CHAPTER 4

The shape of change

Children change in thousands of ways, every day. There are changes in body size, attitudes, food preferences, style of dress, sense of humor, and a host of mental abilities. Sometimes change is slow and gradual. Sometimes it happens so fast that the child is a different person from one hour to the next. And sometimes consistent progress is followed by retreat. Out of all the myriad patterns of change that could be studied in a human child, the field of child development has traditionally focussed on a relatively narrow subset. Simply put, some forms of change are more interesting than others, because they present an important challenge to theories of development and (above all) to the peripatetic issue of nature, nurture and their interaction.

In this section, we want to unpack the term "interesting" into a set of primitives, presenting a taxonomy of growth patterns and an analysis of the theoretical assumptions that accompany each one. This taxonomy has a very specific function in a volume on connectionism and developmental psychology, because connectionism offers alternative explanations for some of the most interesting and exotic patterns of growth on the developmental landscape.

To some extent, the interest value of a particular growth pattern comes from its content rather than its shape. For example, with some exceptions (e.g., Gesell, 1929; Rochat, 1984; Thelen, 1986, 1994), developmentalists have not focussed on physical growth or motor development—except, perhaps, as metaphors for mental and/or behavioral change (e.g., Lenneberg, 1967, on proposed parallels between motor and language milestones; or Chomsky, 1980, who likens the language faculty to a mental organ). We will not dwell here on the inherent interest value of specific content domains, except to note that many of our own preferred examples

come from the perceptual, linguistic, cognitive and social domains that motivate other developmental theories. Our focus here will be on the shape of change *within* these domains, along three nested dimensions: *linearity* (i.e., linear vs. nonlinear), *direction* (monotonic vs. nonmonotonic) and *continuity* (continuous vs. discontinuous). It is our contention that most developmentalists adopt an implicit theory of these three dimensions, a theory in which the least interesting phenomena lie at one end (change that is continuous, monotonic and linear) while the most interesting phenomena lie at the other extreme (change that is nonlinear, nonmonotonic and discontinuous).

In this context, the word "interesting" refers to the nature, number and transparency of the causes that must be invoked to explain each form of change. In particular, it is usually assumed that phenomena at the "uninteresting" linear end can be explained by a variety of simple and transparent causes—so many that their explanation poses no real theoretical challenge (i.e., these domains are unconstrained). By contrast, phenomena at the "interesting" nonlinear end of the spectrum have proven much more difficult to explain, requiring theorists to seek complex and highly-constrained sources of causation that often bear a very indirect and non-obvious relationship to the final product. In the most interesting cases, it seems as though a different set of causal factors may be required at different points along the developmental trajectory. Hence these phenomena provide a greater challenge to developmental theory, and their resolution can be viewed as a substantial victory. ***One of our goals here is to illustrate how the number of causes or mechanisms required to explain complex patterns of change can be reduced within a connectionist framework.*** In particular, there are cases in which several distinct causal mechanisms can be subsumed by a smaller set of explanatory principles. Of course not everyone finds this kind of unification aesthetically pleasing (*"De gustibus non est disputandum"*). But for those who value parsimony and theoretical austerity, this should be viewed as an advance.

Dynamical systems theory will figure importantly in what follows, and for the sake of completeness, we wish to hail the recent trend, both in developmental theory (e.g., Smith & Thelen, 1993;

Thelen & Smith, 1994; van Geert, 1994) and in cognitive science more generally (e.g., Port & van Gelder, 1995) to use dynamical systems approach in understanding behavior. Some of this work (Smith & Thelen, for instance) focuses on motor development, whereas other work (for example, the Port & Van Gelder collection) looks at cognitive and perceptual processes as well. This approach has demonstrated that it is often possible to capture complex patterns with models that have relatively few degrees of freedom.

We now consider six different forms of change, discussing their shapes and the formal models which give rise to them.

1. Linear change. Figure 4.1 illustrates simple linear patterns of change over time, including cases of linear increase (Figure 4.1a) and linear decrease (Figure 4.1b). Hypothetical examples of Figure 4.1a might include the gradual growth of vocabulary that is usually observed from 18 years of age to adulthood (Bates & Carnevale, 1993). Similar claims could be made for any domain of cultural knowledge that depends (or so it is assumed) on accumulative experience, and/or perceptual-motor skills that are characterized by incremental gains. Hypothetical examples of Figure 4.1b might include a gradual loss of speed in motor skills after age 30, or gradual loss of accuracy on memory tasks after age 50 in normal adults. Other examples (inverting the usual developmental convention of plotting upward growth) might include a drop-off in the errors observed on spelling tests for grade school children from 7–12 years of age.

Investigators are rarely surprised to find such patterns of linear change, because it is assumed that (1) they are quite common, and (2) they can be easily explained by a host of simple additive mechanisms (experiential or genetic). Although the second point is undeniably true, the first is questionable. In fact, there are good reasons to believe that truly linear patterns of gain or loss are relatively rare in the study of biology or behavior, and many putative cases of linear change be artifacts of sampling. For example, changes in shoe size may fit a linear pattern if we restrict our attention to children between 5–11 years of age. However, we would see some compelling nonlinear increases in shoe size if the age range were expanded

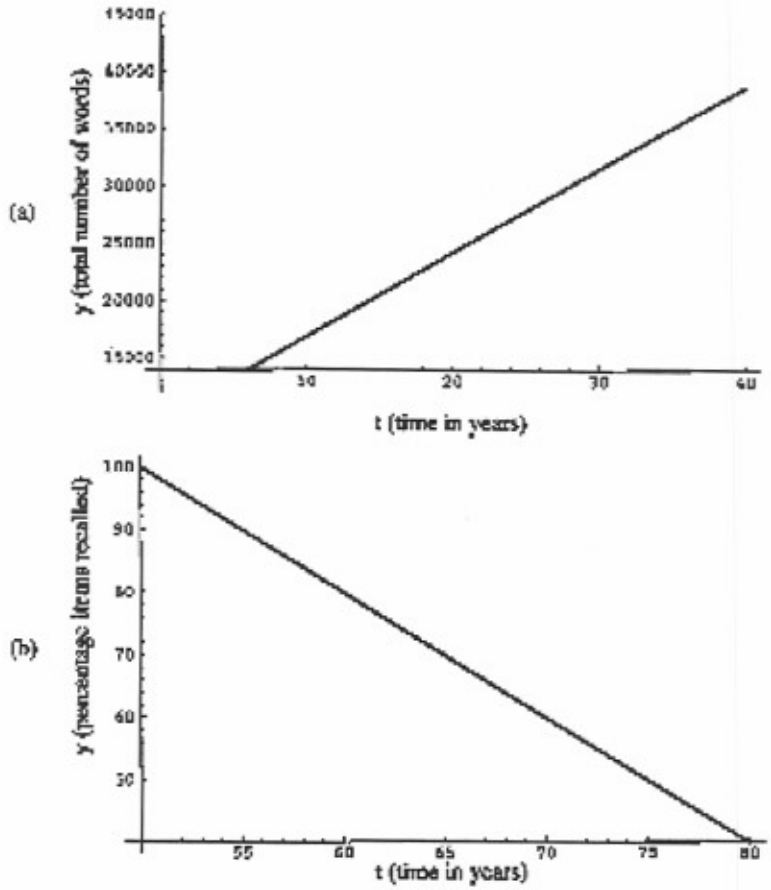


FIGURE 4.1 Hypothetical graphs illustrating linear outcomes (vertical axes) as a function of time (horizontal axes).

to include adolescence. Another source of artifactual linearity comes from summing across individual patterns of growth in group studies. For example, the burst in shoe size displayed by many individual children around adolescence might be lost if we were to graph cross-sectional data from 9–16 years of age in a group of 1,000 children from different cultural, socioeconomic and ethnic groups. To be sure, these data would fit a linear pattern (that is, we have not

made a mistake); however, it would be a mistake to assume that this linear pattern reflects a linear (incremental, accumulative) mechanism.

This brings us to a critical distinction between cause and effect, i.e., between the mechanisms responsible for growth and the patterns of growth that are the output of such a mechanism. Invariably, we are forced to infer the former from the latter—and that is where we often go astray. To make this point, we need to distinguish between two different aspects of change: the **outcome function** (i.e., the state of the system as a function of time) and the **dynamic equation** (i.e., the rule governing the rate of change of the system). These two aspects of change have equally important but logically distinct implications for theories of development, particularly when we move on to cases of nonlinear outcomes, with or without nonlinear dynamics.

By definition, a relationship between two variables is linear if it can be fit by a formula of the type

$$y = at + b \qquad \text{(EQ 4.1)}$$

where y and t are variables, and a and b are constants that are independent of y and t. Any relationship that cannot be fit by a formula of this kind is, by definition, nonlinear. The outcomes (the y's in Equation 4.1) illustrated in Figure 4.1 can each be described by a linear equation (which defines the ascending slope in outcomes over time in Figure 4.1a, and the descending slope in outcomes over time in Figure 4.1b).

Let us make the example in Figure 4.1a more concrete. We will assume that the t in Equation 4.1 refers to time. The y stands for quantities of some measurable behavior. In the specific relationship illustrated in Figure 4.1a, t (on the horizontal axis) stands for an age range from 6–40 years and y (on the vertical axis) represents the estimated number of words in the vocabulary of the average English speaker, from 14,000 words at age 6 (Templin, 1957; see also Carey, 1982) to an estimated average of 40,000 words at age 40 (McCarthy, 1954). What of the constants a and b in Equation 4.1? The constant a represents the rate of change (which is why we want to multiply it by the age, t). The constant b is the starting number of

words (which is why we add that to the other term). It turns out that values of 730 words per year (approximately 2 per day) for *a* and 9620 words for *b* (necessary to produce the estimated value of 14,000 words at age 6) work well.

