. Chaotic attractors in crisis. *Physical Review Letters* **48**(22), 1507-510.

———. 1983. Crises, sudden changes in chaotic attractors, and transient chaos. *Physica D* **7**, 181-200.

———. 1987. Chaos, strange attractors, and fractal basin boundaries in nonlinear dynamics. *Science* **238**(30 October), 632-8.

Grebogi, C., E. Ott, S. Pelikan, J. A. Yorke 1984. Strange attractors that are not chaotic. *Physica D* **13**, 261-8.

Greenside, H. S., A. Wolf, J. Swift, T. Pignataro 1982. Impracticality of a box-counting algorithm for calculating the dimensionality of strange attractors. *Physical Review A* **25**(6), 3453-6.

Grossman, S. & S. Thomae 1977. Invariant distributions and stationary correlation functions of one-dimensional discrete processes. *Zeitschrift für Naturforschung* **32a**(12), 1353-63.

Härdle, W. 1991. *Smoothing techniques* (*with implementation in S*). Berlin: Springer.

Havstad, J. W. & C. L. Ehlers 1989. Attractor dimension of nonstationary dynamical systems from small datasets. *Physical Review A* **39**(2), 845-53.

Hayles, N. K. 1990. Self-reflexive metaphors in Maxwell's demon and Shannon's choice. In *Literature and science: theory and practice*, S. Peterfreund (ed.), 209-237. Boston: Northeastern University Press.

Hénon, M. 1976. A two-dimensional mapping with a strange attractor. *Communications in Mathematical Physics* **50**, 69-77.

Herivel, J. 1975. *Joseph Fourier: the man and the physicist*. Oxford: Oxford University Press.

Hoffman, B. 1975. *About vectors*. New York: Dover.

Hofstadter, D. R. 1981. Metamagical themas. *Scientific American* **245**(5), 22-43.

———. 1985. *Metamagical themas: questing for the essence of mind and pattern*. New York: Basic Books.

Holden, A. V. (ed.) 1986. *Chaos*. Princeton: Princeton University Press.

Holden, A. V. & M. A. Muhammad 1986. A graphical zoo of strange and peculiar attractors. See Holden (1986: 15-35).

Holzfuss, J. & G. Mayer-Kress 1986. An approach to error-estimation in the application of dimension algorithms. See Mayer-Kress (1986: 114-22).

Horgan, J. 1992. Claude E. Shannon. *Institute of Electrical and Electronics Engineers (IEEE) Spectrum* **29**(4), 72-5.

Hunt, B. R. & J. A. Yorke 1993. Maxwell on chaos. *Nonlinear Science Today* **3**(1), 1-4.

Izenman, A. J. 1991. Recent developments in nonparametric density estimation. *Journal of the American Statistical Association* **86**(413), 205-224.

Jackson, E. A. 1991. *Perspectives of nonlinear dynamics*, vol. 1. Cambridge: Cambridge University Press.

Jenkins, G. M. & D. G. Watts 1968. *Spectral analysis and its applications*. San Francisco: Holden-Day.

Jensen, R. V. 1987. Classical chaos. *American Scientist* **75**(March-April), 168-81.

Johnson, B. R. & S. K. Scott 1990. Period doubling and chaos during the oscillatory ignition of $CO+O_2$ reaction. *Journal of the Chemical Society, Faraday Transactions* **86**, 3701-705.

ac, M. 1983. What is random? *American Scientist* **71**(4), 405-406.

Kachigan, S. K. 1986. *Statistical analysis*. New York: Radius.

Kadanoff, L. P. 1983. Roads to chaos. *Physics Today* **36**(12), 46-53.

Kanasewich, E. R. 1981. *Time sequence analysis in geophysics*. Edmonton: University of Alberta Press.

Kaplan, D. T. 1994. Exceptional events as evidence for determinism. *Physica D* **73**, 38-48.

Kaplan, D. T. & L. Glass 1992. Direct test for determinism in a time series. *Physical Review Letters* **68**(4), 427-30.

Kennel, M. B., R. Brown, H. D. I. Abarbanel 1992. Determining embedding dimension for phase space reconstruction using a geometrical construction. *Physical Review A* **45**(6), 3403-411.

Kennel, M. B. & S. Isabelle 1992. Method to distinguish possible chaos from colored noise and to determine embedding parameters. *Physical Review A* **46**(6), 3111-18.

Kolmogorov, A. N. 1959, On entropy per unit time as a metric invariant in automorphisms. *Proceedings of the Academy of Sciences of the USSR* **124**(4), 754-5 (in Russian).

Kostelich, E. J. & J. A. Yorke 1990. Noise reduction: finding the simplest dynamical system consistent with the data. *Physica D* **41**, 183-96.

Kovalevskaya, S. V. 1978. *A Russian childhood* [translated and edited by Beatrice Stillman, with an analysis of Kovalevskaya's mathematics by P. Y. Kochina]. Berlin: Springer.

Kreyszig, E. 1962. *Advanced engineering mathematics*. New York: John Wiley.

Krippendorff, K. 1986. *Information theory: structural models for qualitative data*. Los Angeles: Sage.

Lall, U. & M. Mann 1995. The Great Salt Lake: a barometer of low-frequency climatic variability. *Water Resources Research* **31**(10), 2503-515.

Laskar, J. 1989. A numerical experiment on the chaotic behaviour of the solar system. *Nature* **338**(6212), 237-8.

Lauwerier, H. A. 1986a. One-dimensional iterative maps. See Holden (1986: 39-57).

———. 1986b. Two-dimensional iterative maps. See Holden (1986: 58-95).

Lewin, R. 1992. *Complexity—life at the edge of chaos*. New York: Macmillan.

Liapounoff, M. A. 1949. *Problème général de la stabilité du mouvement*. Princeton, New Jersey: Princeton University Press.

Lichtenberg, A. J. & M. A. Lieberman 1992. *Regular and chaotic dynamics*, 3rd edn. Berlin: Springer.

Liebert, W. & H. G. Schuster 1989. Proper choice of the time delay for the analysis of chaotic time series. *Physics Letters A* **142**(2-3), 107-111.

Liebert, W., K. Pawelzik, H. G. Schuster 1991. Optimal embeddings of chaotic attractors from topological considerations. *Europhysics Letters* **14**(6), 521-6.

Liebovitch, L. S. & T. Toth 1989. A fast algorithm to determine fractal dimensions by box counting. *Physics Letters A* **141**(8-9), 386-90.

Longley, P. A. & M. Batty 1989. Fractal measurement and line generalization. *Computers and Geosciences* **15**(2), 167-83.

Lorenz, E. N. 1963. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences* **20**(2), 130-41.

———. 1976. Nondeterministic theories of climatic change. *Quaternary Research* **6**, 495-506.

———. 1987. Deterministic and stochastic aspects of atmospheric dynamics. In *Irreversible phenomena and dynamical systems analysis in geosciences*, C. Nicolis & G. Nicolis (eds), 159-79. Norwell, Massachusetts: Reidel.

———. 1991. Dimension of weather and climate attractors. *Nature* **253**(19 September), 241-3.

———. 1993. *The essence of chaos*. Seattle: University of Washington Press.

Machlup, F. 1983. Semantic quirks in studies of information. In *The study of information: interdisciplinary messages*, F. Machlup & U. Mansfield (eds), 641-71. New York: John Wiley.

Makridakis, S., S. C. Wheelwright, V. E. McGee 1983. *Forecasting: methods and applications*, 2nd edn. New York: John Wiley.

Mandelbrot, B. B. 1967. How long is the coast of Britain: statistical self-similarity and fractional dimension. *Science* **156**(5 May), 636-8.

———. 1983. *The fractal geometry of nature*. New York: W. H. Freeman.

Mañé, R. 1981. On the dimension of the compact invariant sets of certain nonlinear maps. In *Dynamical systems and turbulence*, D. A. Rand & L-S. Young (eds), 230-42 [Lecture Notes in Mathematics 898]. Berlin: Springer.

Martinerie, J. M., A. M. Albano, A. I. Mees, P. E. Rapp 1992. Mutual information, strange attractors, and the optimal estimation of dimension. *Physical Review A* **45**(10), 7058-7064.

May, R. M. 1974. Biological populations with nonoverlapping generations: stable points, stable cycles, and chaos. *Science* **186**(15 November), 645-7.

———. 1976. Simple mathematical models with very complicated dynamics. *Nature* **261**(June 10), 459-67.

———. 1987. Chaos and the dynamics of biological populations. *Proceedings of the Royal Society of London A* **413**, 27-44.

May, R. M. & G. F. Oster 1976. Bifurcations and dynamic complexity in simple ecological models. *The American Naturalist* **110**(974), 573-99.