Although often it is the outcome function one is interested in, it is also often useful to be able to focus on the way the outcome changes over time. In the example above we say that the rate of change is represented by *a*; in the example, the rate remains constant. We can describe the rate of change using the following linear dynamic equation:

$$dy/dt = a \qquad \text{(EQ 4.2)}$$

(The term *dy/dt* is a special notation, called a derivative, which in this case does not refer to division. Instead, it is to be read as "the rate of change in *y* per unit of change in *t*." The *d*'s here have no other meaning). As we said, the symbol on the right-hand side of the equation is constant, so if we graphed the rate of change over time, we would expect to find the horizontal line shown in Figure 4.2. This simple example illustrates an important thing to

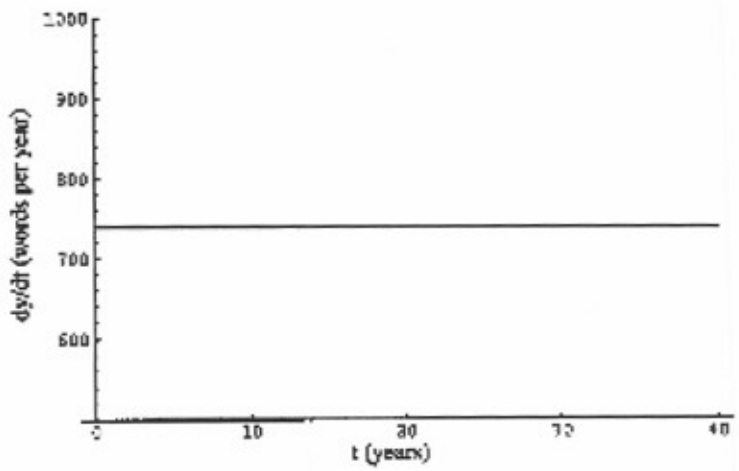


FIGURE 4.2 Dynamic equation underlying the outcome graphed in Figure 4.1a. Because the rate of change (dy/dt) is constant over time, the graph is flat.

be due to some internal clock of the individual or be external and just related to the rate at which the individual is exposed to new information. We could represent such prescribed or externally controlled changes in the rule by including an explicit dependence of f on time. The case in which there is no explicit time dependence is called autonomous. In that case f depends only on y, in other words, the rate of learning at any given time depends only on the state of the system at that time. This may be appropriate if the influx of new words is a constant in time.

In the present context, we could consider the system to be the language learning system of the individual, and say we represent the state of the system by the number of learned words, y. We could postulate that the rate of learning words is given by some efficiency, ε, times g, the number of words presented to the child per unit time. For example, if the number of new words presented to the person in a year is 1000 but the person's learning system has an efficiency of 20% then the number of words learned would be $\varepsilon g = 200$. To simplify matters, let us also assume that the number of new words presented per unit time is constant. Thus we can represent the evolution of the system by

$$\frac{dy}{dt} = \varepsilon(y, t)g \qquad \text{(EQ 4.4)}$$

Such an equation is capable of representing as complicated an evolution as we like simply by specifying a complicated evolution of the efficiency with time. Thus changes in rates of learning could simply be prescribed to occur at different times. On the other hand, it may be that the efficiency of learning evolves mainly due to the internal dynamics of the learning system and is not governed or regulated, at least in any critical way, by outside processes. Such a model would be written as

$$\frac{dy}{dt} = \varepsilon(y)g \qquad \text{(EQ 4.5)}$$

That is, the efficiency of learning here depends explicitly only on the state of the system at any given time. As we shall see, this

system too can give rise to complicated behavior, but now any changes in the efficiency of the system can only arise out of the natural evolution of the system subjected to a constant input of data, as opposed to the case where the changes are explicitly prescribed.

In what follows, we will consider some simple models for $e(y)$ that can give interesting behavior for y. Let us turn now to some more interesting cases.

2. Nonlinear monotonic change with linear dynamics. Now we can move into patterns that have played a more important role in developmental theory. We will start with patterns of growth that follow an "interesting" nonlinear trajectory, patterns that have inspired complex explanations involving two or more distinct causal mechanisms. As we shall see, however, the apparent complexity of these nonlinear outcome patterns is illusory, because they can be parsimoniously explained by a simple linear dynamics. (And this is why we made the distinction above between outcome equations and dynamical equations.)

Figure 4.3 illustrates the nonlinear relationship between vocab-

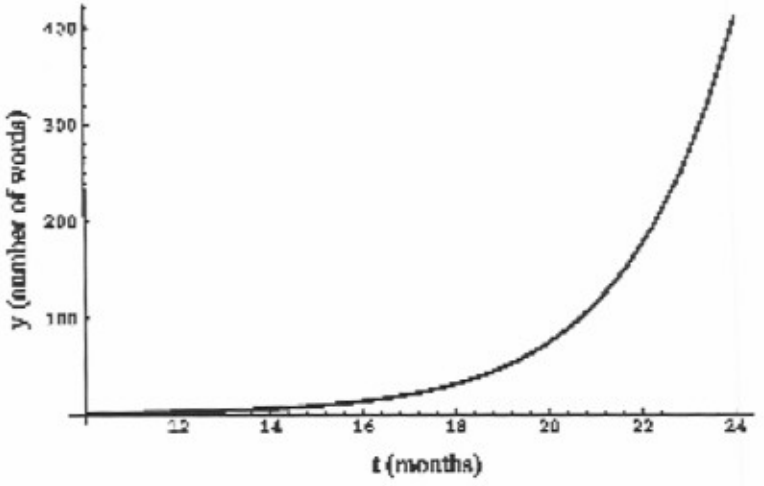


FIGURE 4.3 Hypothetical relationship between a child's expressive vocabulary and age.

ulary and age that has been observed in the earliest stages of language development (from Bates & Carnevale, 1993). In this case, t represents age from 10 to 24 months, and y represents the number of words in the child's expressive vocabulary. This graph is an idealization, but it is patterned after vocabulary growth functions that have been observed in diary studies of real live human children (e.g., Dromi, 1987; Nelson, 1973).

In contrast with the graph of adult vocabulary growth in Figure 4.1a, the infant data graphed in Figure 4.3 illustrate a nonlinear outcome. It appears from a cursory examination of this pattern that there is a marked acceleration somewhere around the 50-word level in the rate at which new words are added (occurring here around 20 months). For example, the child learns only 24 new words between 10 and 14 months, but she learns 328 new words between 20 and 24 months. This nonlinear pattern can be described by the nonlinear function

$$y = y_0 e^{b(t - t_0)} \qquad \text{(EQ 4.6)}$$

Here y_0 (=1 in this case) is the number of words the child knows at 10 months of age (Fenson et al., 1994). The constant b is called the exponential growth rate (which in this case is 43% per month), and e^x is the exponential function. Actually all this function e^x means is that the constant e (e is a constant frequently used in mathematics and its value is approximately 2.718) is raised to the power x. For example, e^2 is approximately 2.718^2 or approximately 7.388, etc.

"Burst" patterns like this one have been reported many times in the language acquisition literature. They are common in vocabulary development between 14 and 24 months of age, and similar functions have been reported for aspects of grammatical development between 20–36 months of age. How should such bursts be interpreted? A number of explanations have been proposed to account for the vocabulary burst. They include "insight" theories (e.g., the child suddenly realized that things have names, Dore, 1974; Baldwin, 1989), theories based on shifts in knowledge (Zelazo & Reznick, 1991; Reznick & Goldfield, 1992), categorization (Gopnik & Meltzoff, 1987), and phonological abilities (Menn, 1971; Plunkett,

1993). Although these theories vary greatly in the causal mechanisms brought to bear on the problem, they all have one thing in common: The proposed cause is located at or slightly before the perceived point of acceleration (i.e., somewhere around 50 words). In a sense, such theories assimilate or reduce the data in Figure 4.3 to a pair of linear relationships, illustrated in Figure 4.4, i.e., two linear functions whose cross-point indexes a sudden, discontinuous change in the rules that govern vocabulary growth.

However, if the function illustrated in Figure 4.3 is accurate,

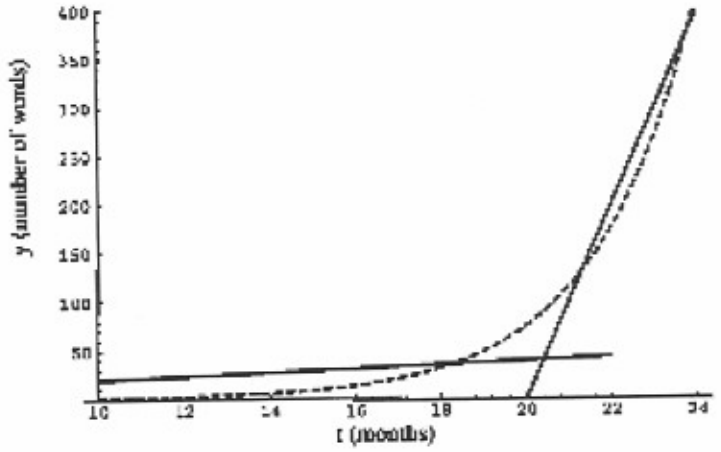


FIGURE 4.4 A piece-wise linear approximation (solid lines) of the data shown in Figure 4.3 (dotted curve).

this perceived point of discontinuity in slope is really an illusion. Figure 4.3 represents a function with a constant fractional rate of increase and can be generated by a linear dynamical equation of the form

$$\frac{dy}{dt} = by \qquad \text{(EQ 4.7)}$$

which tells us that the increase at any given moment is always proportional (the constant b is the percentage increase per time) to total vocabulary size (given by the variable y). This linear dynamical relation between the rate of change and y is graphed in Figure 4.5.

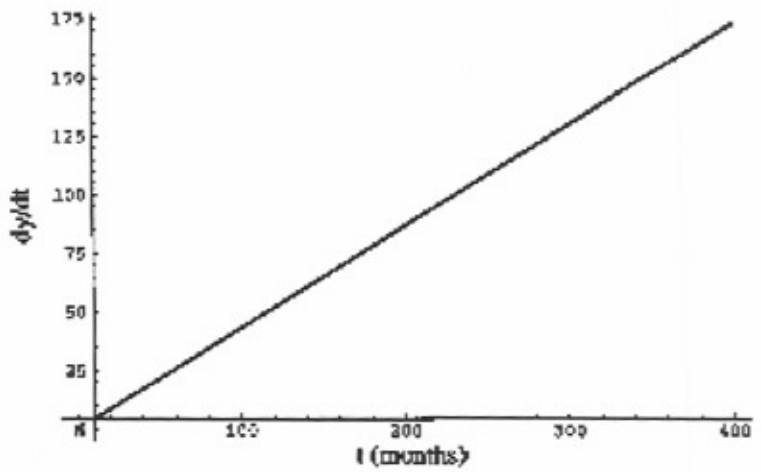


FIGURE 4.5 Graph of the linear dynamic equation Equation 4.7, which underlies the dotted curve in Figure 4.4.

As van Geert (1991) has argued in a paper on the equations that govern developmental functions, we do not need to invoke intervening causes to explain this kind of growth. The real cause of the acceleration that we commonly observe between 50–100 words may be the growth equation that started the whole process in the first place, i.e., conditions that have operated from the very beginning of language development. Of course this does not rule out the possibility that other factors intervene along the way. Environmental factors may act to increase or reduce the rate of gain, and endogenous events like the "naming insight" and/or changes in capacity could alter the shape of learning. Our point is, simply, that such factors are not necessary to account for nonlinear patterns of change.