Mayer-Kress, G. (ed.) 1986. *Dimensions and entropies in chaotic systems*. Berlin: Springer.

———. 1987. Application of dimension algorithms to experimental chaos. In *Directions in chaos* (vol. 1), B-L. Hao (ed.), 122-47. Singapore: World Scientific.

McAuliffe, K. 1990. Get smart: controlling chaos. *OMNI* **12**(5), 43-8, 86-92.

Mende, W., H. Herzel, K. Wermke 1990. Bifurcations and chaos in newborn infant cries. *Physics Letters A* **145**(8-9), 418-24.

Middleton, G. V. 1995. Appendix I: chaossary—a short glossary of chaos. In *Nonlinear dynamics and fractals-new numerical techniques for sedimentary data*, G. V. Middleton. R. E. Plotnick, D. M. Rubin (eds), 133-58. Short Course 36, Society for Sedimentary Geology (SEPM), Tulsa, Oklahoma.

Moon, F. C. 1992. *Chaotic and fractal dynamics*. New York: John Wiley.

Morrison, F. 1988. On chaos. EOS, *Transactions of the American Geophysical Union* 69(25), 668-9.

Mullin, T. 1993. A dynamical systems approach to time-series analysis. In *The nature of chaos*, T. Mullin (ed.), 23-50. New York: Oxford University Press.

Mundt, M. D., W. B. Maguire II, R. R. P. Chase 1991. Chaos in the sunspot cycle: analysis and prediction. *Journal of Geophysical Research* **96**(A2), 1705-716.

Murray, J. D. 1991. Mathematics—biology—nonlinearity. *Nonlinear Science Today* **1**(3), 1-5.

Nerenberg, M. A. H. & C. Essex 1990. Correlation dimension and systematic geometric effects. *Physical Review A* **42**(12), 7065-7074.

Nese, J. M. 1989. Quantifying local predictability in phase space. *Physica D* **35**, 237-50.

Nicolis, J. S., G. Mayer-Kress, G. Haubs 1983. Non-uniform chaotic dynamics with implications to information processing. *Zeitschrift für Naturforschung* **38a**, 1157-69.

Nychka, D., S. Ellner, A. R. Gallant, D. McCaffrey 1992. Finding chaos in noisy systems. *Journal of the Royal Statistical Society*, Series B **54**(2), 399-426.

Olsen, L. F. & H. Degn 1985. Chaos in biological systems. *Quarterly Review of Biophysics* **18**(2), 165-225.

Orford, J. D. & W. B. Whalley 1987. The quantitative description of highly irregular sedimentary particles: the use of the fractal dimension. In *Clastic particles-scanning electron microscopy and shape analysis of sedimentary and volcanic clasts*, J. R. Marshall (ed.), 267-80. New York: Van Nostrand Reinhold.

Osborne, A. R. & A. Provenzale 1989. Finite correlation dimension for stochastic systems with power-law spectra. *Physica D* **35,** 357-81.

Oseledec, V. 1. 1968. A multiplicative ergodic theorem: Ljapunov characteristic numbers for dynamical systems. *Transactions of the Moscow Mathematical Society* **19,** 197-231 (in English).

Ottino, J. M., F. J. Muzzio, M. Tjahjadi, J. G. Franjione, S. C. Jana, H. A. Kusch 1992. Chaos, symmetry, and self-similarity: exploiting order and disorder in mixing processes. *Science* **257**(7 August), 754-60.

Packard, N. H., J. P. Crutchfield, J. D. Farmer, R. S. Shaw 1980. Geometry from a time series. *Physical Review Letters* **45**(9), 712-16.

Paluš, M. 1993. Identifying and quantifying chaos by using information-theoretic functionals. See Weigend & Gershenfeld (1993: 387-413).

Parker, T. S. & L. O. Chua 1989. *Practical numerical algorithms for chaotic systems*. Berlin: Springer.

Paulos, J. A. 1990. *Innumeracy—mathematical illiteracy and its consequences*. New York: Vintage.

Peitgen, H-O. & D. Saupe (eds) 1988. *The science of fractal images*. Berlin: Springer.

Peitgen, H-O., H. Jürgens, D. Saupe 1992. *Chaos and fractals: new frontiers of science*. Berlin: Springer.

Percival, I. 1989. Chaos: a science for the real world. *New Scientist* **124**(1687), 42-7.

Peters, E. E. 1991. *Chaos and order in the capital markets*. New York: John Wiley.