The most general linear dynamical (autonomous) equation in one dynamical variable, y, is

$$\frac{dy}{dt} = by + c \qquad \text{(EQ 4.8)}$$

Here b and c are constants independent of time and y. Again, by linear dynamics, we mean that the relationship between the rate of change of y and y itself is linear. This relationship when plotted is

simply a straight line like that in Figure 4.5. If $y = 0$, then the rate of word accumulation is the constant c. In other words, this dynamics assumes that even when you have no words you have the ability to learn some. The rate at which you learn those first words would be c, the product of a nonzero efficiency times the number of words per unit time to which you were exposed. If b is positive, the term by implies that the more words you know, the easier it is to accumulate more. This one term by can be thought of as the net of an increase in ability to learn words resulting from previous accumulation and the rate of loss of words due to forgetfulness which may also be proportional to the total number of words. In a sense, the two terms on the right hand side of Equation 4.8 can be thought of as two mechanisms that compete with each other at all times. The term that dominates depends on the size of y which changes in time. Thus the behavior of the system can have different characteristics at different times although the same mechanisms are always operating. In such autonomous models all of the complexity is assumed to come from the natural dynamics of the system itself under a constant environmental forcing.

The general solution of Equation 4.8 is

$$y(t) = \left(y_0 + \frac{c}{b}\right)e^{b(t-t_0)} - \frac{c}{b} \qquad \text{(EQ 4.9)}$$

The solution is simply an exponential plus a constant. Whether the solution increases or decreases with t depends on the sign of $by_0 + c$. Four examples of the shape of this function are shown in Figure 4.6.

This differs from the simpler solution Equation 4.6 in that one can have growth even if the initial number of words is zero. Also, if b is negative, this solution can represent the growth or decay of the number of words to some constant value (explicitly, y tends toward $\frac{-c}{b}$, as t increases indefinitely).

The simple vocabulary burst example in Figure 4.4 illustrates how a two-cause theory can be replaced by a theory based on a single mechanism (a linear evolution equation). More complicated data may suggest more mechanisms. But, as in the general linear dynamical case, it is interesting to consider the possibility that all

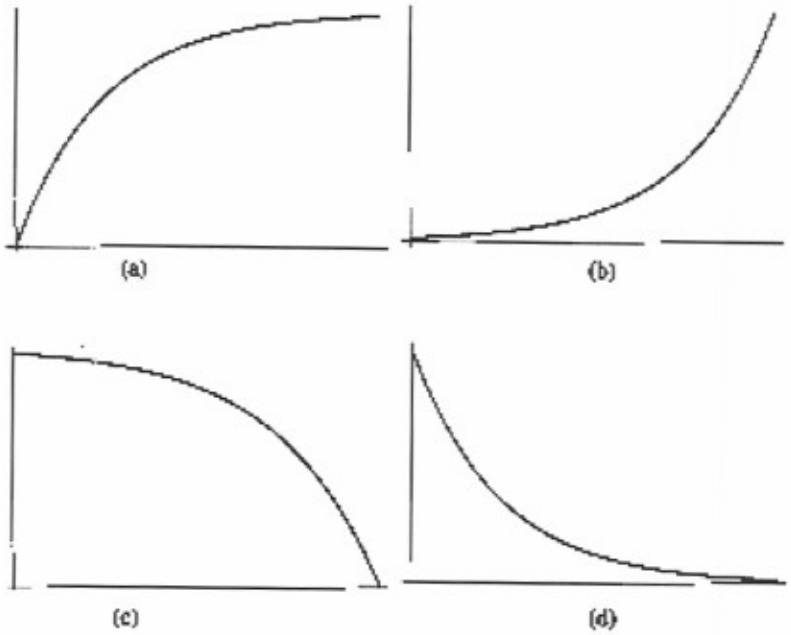


FIGURE 4.8 Hypothetical graphs illustrating nonlinear outcomes (vertical axes) as a function of time (horizontal axes).

the mechanisms are always operating and their relative importance depends only on the state of the system given by y, rather than invoking some ad hoc change prescribed at a certain time or age.

As an interesting example of this possibility, we consider the result found by Johnson and Newport (1989), who have focussed on second language learning by children and adults. Johnson and Newport set out to test the critical period hypothesis, i.e., the idea that there are maturational constraints on language learning that make it easy for children but difficult for adults to acquire a second language (we discuss a connectionist model which makes similar assumptions in Chapter 6).

To test this hypothesis, Johnson and Newport studied a sample of Chinese-English bilinguals who arrived in the United States (and

began the process of second language learning) at various points from infancy to the early adult years. To overcome some of the methodological problems encountered in earlier studies comparing child and adult second language learners, they focussed entirely on the end state of language learning in adult bilinguals, approximately thirty years after arrival. In adult bilinguals who arrived between 0–16 years of age, they report a significant negative correlation between age of arrival and accuracy scores on a test of sensitivity to English grammar (r = -.87). In adults who arrived somewhere between 16–40 years, there was no significant relationship between age of arrival and grammar scores (p < -.16), although as a group the late-arrivers performed at a lower level than those who arrived before age 16.

Because these two regression lines are so different, Johnson and Newport conclude that different forces are operating in these two periods of development. Specifically, Johnson and Newport supposed that there is a change in maturational state, from a plasticity or readiness for language learning under age 16 (a state that undergoes a linear decrease from birth) to a steady state of limited success in second language learning after age 16.

But this is not the only possible explanation for these data. In Figure 4.7, we have replotted the data from Johnson and Newport on a single graph that spans 0–40 years.[1] The two straight dotted lines in Figure 4.7 represent the two regression lines, for 0-16 years and for 16–40 years, respectively. These two lines illustrate the "change of state" view presented by Johnson and Newport.

However, Figure 4.7 also includes a single curvilinear function (solid line) that we have fit to the same data, spanning the arrival range from 0–40. This curvilinear outcome function is a solution of the linear dynamical equation Equation 4.8, given by Equation 4.9. Here y is the percentage of correct performance. The simple relationship between the rate of change and y is illustrated in Figure 4.8. In other words, the "two-state" maturational effects described by Johnson and Newport can be fit by a theory in which

1. We thank Jacqueline Johnson for generously making these data available to us.

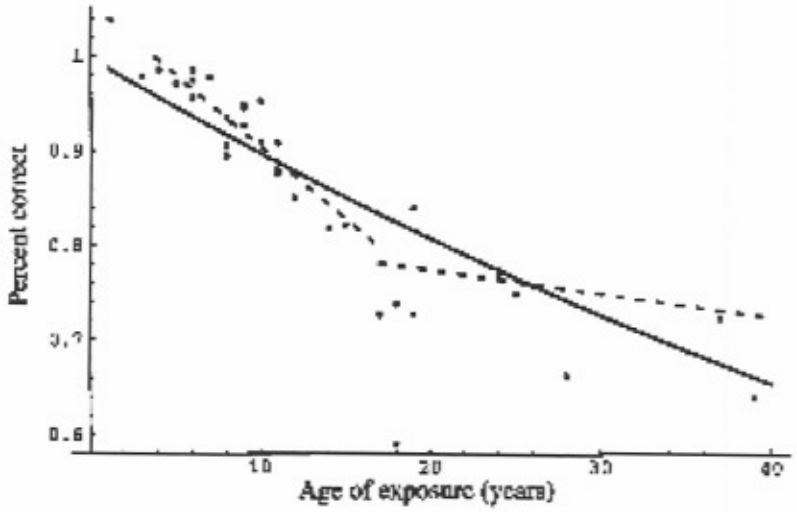


FIGURE 4.7 Data from Johnson & Newport (1989), fit by two models. In the two-stage model proposed by Johnson and Newport, language learning below the age of 16 is facilitated by a state of plasticity which is lost after the age of 16. The two phases are modeled by two different linear regression equations (graphed with dotted lines). The mean percentage of variance accounted for by these two lines is 39.25%. Alternatively, the same data may be fit by a single nonlinear function, shown by the solid line (the nonlinearity is slight, so that visually, within the range of values shown, the curve looks linear). The percentage of variance accounted for by the single stage nonlinear model is 63.1%

two simple mechanisms constantly compete to produce a smooth curve of changing slope. The changes are all determined by the direct response of the learning system to the constant environment and the periods of rapid versus slow change are determined by the state of the learning system itself, and not by any system external to it. Of course these are not the only facts in the Johnson and Newport data. One must also deal, for example, with a marked increase in the variance around the curve after 16 years. Our point here is not to dispute the conclusions offered by Johnson and Newport, but to point out that a simple linear dynamical alternative is available to account for the same nonlinear outcomes.

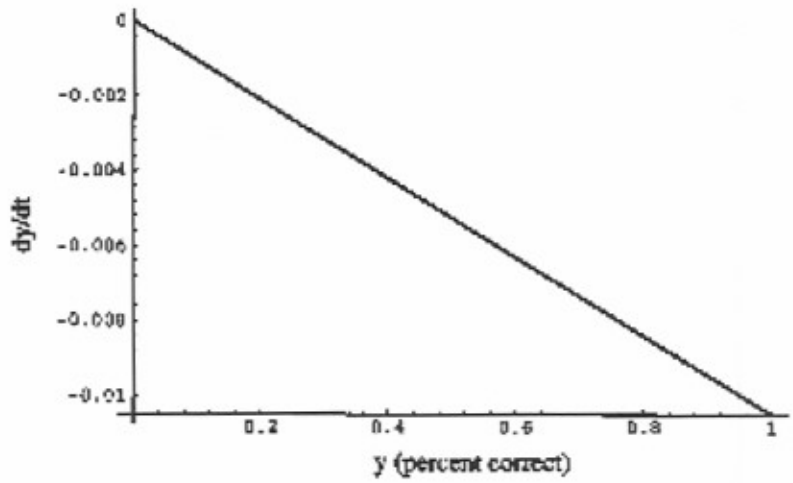


FIGURE 4.8 The linear dynamic equation underlying the outcome graphed in the nonlinear model of the Johnson and Newport data shown in Figure 4.7

It turns out that a similar analysis applies to another domain in which apparently abrupt changes in learning have been interpreted as evidence for two-state models: the acquisition of bird-song. Although birds and humans are clearly different, tantalizing similarity in pat- terns of acquisition of song by young birds have long been cited as evidence for the inherent plausibility critical period models.

Marler and Peters (1988) presented new data which they argue bolsters the claim that certain species of birds undergo a critical period of plasticity during which exposure to adult models will result in the acquisition of normal song. If young birds are deprived of the necessary input during the first 100 days after hatching, the percentage of correct song structure learned drops dramatically.