Peterson, I. 1988. In the shadows of chaos. *Science News* **134**(3 December), 360-61.

Pielou, E. C. 1969. *An introduction to mathematical ecology*. New York: John Wiley.

Pineda, F. J. & J. C. Sommerer 1993. Estimating generalized dimensions and choosing time delays: a fast algorithm. See Weigend & Gershenfeld (1993: 367-85).

Pippard, B. 1982. Instability and chaos: physical models of everyday life. *Interdisciplinary Science Reviews* 7(2), 92-101.

Plotnick, R. E. & K. Prestegaard 1993. Fractal analysis of geologic time series. In *Fractals in geography*, N. S. Lam & L. de Cola (eds), 193-210. Englewood Cliffs, New Jersey: Prentice-Hall.

Pomeau, Y. & P. Manneville 1980. Intermittent transition to turbulence in dissipative dynamical systems. *Communications in Mathematical Physics* **74**, 189-97.

Pool, R. 1989a. Is it chaos, or is it just noise? *Science* **243**(6 January), 25-8.

———. 1989b. Ecologists flirt with chaos. *Science* **243**(20 January), 310-13.

———. 1989c. Is it healthy to be chaotic? *Science* **243**(3 February), 604-607.

———. 1989d. Quantum chaos: enigma wrapped in a mystery. *Science* **243**(17 February), 893-5.

———. 1989e. Is something strange about the weather? *Science* **243**(10 March), 1290-93.

———. 1989f. Chaos theory: how big an advance? *Science* **243**(7 July), 26-8.

———. 1990a. Fractal fracas. *Science* **249**(27 July), 363-4.

———. 1990b. Putting chaos to work. *Science* **250**(2 November), 626-7.

Prichard, D. 1994. The correlation dimension of differenced data. *Physics Letters A* **191**(3-4), 245-50.

Prichard, D. & C. P. Price 1992. Spurious dimension estimates from time series of geomagnetic indices. *Geophysical Research Letters* **19**(15), 1623-6.

———. 1993. Is the AE index the result of nonlinear dynamics? *Geophysical Research Letters* **20**(24), 2817-20.

Priestley, M. B. 1981. *Spectral analysis and time series* [2 volumes]. London: Academic Press.

Provenzale, A., L. A. Smith, R. Vio, G. Murante 1992. Distinguishing between low-dimensional dynamics and randomness in measured time series. *Physica D* **58**, 31-49.

Ramirez, R. W. 1985. *The EFT fundamentals and concepts*. Englewood Cliffs, New Jersey: Prentice-Hall.

Ramsey, J. B. & H-J. Yuan 1989. Bias and error bars in dimension calculations and their evaluation in some simple models. *Physics Letters A* **134**(5), 287-97.

Rapp, P. E. 1992. Dynamical analysis of biological systems: a case for guarded optimism. In National Institutes of Health, workshop on "The head and heart of chaos: nonlinear dynamics in biological systems," Executive Report (15-16 June), 43-6.

Rapp, P. E., T. R. Bashore, J. M. Martinerie, A. M. Albano, I. D. Zimmerman, A. I. Mees 1989. Dynamics of brain electrical activity. *Brain Topography* **2**(1-2), 99-118.

Rapp, P. E., A. M. Albano, T. I. Schmah, L. A. Farwell 1993. Filtered noise can mimic low-dimensional chaotic attractors. *Physical Review E* **47**(4), 2289-97.

Rasband, N. S. 1990. *Chaotic dynamics of nonlinear systems*. New York: John Wiley.

Rayner, J. N. 1971. *An introduction to spectral analysis*. London: Pion.

Read, P. L., M. J. Bell, D. W. Johnson, R. M. Small 1992. Quasi-periodic and chaotic flow regimes in a thermally driven, rotating fluid annulus. *Journal of Fluid Mechanics* **238**, 599-632.

Richardson, L. F. 1961. The problem of contiguity: an appendix to *Statistics of deadly quarrels*. *General Systems Yearbook* **6**(1), 139-87.

Robinson, A. L. 1982. Physicists try to find order in chaos. *Science* **218**(5 November), 554-6.

Rössler, O. E. 1976. An equation for continuous chaos. *Physics Letters* **57A**(5), 397-8.

———. 1979. An equation for hyperchaos. *Physics Letters* **71A**, 155-7.

Rothman, T. 1991. *A physicist on Madison avenue*. Princeton, New Jersey: Princeton University Press.