In Figure 4.9 we have graphed the data from Marler and Peters, along with two ways to model the data. The model Marler and Peters propose is shown by the dotted lines, which illustrate a very extreme form of the two state hypothesis. During the first state (roughly the initial 75 days of a bird's life) learning is at a high level. If learning is delayed until this initial critical period is passed,

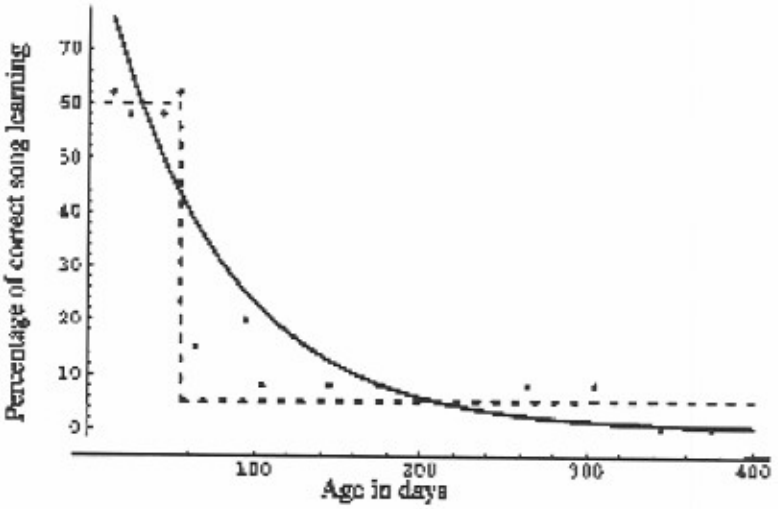


FIGURE 4.9 Data from Marler and Peters (1988), fit by two models. The two-stage model (dotted line) assumes an abrupt loss of plasticity at approximately 75 days; this model accounts for 59.9% of the variance in the data. The solid curve corresponds to the single-stage model, which assumes a nonlinear drop in learning; this model accounts for 73% of the variance.

then performance drops to a very low level. This two state model accounts for 59.9% of the variance. However, the same data can be accounted for by an exponential function plus a constant graphed in the solid line. This single-stage nonlinear function accounts for 73% of the variance. Once again, we see that a single nonlinear function arising from a simple linear mechanism or dynamics may produce effects which look as if they arise from multiple state systems.

On the other hand, there are excellent reasons to believe that the linear dynamical equation shown in Figure 4.4 tells only part of the story (see also van Geert, 1991). Let us suppose for a moment that vocabulary growth continued to follow this dynamic function for a few more years. At this rate of growth, our hypothetical child would have a vocabulary of approximately 68,000 words at 3 years of age, 12 million words at four years, and 2 billion words by the time she enters kindergarten! Since there are no known cases of this

sort, we are forced to one of two conclusions: (1) some exogenous force intervenes to slow vocabulary growth down, or (2) the initial acceleration and a subsequent deceleration were both prefigured in the original growth equation. We will turn to this point shortly, in a discussion of nonlinear dynamics.

(3) **Nonlinear monotonic functions with nonlinear dynamics.** Figure 4.10 illustrates a relatively simple nonlinear pattern known as the logistic function. Functions of this kind are quite common in

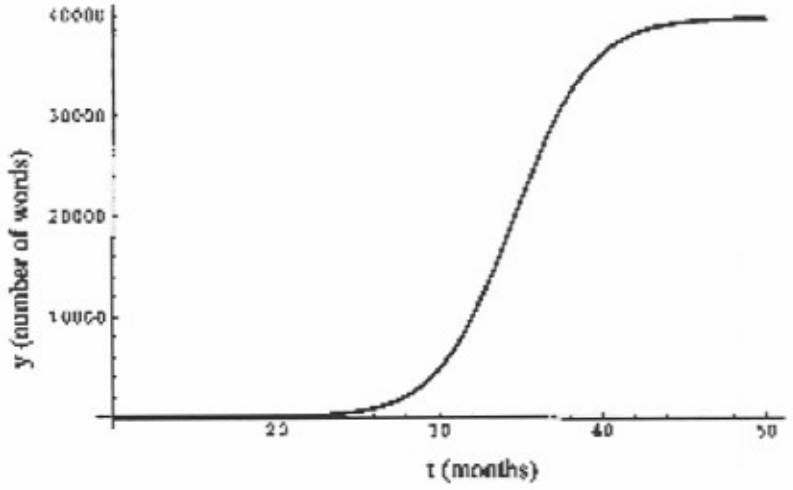


FIGURE 4.10 A nonlinear form of change in vocabulary (y axis) as a function of time (x axis).

behavioral research, and they are also common in neuroscience (where they can be used to describe the probability of firing for a single neuron or population of neurons, assuming some threshold value).

The particular example in Figure 4.10 reflects a hypothetical example of vocabulary growth from approximately 10–48 months of age. The first part of the graph from 10–24 months is almost identical to the growth function in Figure 4.6b (although that is difficult to

see because of the change in scale). In contrast with the ever-increasing exponential burst in Figure 4.6b, Figure 4.10 does have a true inflection point (half-way up the curve), defined as the point at which the rate of change stops increasing, and starts to slow down. This pattern can be described by the equation

$$y = \frac{y_0 e^{b(t-t_0)}}{\left(1 + \left(\frac{y_0}{y_{max}}\right)(e^{b(t-t_0)} - 1)\right)} \qquad \text{(EQ 4.10)}$$

Although this is a more complicated equation than the ones we have seen so far, the only new term here (in addition to the ones introduced in the previous example) is the constant parameter y_{max}, which stands for an estimated upper limit on adult vocabulary of 40,000 words. (We do not have to assume that the child knows this upper limit in advance; instead, the limit might be placed by the available data base or by some fixed memory capacity). The nonlinear dynamic equation that underlies this nonlinear pattern of change is

$$\frac{dy}{dt} = ay^2 + by \qquad \text{(EQ 4.11)}$$

where a is defined as $\frac{-b}{y_{max}}$.

This dynamic relationship between rate of growth and the variable undergoing change is graphed in Figure 4.11. The main thing to notice here is the changing relationship between ay^2 and by which explains why the initial acceleration and subsequent decline in growth are both contained in the same equation. To obtain the shape in figure Figure 4.10, we make b positive and a negative. Early in the evolution, say from 10 to 24 months, by is much larger than $-ay^2$, and so the evolution during that period is almost identical to that in Figure 4.6b. As time proceeds, and y increases in size, ay^2 becomes closer in size to by. Because the growth rate is defined as the difference between these two terms of opposite sign, the rate of growth approaches zero at the specified vocabulary maximum, and that vocabulary maximum, $y_m = \frac{-b}{a}$, is predicted by balancing

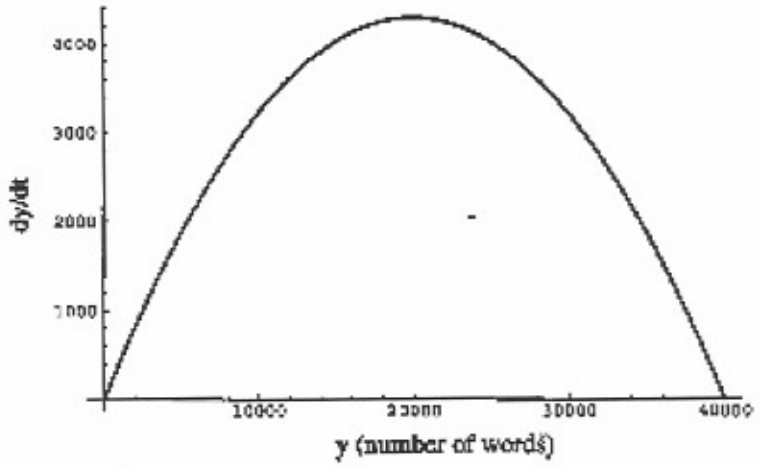


FIGURE 4.11 The graph of the nonlinear dynamic equation (Equation 4.11) underlying the outcome shown in Figure 4.10.

the two terms. Here we see the two different "stages" in development are given by a continuous change in the relative magnitude of these two terms.

This dynamic equation provides a better overall match to the child vocabulary growth data than the exponential in Figure 4.3 (i.e., the function that predicts a 2 billion word vocabulary when the child enrolls in kindergarten). However, it is still grossly inadequate. In Figure 4.10, our hypothetical child is already close to adult vocabulary levels at 4 years of age—unlikely under any scenario. In other words, as van Geert also concludes in his analysis of lexical growth patterns, the symmetrical properties of the logistic function cannot capture the apparent asymmetries in rate of growth evidenced across the human life time. Additional mechanisms are required to explain the asymmetries that are actually observed in infant vocabulary development, in particular the asymmetric damping or slowing down of vocabulary learning between 2–4 years of age. For example, Dromi (1987) has proposed that this damping occurs because the child has switched her attention from vocabulary to grammar. In this case, we have formal justification for

adopting a more complex developmental model. This contrasts with the vocabulary burst example described above, where a simple linear dynamic is sufficient to handle all the data. The solution proposed by van Geert was to use a delay equation, that is, the rate of change of the system depends on its state at an earlier time. An alternate solution would be to introduce a slightly more complicated version of the dynamical Equation 4.11. In that equation the rate of change is zero if $y = 0$. It is more reasonable to assume that there is a mechanism for learning even if you have no words to start with. This can be achieved by adding a constant to the right hand side of Equation 4.11. Thus we have

$$\frac{dy}{dt} = ay^2 + by + c \qquad \text{(EQ 4.12)}$$

and the general solution to this evolution equation is

$$y(t) = y_m \frac{1 - e^{2a(y_m - y_i)t}}{1 - \frac{y_m}{2y_i - y_m} e^{2a(y_m - y_i)t}} \qquad \text{(EQ 4.13)}$$

where y_m is the maximum number of words that the curve tends toward, y_i is the inflection point where the rate of growth stops increasing and begins to slow down. Both y_m and y_i are defined by the values of the constants a, b, and c. Also a can be defined in terms of y_M, y_I, and c, which is the initial slope of the solution curve. The formula for this is

$$a = \frac{c}{2y_i y_m - y_m^2} \qquad \text{(EQ 4.14)}$$

Thus with this dynamics, one can choose the initial slope of the graph, the inflection point, and the maximum number of words all independently to match the data. Figure 4.12 shows this solution with these parameters all chosen to make a reasonable fit to typical vocabulary growth over a lifetime.