Rowlands, G. & J. C. Sprott 1992. Extraction of dynamical equations from chaotic data. *Physica D* **58**, 251-9.

Ruelle, D. 1990. Deterministic chaos: the science and the fiction. *Proceedings of the Royal Society of London A* **427**, 241-8.

———. 1991. *Chance and chaos.* Princeton, New Jersey: Princeton University Press.

———. 1994. Where can one hope to profitably apply the ideas of chaos? *Physics Today* **47**(7), 24-30.

Salas, J. D., J. W. Delleur, V. Yevjevich, W. L. Lane 1980. *Applied modeling of hydrologic time series*. Littleton, Colorado: Water Resources Publications.

Sangoyomi, T. B., U. Lall, H. D. I. Abarbanel 1996. Nonlinear dynamics of the Great Salt Lake: dimension estimation. *Water Resources Research* **32**(1), 149-59.

Sano, M. & Y. Sawada 1985. Measurement of the Lyapunov spectrum from a chaotic time series. *Physical Review Letters* **55**(10), 1082-1085.

Sauer, T.,J. A. Yorke, M. Casdagli 1991. Embedology. *Journal of Statistical Physics* **65**(3/4). 579-616.

Saupe, D. 1988. Algorithms for random fractals. See Peitgen & Saupe (1988: 71-136).

Schaffer, W. M. & M. Kot 1985. Do strange attractors govern ecological systems? *BioScience* **35**(6), 342-50.

———. 1986. Chaos in ecological systems: the coals that Newcastle forgot. *Trends in Ecology and Evolution* **1**(3), 58-63.

Schaffer, W. M., S. Ellner, M. Kot 1986. Effects of noise on some dynamical models in ecology. *Journal of Mathematical Biology* **24**, 479-523.

Schaffer, W. M., L. F. Olsen, G. L. Truty, S. L. Fulmer 1990. The case for chaos in childhood epidemics. In *The ubiquity of chaos*, S. Krasner (ed.), 138-66. Washington DC: American Association for the Advancement of Science.

Schaffer, W. M., G. L. Truty, S. L. Fulmer 1992. *Dynamical software professional-user's manual and introduction to chaotic systems*, vol. 2 (version 11.2). Tucson: Dynamical Systems [distributed by Academic Computing Specialists, Salt Lake City, Utah].

Schneider, E. D. 1988. Thermodynamics, ecological succession, and natural selection: a common thread. In *Entropy, information, and evolution: new perspectives on physical and biological evolution*, B. H. Weber, D. J. Depew, J. D. Smith (eds), 107-138. Cambridge, Massachusetts: Massachusetts Institute of Technology Press.

Schuster, H. G. 1988. *Deterministic chaos: an introduction*, 2nd edn. Weinheim, Germany: VCH.

Scott, D. W. 1992. *Multivariate density estimation*. New York: John Wiley.

Seydel, R. 1988. *From equilibrium to chaos*. Amsterdam: Elsevier.

Shannon, C. E. & W. Weaver 1949. *The mathematical theory of communication*. Urbana-Champaign: University of Illinois Press.

Shaw, R. 1984. *The dripping faucet as a model chaotic system*. Santa Cruz, California: Aerial Press.

Shinbrot, T. 1993. Chaos: unpredictable yet controllable? *Nonlinear Science Today* **3**(2), 1-8.

Shinbrot, T., C. Grebogi, E. Ott, J. A. Yorke 1993. Using small perturbations to control chaos. *Nature* **363**(3 June), 411-17.

Shumway, R. H. 1988. *Applied statistical time series analysis*. Englewood Cliffs, New Jersey: Prentice-Hall.

Silverman, B. W. 1986. *Density estimation for statistics and data analysis*. London: Chapman & Hall.

Sinai, Ya. 1959. On the concept of entropy in a dynamical system. *Proceedings of the Academy of Sciences of the USSR* **124**(4), 768-71 (in Russian).

Smith, L. A. 1988. Intrinsic limits of dimension calculations. *Physics Letters A* **133**(6), 283-8.

Stewart, H. B. 1984. *The geometry of chaos*. Brookhaven Lecture Series 209, Report BNL-51892, Brookhaven National Laboratory.

Stewart, I. 1989a. *Does God play dice?: the mathematics of chaos*. Cambridge, Massachusetts: Basil Blackwell.

———. 1989b. Portraits of chaos. *New Scientist* **124**(1689), 42-7.

Stoop, R. & J. Parisi 1991. Calculation of Lyapunov exponents avoiding spurious elements. *Physica D* 50, 89-94.

Sussman, G. J. & J. Wisdom 1992. Chaotic evolution of the solar system. *Science* **257**(3 July), 56-62.

Swinney, H. L. 1986. Experimental observations of order and chaos. In *Nonlinear dynamics and chaos: geometrical methods for engineers and scientists*, J. M. T. Thompson & H. B. Stewart (eds), 332-49. New York: John Wiley.

Takens, F. 1981. Detecting strange attractors in turbulence. In *Dynamical systems and turbulence*, D. A. Rand & L-S. Young (eds), 366-81 [Lecture Notes in Mathematics 898]. Berlin: Springer.

———. 1985. On the numerical determination of the dimension of an attractor. In *Dynamical systems and bifurcations*, B. L. J. Braaksma, H. W. Broer, F. Takens (eds), 99-106. Berlin: Springer.

Tashman, L. J. & K. R. Lamborn 1979. *The ways and means of statistics*. New York: Harcourt Brace Jovanovich.

Theiler, J. 1986. Spurious dimension from correlation algorithms applied to limited time-series data. *Physical Review A* **34**(3), 2427-32.

———. 1987. Efficient algorithm for estimating the correlation dimension from a set of discrete points. *Physical Review A* **36**(9), 4456-62.

———. 1990a. Estimating fractal dimension. *Journal of the Optical Society of America A* **7**(6), 1055-1073.

———. 1990b. Statistical precision of dimension estimators. *Physical Review A* **41**(6), 3038- 3051.

Theiler, J., B. Galdrikian, A. Longtin, S. Eubank, J. D. Farmer 1992. Using surrogate data to detect nonlinearity in time series. In *Nonlinear modeling and forecasting*, M. Casdagli & S. Eubank (eds), 163-88. Redwood City, California: Addison-Wesley.

Theiler, J. & S. Eubank 1993. Don't bleach chaotic data. *Chaos* **3**(4), 771-82.

Theiler, J., P. S. Linsay, D. M. Rubin 1993. Detecting nonlinearity in data with long coherence times. See Weigend & Gershenfeld (1993: 429-55).

Thompson, J. M. T. & H. B. Stewart 1986. *Nonlinear dynamics and chaos: geometrical methods for engineers and scientists*. New York: John Wiley.

Tredicce, J. R. & N. B. Abraham 1990. Experimental measurements to identify and/or characterize chaotic signals. In *Lasers and quantum optics*, L. M. Narducci, E. J. Quel, J. R. Tredicce (eds), 148-98. Singapore: World Scientific.

Tribus, M. & E. C. McIrvine 1971. Energy and information. *Scientific American* **224**(3), 179-88.

Troutman, B. M. & G. P. Williams 1987. Fitting straight lines in the Earth sciences. In *Use and abuse of statistical methods in the earth sciences*, W. B. Size (ed.), 107-128. New York: Oxford.

Tsonis, A. A. 1989. Chaos and unpredictability of weather. *Weather* **44**, 258-63.

Tsonis, A. A., G. N. Triantafyllou, J. B. Elsner, J. J. Holdzkom II, A. D. Kirwan Jr 1994. An investigation of the ability of nonlinear methods to infer dynamics from observables. *Bulletin of the American Meteorological Society* **75**(9), 1623-633.

Tufillaro, N. B., T. Abbott, J. Reilly 1992. *An experimental approach to nonlinear dynamics and chaos*. Redwood City, California: Addison-Wesley.

Vassilicos, J. C., A. Demos, F. Tata 1993. No evidence of chaos but some evidence of multi-fractals in the foreign exchange and the stock markets. *In Applications of fractals and chaos*, A. J. Crilly, R. A. Earnshaw, H. Jones (eds), 249-65. Berlin: Springer.

Vautard, R. & M. Ghil 1989. Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series. *Physica D* **35**, 395-424.

Voss, R. F. 1988. Fractals in nature: from characterization to simulation. See Peitgen & Saupe (1988: 21-70).

Waldrop, M. M. 1992. *Complexity: the emerging science at the edge of order and chaos*. New York: Simon & Schuster.

Wastler, T. A. 1963. *Application of spectral analysis to stream and estuary field surveys, 1. individual power spectra*. Public Health Service Publication 999-WP-7, us Department of Health, Education, and Welfare, Cincinnati.