We can also write the rate of change as an efficiency of learning

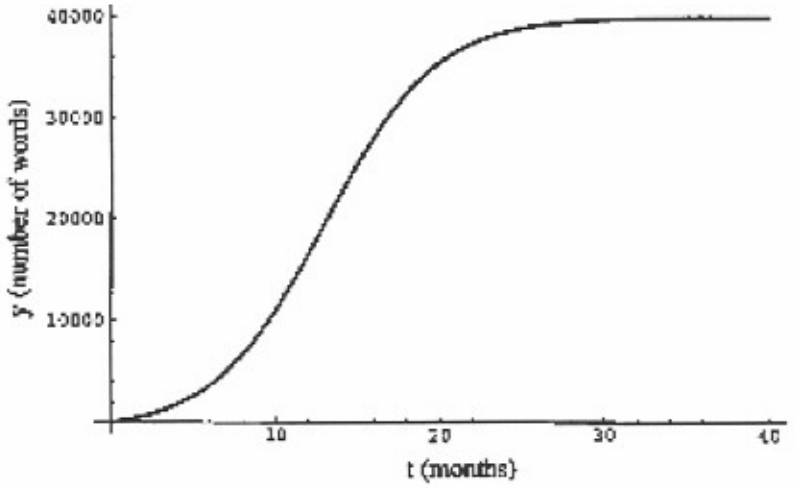


FIGURE 4.12 Modified model of relationship between change in vocabulary (y axis) as a function of time (x axis).

$e(y)$ times the rate of input of new words g. That is,

$$\frac{dy}{dt} = e(y)g \qquad \text{(EQ 4.15)}$$

where

$$e(y) = \frac{ay^2 + by + c}{g}. \qquad \text{(EQ 4.16)}$$

The c term corresponds to the mechanism that permits learning at a constant rate for the first words. The term by, proportional to the number of words known, increases the efficiency of learning because of the familiarity with other words. The term ay^2 is negative, thus it decreases the efficiency of learning. A possible mechanism that would require such a quadratic term would be interference between words. Since the number of pairs of words that can interfere increases as $y(y-1)$ or approximately y^2, it is reasonable to include such a quadratic term in the dynamics.

(4) **Nonmonotonic change.** Now we come to the patterns that have always provided the biggest challenge to theories of development. These involve nonmonotonic outcomes, in which behavior moves in one direction for a while, and then (perhaps only temporarily) reverses.

The first example, shown in Figure 4.13 is actually quite com-

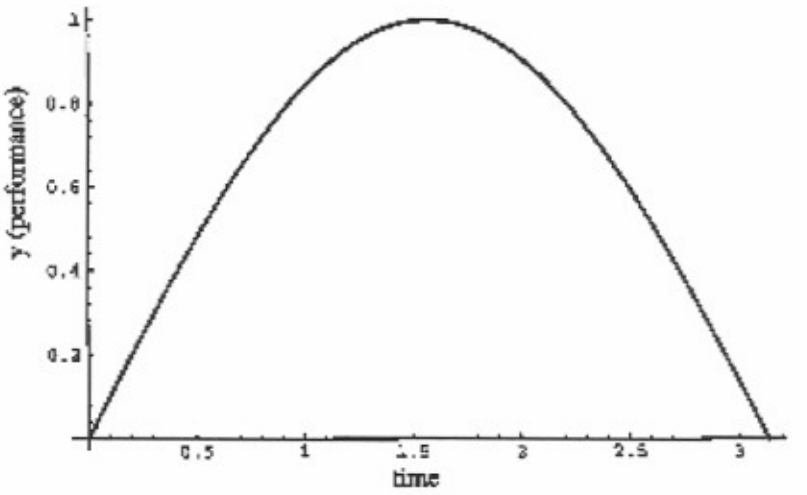


FIGURE 4.13 Hypothetical performance curve in which performance increases early during life, peaks, and then declines during the remainder of life.

mon if we take the whole human life-span into account: An increase in outcomes (performance, size, ability, etc.) in the first half of life, followed by a symmetrical decline in the second half. The outcome function shown in Figure 4.13 is actually

$$y(t) = \sin bt \qquad \text{(EQ 4.17)}$$

The dynamical equations of the type that we have been dealing with take the form

$$\frac{dy}{dt} = f(y) \qquad \text{(EQ 4.18)}$$

that is, autonomous equations for the rate of change of a system described by only one dynamical variable y. These kind of equations can only represent monotonic behavior. This is because the rate of change is a function only of y. Imagine at two different times the function $y(t)$ has the same value, for example once on the increasing side of Figure 4.13 and once on the decreasing side. Since the value of y is the same at both of these points, the function $f(y)$ will be the same at both of those times, thus $\frac{dy}{dt}$ must be the same at both points, but then they must both be points where the curve is increasing or decreasing because the slope must be the same at two points that have the same value of y.

One way to deal with nonmonotonic functions would be to introduce explicit time dependence into the equation,

$$\frac{dy}{dt} = f(y, t) \qquad \text{(EQ 4.19)}$$

that is, to make the equation nonautonomous and change the slope by prescription. For example, we could generate the $\sin bt$ curve from the equation

$$\frac{dy}{dt} = b\cos bt \qquad \text{(EQ 4.20)}$$

(the cosine function has positive/negative values just where the slope of $\sin bt$ is positive/negative). But this would be a poor model of the possible interesting dynamics of the system that produce the evolution. This kind of equation would be better reserved for cases where the systems evolution really is controlled in detail by external driving.

Another way to approach the problem is to realize that using one dynamical variable to describe a complex system may be inadequate. In reality the learning system is highly complex and many components would be needed to accurately describe its behavior. We would then want to represent the evolution of a dynamical system with a vector equation

$$\frac{d\vec{y}}{dt} = f(\vec{y}) \qquad \text{(EQ 4.21)}$$

where

$$\vec{y} = (y_1, y_2, \ldots, y_n) \qquad \text{(EQ 4.22)}$$

is a vector of n variables that represent different components of the system. Even in linear dynamics, then, nonmonotonic change is possible.

For example, the dynamical equation which generated the curve in Figure 4.13 is

$$\begin{aligned} \frac{dy_1}{dt} &= by_2 \\ \frac{dy_2}{dt} &= -by_1 \end{aligned} \qquad \text{(EQ 4.23)}$$

where b is a constant. This is a system of coupled linear equations. It is linear because no higher than the first powers of the dynamical variables are involved, and the two equations are coupled because the growth of y_1 depends on the value of y_2 and vice versa. For example, in this case if b were positive, the first equation says that y_1 would be increasing if y_2 were positive and decreasing if it were negative. Similarly, the second equation says y_2 is increasing/decreasing in time when y_1 is negative/positive. The result is that $y_1 = \sin bt$ and $y_2 = \cos bt$.

Dynamical systems can be modeled at many different levels. For example, the y_i may be some macroscopic variables such as the number of words accumulated and the mean word length of utterances. At the other microscopic extreme the system may be described by its most fundamental units. For example, y_i could be the level of activation of neuron i in the brain. Actually, by increasing the number of components contributing to the behavior, arbitrarily complicated behavior could be represented even with just linear dynamics. By including the possibility of nonlinear interactions extremely complicated behavior may appear even for systems with as little as three interacting components such as in the case of

the Lorenz attractor that we will describe in a later section of this chapter.

The important insight here is that we need no external cause to change the course of development halfway through life. The rise and fall of this developmental system are prefigured in the same dynamic equation, set into motion at the beginning of its life. Of course we would not want to insist that this is the necessary explanation for the unhappy events in Figure 4.13. Alternative accounts are possible, but they would necessary involve external influences to impinge on the learning system rather than an uninterrupted interaction between the system and the data presented.

Now we come to the parade case of nonlinear development: the so-called U-shaped functions in which things get worse for just a little while before they get better. An outcome function of this kind is illustrated in Figure 4.14.

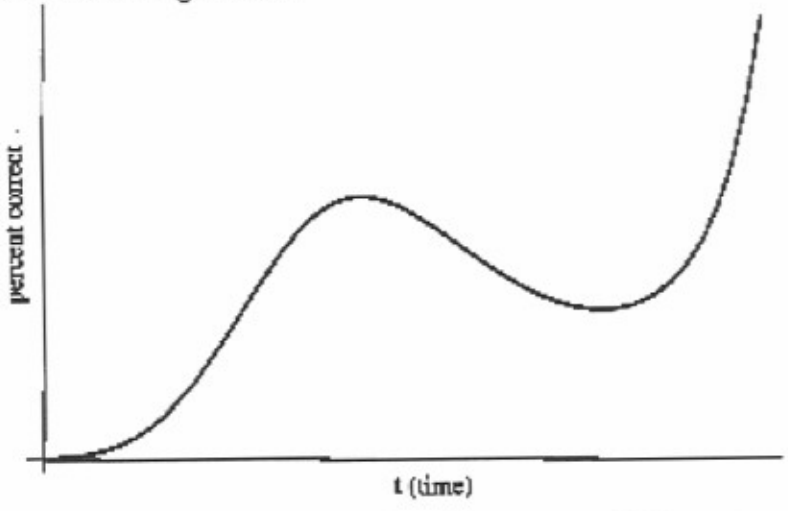


FIGURE 4.14 Hypothetical U-shaped curve, in which an initial improvement in performance is followed by a temporary decline, followed by mastery of the task.

This particular function was chosen to approximate the famous U-shaped pattern for production of the English past tense (Berko, 1958; Brown, 1973; Bybee & Slobin, 1982; Ervin, 1964; Marchman, 1988; Marcus et al., 1992; Pinker & Prince, 1988; Plunkett & March-

man, 1991; Rumelhart & McClelland, 1986), discussed at length in Chapter 3, and which will figure prominently in our arguments.

A graph such as this could be obtained either by making our two variable model nonlinear, or adding a third variable, y_3, in a linear dynamical system. A simple nonlinear model could be constructed as follows:

$$\begin{aligned} \frac{dy_1}{dt} &= h(ay_1^2 - by_2 + c)y_1 \\ \frac{dy_2}{dt} &= h \end{aligned} \qquad \text{(EQ 4.24)}$$

The solution to this system of equations is

$$\begin{aligned} y_1(t) &= y_{10}e^{\frac{ah^3}{3} - \frac{bh^2}{2} + ch} \\ y_2(t) &= ht \end{aligned} \qquad \text{(EQ 4.25)}$$

where we have assumed $y_2(0) = 0$. The idea behind the dynamical model is that y_1 represents the capability for getting the past tense correct and y_2 is some measure of the comprehension of the regular rule. In this crude model we just assume y_2 keeps growing with time. Now if we assume y_1 should grow at a rate proportional to the examples of the past tense that the child has already absorbed, we would write

$$\frac{dy_1}{dt} = ry_1 \qquad \text{(EQ 4.26)}$$

But we also want to build in some loss of ability when the knowledge of the regulars starts destroying some of the irregular patterns that had originally been correctly incorporated. This happens as y_2 grows. Finally we want y_1 to have a positive growth rate again when the knowledge or the regular rules has reached a sufficiently high level. This is accomplished by writing the growth rate as

$$r = h(ay_2^2 - by_2 + c) \qquad \text{(EQ 4.27)}$$

Thus when y_2 is small y_1 grows at the rate $h = dc$, at some intermediate value of y_2 the $-by_2$ term dominates and makes r negative and y_1 decreases, then when y_2 is sufficiently large, the growth rate is again positive. So this is one way to model such behavior.

Now, let us consider for a moment the usual treatment of outcomes like the U-shaped development of the English past tense. The basic story (grossly oversimplified) is something like this:

(1) Children begin to talk about actions by using an unmarked form, usually in the present tense (e.g., "Daddy come").

(2) The child's first systematic efforts to explicitly mark the past tense usually involve high frequency irregulars, produced correctly (e.g., "Daddy came").

(3) After a period of correct production, the same child begins to produce errors in which the regular past is over-generalized of the irregular past (e.g., "Daddy comed"). Note that these errors are usually relatively rare (under 20% of all past tense forms), and they coexist with continued usage of the correct form for a period that can range from weeks to years. In the same period, some investigators report the existence of a very rare error in the opposite direction, i.e., "irregularizations" in which an irregular pattern is over-generalized to regular and/or to other irregulars (e.g., "pick"→ "pack", perhaps by analogy to "sit"→ "sat").

(4) Eventually these errors of over-generalization phase out, and the child enters into the mature state of correct regular and irregular verb marking.

When these phenomena were first described by Roger Brown and his colleagues (Brown, 1973), investigators proposed an explanation based on two developing mechanisms: rote memory (which is responsible for the early mastery of high frequency irregulars), and rule extraction (which is responsible for the subsequent extraction of the "-ed" rule and its over-generalization to other forms). Presumably, these two mechanisms compete across the period in which over-generalization errors occur, until the boundary conditions are established for each one.

This behavioral example and its two-mechanism explanation are familiar to most developmental psychologists. Indeed, this may be the best-known and most-cited case in the field of language

acquisition. It is perhaps for this reason that the famous Rumelhart and McClelland (1986) simulation of the past tense had such a dramatic impact, inspiring a wave of criticisms (Marcus et al., 1992; Pinker & Prince, 1988 and counter-attacks (Daugherty & Seidenberg, 1992; Hare & Elman, 1995; Marchman, 1988, 1993; Plunkett & Marchman, 1991, 1993). Rumelhart and McClelland were able to replicate many of the details of past tense learning in a neural network with no explicit division between rote and rule. Although this initial work made a number of important points, the model contained several flaws, a fact that is acknowledged by most connectionists today. But subsequent studies by other investigators have overcome the initial weaknesses in the R&M model, and it now seems clear that the basic pattern in Figure 4.14 can be obtained in a homogeneous neural network.

So where does the U-shaped pattern come from, if a single mechanism is responsible for the entire developmental trajectory? As P&M have pointed out, the U-shaped function reflects a dynamic competition between regular and irregular mappings within a single system. As the number of verbs in the competition pool expands across the course of learning, there are shifts in the relative strength of regular and irregular forms. The U-shaped dip in the learning curve occurs around the point in development in which there is a change in the proportional strength of regular "-ed" compared to other mappings.

Here is the basic irony that we want to underscore. At some level, as pointed out, the two-mechanism account of the U-shaped function is correct. That is, to produce the pattern in Figure 4.14, we need two competing terms in the equation. The difference between this account and the classic two-mechanism view lies in the fact that our two competing terms are contained within the same equation, and hence (by implication) within a single system. In other words, there is fundamental difference between the classic account and the connectionist view at the level of cognitive architectures. However, the basic intuition that "two things are in competition" is necessary in both accounts.

(5) True discontinuities. In the examples presented so far we have tried to illustrate that changes in the behavior of a system can be due to competition between different mechanisms within the system and do not need to be prescribed ahead of time nor imposed by changes in the environment. Depending on the kind of nonlinearity employed the autonomous dynamical system described by

$$\frac{dy}{dt} = f(y) \qquad \text{(EQ 4.28)}$$

can exhibit behavior with sharp changes in the rate of growth. A simple way to introduce such changes into the models is to use nonlinear functions that change rapidly for some small variation of the dynamical variables. The logistic function is such a function.

Even more extreme is the step function, $H(y)$. This function is 0 when its argument is negative and 1 when its argument is positive. Thus if we consider the nonlinear evolution equation of the form

$$\frac{dy}{dt} = a + bH(y - y_c) \qquad \text{(EQ 4.29)}$$

the solution will have two parts:

$$\begin{aligned} y(t) &= at && \text{for} \quad (t < t_c) \\ y(t) &= (a+b)(t - t_c) + y_c && \text{for} \quad (t \geq t_c) \end{aligned} \qquad \text{(EQ 4.30)}$$

where we have assumed $y(0) = 0$. There is a discontinuous change of slope at the critical time defined by

$$t_c = \frac{y_c}{a} \qquad \text{(EQ 4.31)}$$

Thus, sharp changes in behavior can be due to the natural evolution of a nonlinear system even when the external forcings are constant.

On the other hand, it may be that discontinuous changes are in fact caused by the influence of some other system, such as the hormonal system, on the learning system. How would we describe that? One way would be to introduce new dynamical variables, $x(t)$, with their own laws of evolution and allow them to play a role

in the evolution of the learning system represented by $y(t)$. Thus we would write

$$\frac{dy}{dt} = f(x, y)$$
$$\frac{dx}{dt} = k(x) \qquad \text{(EQ 4.32)}$$

where we continue to assume autonomous dynamics for simplicity. The system represented by x then can act as an internal clock, having its own evolution but not affected by the system represented by y, it can still influence y. The influence of x need not introduce discontinuous changes in $\frac{dy}{dt}$, but it might. An example of this would be the system

$$\frac{dy}{dt} = a + bH(x - c)$$
$$\frac{dx}{dt} = 1 \qquad \text{(EQ 4.33)}$$

With $x(0) = 0$, the solution for x is $x(t) = t$, that is, x is just time. Thus the rate of change of y for $t < c$ would be a, and for $t > c$ it would be $(a + b)$. If $c = t_c$, the result for $y(t)$ is precisely the same as the example given in Equation 4.33 above, although the conceptual models are very different. In the first case it is the growth of the variables describing the learning system that triggers the discontinuous change. No external influences are imposed, no set time is prearranged, and the change occurs when a certain level of learning is achieved. In the second, it is the interference of one system which has nothing inherently to do with learning (when growth is independent of learning) that can trigger the change at a prescribed age.

In many of those cases we know with some precision how and why a continuous shift in one parameter works to produce a discontinuous outcome. Take the continuous dimension of temperature: Ice turns to water, and water turns to gas, all with the turn of a single dial. Where does the "new stuff" come from? The answer lies in the interaction between temperature and the molecular properties of water, where the bonds between molecules are made and broken. The key insight here is that a continuous shift along a single param-

eter can put the entire system into a different problem space. In the second part of this chapter we discuss the notion of bifurcations, which provide one way in which small changes in a parameter which governs a dynamical system can lead to dramatically different behavioral regimes. Once the system is boot-strapped into this new problem space, a different set of forces take over to create the final outcome. In cases like this, we cannot even begin to count entries along the y axis until a certain threshold is reached along the x axis—a critical point t_c. To what extent is it meaningful to say that the phenomenon in question was caused at t_c? It appeared at t_c, but its causal history must include everything that led up to the critical value.

(6) Interacting patterns. So far we have talked about the causal/temporal history of individual outcomes. In the last example (true discontinuity) we suggest that certain outcomes cannot occur until the system passes a threshold value, enough to permit entry into a new problem space. The new problem space, in turn, can be described (or so we hope) by principles and laws that dictate the eventual outcome. We have, therefore, introduced the notion of interacting patterns of growth. This brings us to a final point concerning the proper interpretation of change in two or more systems with different developmental trajectories, something we will call "co-development."

The simplest version of co-development is a linear correlation: two systems increase or decrease together over time, in a relationship that can be described by a line. It is quite common in developmental psychology to find claims about the causal relationship between two variables that change together in this way. A case in point is Lenneberg's (1967) argument for the connection between linguistic and motor milestones (e.g., children walk and talk around the same time). Of course no one is trying to suggest that language causes motor development, or vice-versa. The argument is usually cast in terms of a third force, i.e., a maturational clock of some kind that sets the pace for disparate domains. Similar arguments can be found in research on the cognitive bases of language, where investigators have pointed out analogies in the onset times and develop-

mental trajectories of verbal and non-verbal events, e.g., parallels between language and symbolic play in the second year of life. Here too, no one wants to argue that first words are "caused by" symbolic play, or vice-versa. The claim is, instead, that these two representational domains depend upon the same underlying ability.

This brings us to the "layers of causality" issue that we introduced in our discussion of linear development. Let us suppose, for the moment, that many aspects of human development depend (at least in part) upon resources that grow continuously over time. These resources might include aspects of attention, memory, perceptual acuity, motor speed—for present purposes, the identity of this capacity parameter is not crucial. Assuming a continuous dimension of processing capacity, we can quantify the amount of capacity required for progress in any given task. In the simplest cast, there will be a linear relationship between capacity and output. Thus a 50% cut in capacity would result in a 50% deficit in behavior; and a 50% gain in capacity would result in a 50% jump in behavior. This relationship is illustrated by Function 1 in Figure 4.15 (from Bates, McDonald, MacWhinney & Appelbaum, 1991).