Wayland, R., D. Bromley, D. Pickett, A. Passamante 1993. Recognizing determinism in a time series. *Physical Review Letters* **70**(5), 580-82.

Wegman, E. J. 1988. On randomness, determinism and computability. *Journal of Statistical Planning and Inference* **20**, 279-94.

Weigend, A. S. 1994. Paradigm change in prediction. *Philosophical Transactions of the Royal Society of London A* **348**(1688), 405-420.

Weigend, A. S. & N. A. Gershenfeld (eds) 1993. *Time-series prediction: forecasting the future and understanding the past*. Reading, Massachusetts: Addison-Wesley.

West, B. J. 1990. *Fractal physiology and chaos in medicine*. Singapore: World Scientific.

Whitney, H. 1936. Differentiable manifolds. *Annals of Mathematics* **37**(3), 645-80.

Wilkes, M. V. 1966. *A short introduction to numerical analysis*. Cambridge: Cambridge University Press.

Wolf, A. 1983. Simplicity and universality in the transition to chaos. *Nature* **305**(15 September), 182-3.

———. 1986. Quantifying chaos with Lyapunov exponents. In *Chaos*, A. V. Holden (ed.). 273-90. Princeton, New Jersey: Princeton University Press.

Wolf, A., J. B. Swift, H. L. Swinney, J. A. Vastano 1985. Determining Lyapunov exponents from a time series. *Physica D* **16**, 285-317.

Wonnacott, T. H. & R. J. Wonnacott 1984. *Introductory statistics for business and economics*, 3rd edn. New York: John Wiley.

Wright, J. 1984. Method for calculating a Lyapunov exponent. *Physical Review A* **29**(5), 2924-7.

Yorke, J. A., C. Grebogi, E. Ott, L. Tedeschini-Lalli 1985. Scaling behavior of windows in dissipative dynamical systems. *Physical Review Letters* **54**(11), 1095-1098.

Young, L-S. 1983. Entropy, Lyapunov exponents, and Hausdorff dimension in differentiable dynamical systems. *Institute of Electrical and Electronics Engineers (IEEE) Transactions on Circuits and Systems* **CAS-30**(8), 599-607.

Yule, G. U. 1927. On a method of investigating periodicities in disturbed series, with special reference to Wolfe's sunspot numbers. *Philosophical Transactions of the Royal Society of London A* **226**, 267-98.

Zeng, X., R. Eykholt, R. A. Pielke 1991. Estimating the Lyapunov-exponent spectrum from short time series of low precision. *Physical Review Letters* **66**(25), 3229-32.

Zeng, X., R. A. Pielke, R. Eykholt 1992. Extracting Lyapunov exponents from short time series of low precision. *Modern Physics Letters B* **6**(2), 55-75.

———. 1993. Chaos theory and its applications to the atmosphere. *Bulletin of the American Meteorological Society* **74**(4), 631-44.

SELECTED FURTHER READING

Baker, G. L. & J. P. Gollub 1990. *Chaotic dynamics: an introduction*. Cambridge: Cambridge University Press.

Devaney, R. L. 1990. *Chaos, fractals, and dynamics: computer experiments in mathematics*. Menlo Park, California: Addison-Wesley.

———. 1992. *A first course in chaotic dynamical systems: theory and experiment*. Reading, Massachusetts: Addison-Wesley.

Hall, N. (ed.) 1991. *Exploring chaos-a guide to the new science of disorder*. New York: Norton.

Hao, B-L. (ed.) 1984. *Chaos.* Singapore: World Scientific.

Hayles, N. K. (ed.) 1991. *Chaos and order-complex dynamics in literature and science*. Chicago: University of Chicago Press.

Hilborn, R. C. 1994. *Chaos and nonlinear dynamics*. New York: Oxford University Press.

Marek, M. & I. Schreiber 1991. *Chaotic behavior of deterministic dissipative systems*. Cambridge: Cambridge University Press.

Ott, E. 1993. *Chaos in dynamical systems*. Cambridge: Cambridge University Press.

Peterson, I. 1988. *The mathematical tourist-snapshots of modern mathematics*. New York: Freeman.

Strogatz, S. H. 1994. *Nonlinear dynamics and chaos*. Reading, Massachusetts: Addison-Wesley.

Tsonis, A. A. 1992. *Chaos-from theory to applications*. New York: Plenum.