The linear relationship between performance and capacity in Figure 4.15 is, in fact, the assumption that underlies a great deal of research in neuropsychology (for a detailed discussion, see Bates et al. 1991, and Shallice, 1988). It is this assumption that gives force to arguments about the dissociation between cognitive domains. Say, for example, that we find a patient who performs normally on Task A (e.g., a spatial cognitive task) and abysmally on Task B (e.g., a language task). If language and space depended upon the same underlying resource, and there is a linear relationship between resource and capacity for all domains, then a dissociation of this kind could not happen. Since the dissociation does happen, we may conclude with some confidence that this patient has a selective deficit affecting only one domain. That is, language and spatial cognition must depend on two distinct sources that can be independently impaired. Suppose, however, that Task A and Task B are governed by different performance/capacity curves? For example, suppose that Task A follows a linear function (Function 1) while Task B involves a non-

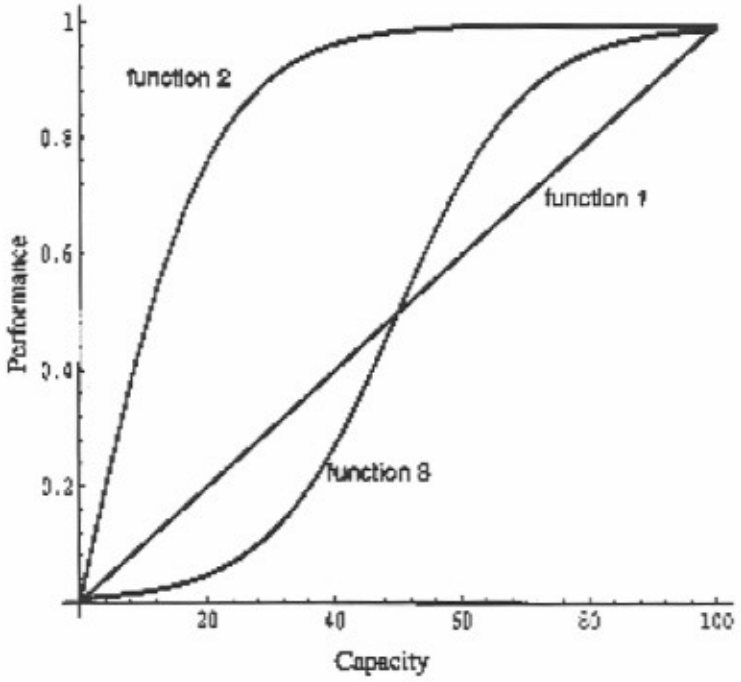


FIGURE 4.15 Function 1 illustrates a linear relationship between capacity and performance. Function 2 illustrates a nonlinear relationship in which rate of increase in performance drops rapidly after approximately a 30% threshold is reached. Function 3 illustrates a sigmoidal relationship in which performance changes rapidly in the region of a 50% threshold

linear relationship of the kind illustrated by Function 2 in Figure 4.15, i.e., a threshold model in which performance approaches ceiling after capacity exceeds 30%, and drops rapidly when capacity falls below 30%. If this is an accurate description of the situation, then a patient who has lost 50% of the resource in question will look very bad on Task A but close to normal on Task B. In other words, a single cause is responsible for both patterns.

Suppose, however, that we now find a patient who shows the opposite symptom profile: normal performance on Task B with serious deficits on Task A. Putting our two patients together, we have a classic double-dissociation, the sine qua non of cognitive neuropsy-

chology. Surely, in this case, we can conclude that A and B are independent systems? In fact, as Shallice (1988) has pointed out, double dissociations can also arise from a single cause if the two tasks differ in the shape of their performance/capacity curves. Compare Function 1 (the linear function) in Figure 4.15 with Function 3 (the S-shaped curve). If our first patient has lost approximately 25% of the shared cognitive resource, he will show a performance profile in which Task B > Task A. If our second patient has lost up to 75% of the same shared resource, he will show a performance profile in which Task A > Task B. In short, claims about the dissociation between two cognitive systems must be evaluated with great caution, with detailed knowledge about the performance/capacity functions that underlie all the domains in question. If we run the arrow forward (capacity goes up, not down), then the same lesson applies to the study of normal development. Two domains can display markedly different developmental profiles, even though they both depend on the same underlying resource. Continuous shifts along that resource dimension can pull these domains apart and bring them back together again, in interesting patterns that tempt us to infer more independence and more discontinuity than we are entitled to claim.

Perhaps we can move our claims of independence to another level. Why, we might ask, do two domains obey such different laws? Why, for example, does Task A follow a linear pattern of growth while Task B reaches asymptote *months* or years before? Isn't this enough evidence, in and of itself, to stake out a modular claim? Yes, and no. We need to know, first of all, the extent to which these disparities in the shape of change derive from the ability in question (e.g., language vs. spatial cognition), or from the tasks we use to measure those abilities. We also need to know more than we do right now about the nature and availability of the input that drives learning in each of these domains. These are difficult questions, but they are the kinds of questions that are easier to answer if we are in a position to simulate learning under different parametric assumptions, to narrow down the critical range of possibilities for study in the human case. Indeed, if it is the case that cognitive change follows nonlinear laws, then simulation may become a cru-

cial tool for developmental research.

The various examples of nonlinearity that we have described here constitute only a small subset of a huge class of possible nonlinear dynamic systems. This class includes an exotic and celebrated form of nonlinearity called chaos, famous because it seems to elude our understanding altogether, being (by definition) completely deterministic but completely unpredictable. The behavior of nonlinear systems is difficult to predict because they have a property called *sensitivity to initial conditions*. In particular, because there is a nonlinear relationship between rate of change and the variable that is undergoing change, the outcomes that we ultimately observe are not proportional to the quantities that we start with. A very small difference in the starting points of two otherwise similar systems can lead (in some cases) to wildly different results. In principle, all these systems are deterministic; that is, the outcome could be determined if one knew the equations that govern growth together with all possible details regarding the input, out to an extremely large number of decimal places. In practice, it is hard to know how many decimal places to go out to before we can relax (and in a truly chaotic system, the universe does not contain enough decimal places). For this and other reasons, nonlinear dynamic systems constitute a new frontier in the natural sciences. For those of us who are interested in applying such systems to problems in behavioral development, this is both the good news and the bad news. The good news is that nonlinear dynamic systems are capable of a vast range of surprising behaviors. The bad news is that they are much harder to understand analytically than linear systems, and their limits remain largely unknown.

Whether or not we are comfortable with this state of affairs, it is likely that developmental psychology will be forced to abandon many of the linear assumptions that underlie current work. We have already shown how linear assumptions may have distorted our understanding of monotonic phenomena like the vocabulary burst, or the famous U-shaped curve in the development of grammatical morphology. It is hopefully clear by now why this assumption is unwarranted. Discontinuous outcomes can emerge from continuous change within a single system. Fortunately, connection-

ist models provide some of the tools that we will need to explore this possibility. Let's now take a look at how some of the concepts we have been discussing are implemented in connectionist networks.

Dynamical systems

Before proceeding to talk about how the above ideas may be implemented in connectionist networks, there is one additional set of concepts which we need to introduce which are central to notions of change: Dynamical systems.

A dynamical system is, very simply, any system whose behavior at one point in time depends in some way on its state at an earlier point in time. There are many systems for which this is not true. When a coin is tossed into the air, whether or not it lands heads up or tails up is totally independent from how it landed on previous tosses (the Gambler's Fallacy is to believe that after a run of heads, a coin is more likely to land tails up—in other words, to believe that this is a dynamical system, when it's not). Mathematically, we could express the outcome with the equation

$$P(\text{heads}) = 0.5 \qquad \text{(EQ 4.34)}$$

(where P denotes probability). The right-hand side of the equation does not in any way take into account prior outcomes.

But clearly there are many cases where behavior does depend on prior states. Suppose we catch a glimpse—so brief that we see only an instant—of a child on a swing, and see that the swing is right below the crossbar (that is, aligned perpendicular to the earth). If all we have to go on is this instantaneous view, we have no way of knowing what the position of the swing will be a second later. The swing might have been at rest, in which case we would expect it to remain there. But it might have been in motion when we took our brief look. In that case it will be in a different position a second later. Mathematically, the change in the position of the swing is given by the following pair of differential equations:

$$\frac{dp}{dt} = \frac{-g\sin\theta}{L} - rp$$
$$\frac{d\theta}{dt} = p \qquad \text{(EQ 4.35)}$$

The length of the swing is denoted by the symbol L, the angular displacement of the swing from the vertical position is represented by θ, p represents speed, and r stands for the coefficient of friction. The second equation tells us the speed (in terms of change in angular displacement, θ), and the first equation tells us how, given that information, the speed will change over time.

The are various ways of classifying dynamical systems, but one useful way is in terms of the kind of behavior they exhibit over time. There are three kinds of behaviors one can distinguish: systems can exhibit *fixed points*, *limit cycles*, and *chaos*.

A system which has an attracting fixed point is one which over time eventually converges on a static behavior. A swing or pendulum, for instance, will ultimately come to rest (assuming friction). The position at rest is called the *attracting fixed point*. Let's look at a very simple equation which we will use to demonstrate a fixed point in a dynamical system.

Let us begin with the equation

$$f(x) = \frac{1}{(1+x)} \qquad \text{(EQ 4.36)}$$

If we graph this for values of x ranging from 0 to 10 we get the curve shown in Figure 4.16.

Suppose we iterate this function. In other words, we begin with some value of x (we'll use 2) and calculate the result: 1/(1+2) or 0.33. Now let us take the output, 0.33, and use that as the new value for x. This time we get 1/(1+0.33) or 0.75. Continuing again eight more times gives us successive values of 0.571429, 0.636364, 0.611111, 0.62069, 0.617021, 0.618421, 0.617886, 0.61809. If we trace the pattern (using what's called a "cobweb graph") we see the trajectory shown in Figure 4.17.

It looks very much as if repeated iterations of this function lead to a single value. Indeed, if we calculate the result of iterating

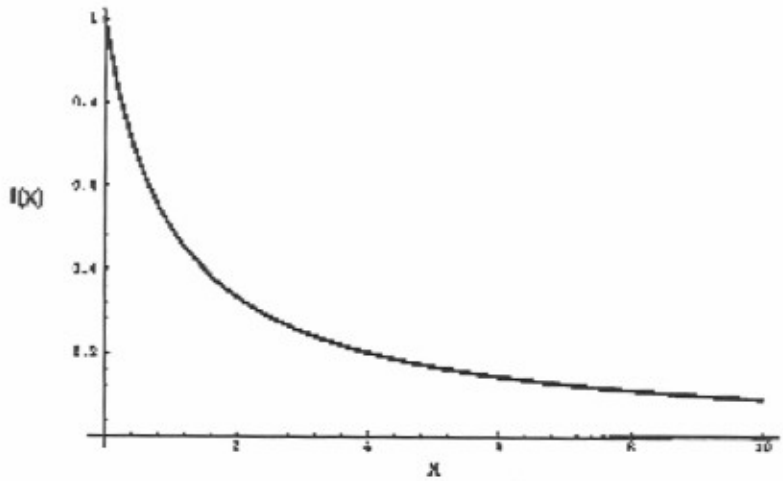


FIGURE 4.19 Graph of the equation $f(x) = \frac{1}{(1+x)}$ for values of x from 0 to 10.

Equation 4.36 on itself, starting with an initial value for x of 2, we find that after 50 iterations we converge on a value of 0.682328. Iterating another 25 or 100 or 1,000 times does not change this value. The system has settled into a fixed point.

Suppose we began with another initial value, say -10? Interestingly, on each iteration we get closer and closer to the value

$$x_{\sim} = \frac{1+\sqrt{5}}{2}$$

which, for the precision available on the machine we used, is approximately = 0.682328. This is true no matter what value we start with. In a way, this equation is like a swing: No matter how hard our initial push, the swing will sooner or later always come to rest (unless, of course, we continue to push it; in this case we have what is called an external *forcing function*).

The fixed point for this function is said to be a global *attractor*, because no matter what initial value we start with, the function ultimately "attracts" us into a final value of 0.682328 (for the given precision available on our machine). (There are also repelling fixed points, which are regions where we remain if we are exactly at the

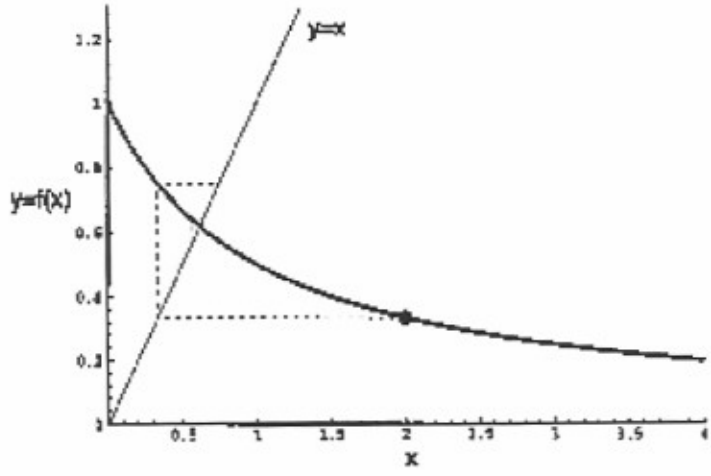


FIGURE 4.17 Graph showing result of iterating Equation 4.36 on itself 10 times, starting with an initial value of 2.0 for x (indicated by the dot). The dashed line traces the changing values of x. On the first iteration, the dotted line crosses the function at x=2.0 and $f(x)$=0.33. On the second iteration, the dotted line crosses the function at x=0.33 and $f(x)$=0.571429. After many iterations, the function settles at values of x=0.682328 and $f(x)$=0.682328, which is an attracting fixed point for this equation.

fixed point, but if we are anywhere else close by, we are pushed away from that value on subsequent iterations of the function.) We might wonder whether a function can only have one attracting fixed point (as is obviously true of the swing example). The answer is no; functions may have any number of fixed points, although it is typical for a function to have only a small number of fixed point attractors. For example, the function:

$$f(x) = \frac{1}{1 + e^{(3 + (-6x))}} \qquad \text{(EQ 4.37)}$$

has three fixed points (two attracting fixed points and one repelling fixed point). Initial values of x which are greater than 0.5 converge toward 0.92928; initial values less than 0.5 converge toward

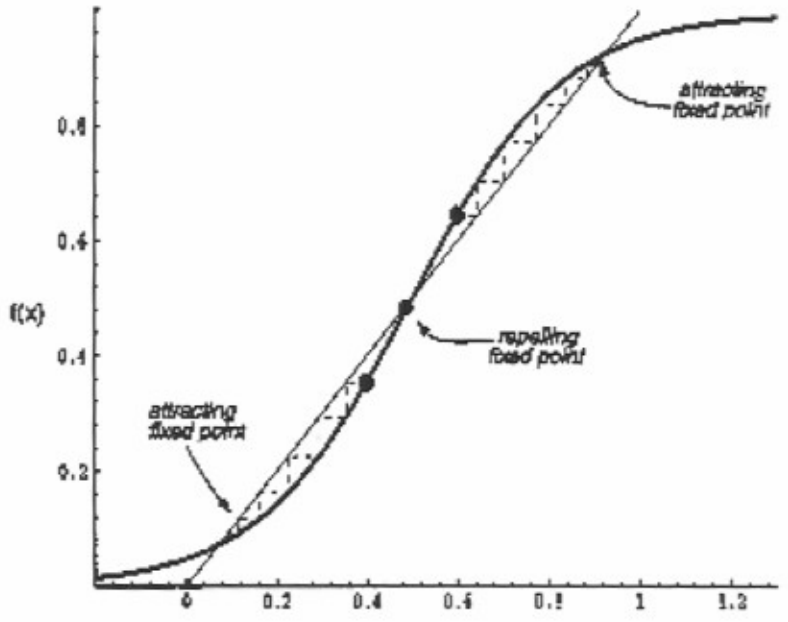


FIGURE 4.18 Graph showing result of iterating Equation 4.37 on itself 10 times, starting with three different initial values for x, shown by dots at 0.4, 0.5, and 0.6. Initial values greater than 0.5 converge toward a fixed point attractor of 0.92928; values less than 0.5 converge toward 0.0707202; and values of 0.5 remain at the fixed point 0.5. The first two are attracting fixed points and the third is a repelling fixed point.

0.0707202; and initial values of exactly 0.5 remain at 0.5 but slight deviations from 0.5 are forced away from that value. This is shown in Figure 4.18.

Fixed points are not the only kind of attractors one finds in dynamical systems. A second type of attractor is the *limit cycle*. In this case, the system exhibits a behavior which repeats periodically. A classic example of a function which can have limit cycles is the quadratic map, given in

$$f(x) = rx(1-x) \qquad \text{(EQ 4.38)}$$

where r is a constant. For example, if we let r be 3.2 and set the initial value of x to 0.5, then on successive iterations, this equation yields 0.8, 0.512, 0.799539, 0.512884, 0.799469, 0.513019, 0.799458, 0.51304, 0.799456, 0.513044, 0.799456, 0.513044, 0.799455, 0.513044, 0.799455, 0.513045, 0.799455, 0.513045, 0.799455, 0.513045. If we example these numbers closely, we see that it looks like we end up with a repeating sequence of two alternating numbers. If we graph the changing values, as is shown in Figure 4.19, it is easy to visualize this cyclic behavior.

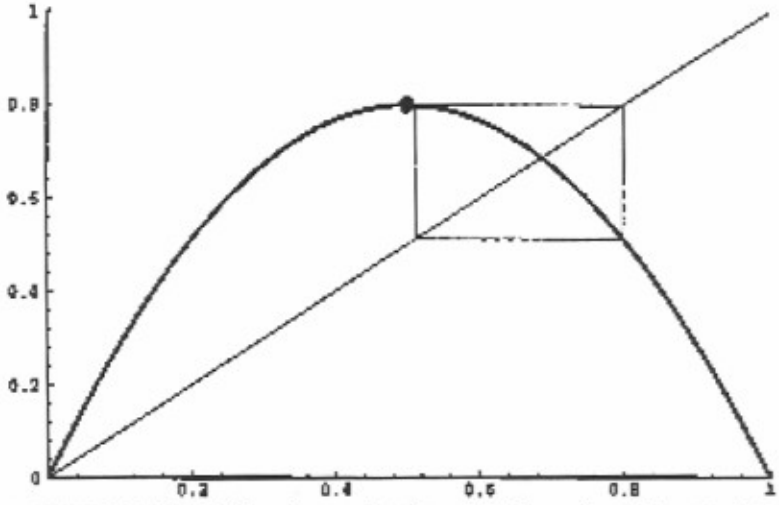
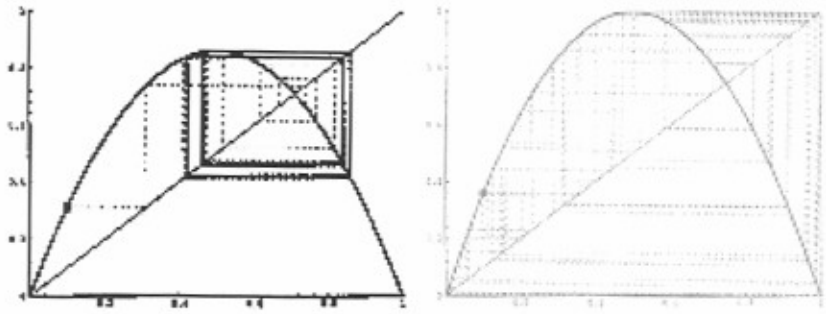


FIGURE 4.19 Graph showing result of iterating Equation 4.38 on itself 10 times, starting with an initial values of 0.5 for x and setting r=3.2. The function cycles through two different values (shown where the cobweb intersects the parabola).

More interesting is the fact that if we give x an initial value of 0.1, or 0.9, we end up with the same results. After many iterations, this equation (with r as we have set it) ends up cycling between the same two numbers: 0.513045 and 0.799455. This means that this limit cycle serves as an *attractor*, meaning that other initial values of the variable x end up being pulled into the same repeating pattern.

The quadratic equation given in Equation 4.38 has been well-studied because it has other interesting properties. If we change the

value of the constant r (remember that in the above examples we kept it at 3.2), then the dynamics of this equation change dramatically. Figure 4.20 shows the behavior of the iterated quadratic equation with values of r=3.45 (left panel) and r=4, in both cases starting with x set to 0.1. In (a) the system soon settles on a repeat-



(a) (b)

FIGURE 4.20 Graph showing result of iterating Equation 4.38 on itself 50 times, setting r=3.45 (a) and r=4 (b), in both cases starting with an initial values of 0.1 for x. (The dots indicate starting positions.) In (a) the system settles into a limit cycle with four possible values. In (b) the system enters a chaotic regime, never exactly repeating a prior state.

ing pattern of period four. In (b) the result is a *chaotic* attractor. Chaotic attractors are the third class of dynamical systems.

In this chaotic case, the system never returns to exactly the same state it was in before. For example, if we iterate Equation 4.38 (setting r=4) 50 times, starting with x=0.1, we end up on the last iteration with x=0.54. But if we start with a very slightly different initial value for x—say, 0.1001—after 50 iterations we have x=0.08. So not only does the attractor not repeat itself, but its behavior varies dramatically depending on the initial values for variables. This property of chaotic systems is called *sensitivity to initial conditions*. Note also that chaotic behavior is not the same as random behavior; if we start with the same initial conditions, we repeat the whole sequence exactly, unlike what happens when we toss a coin. One of the best

known chaotic attractors is the Lorenz attractor (which was developed as a model for changes in weather), shown in Figure 4.21. The

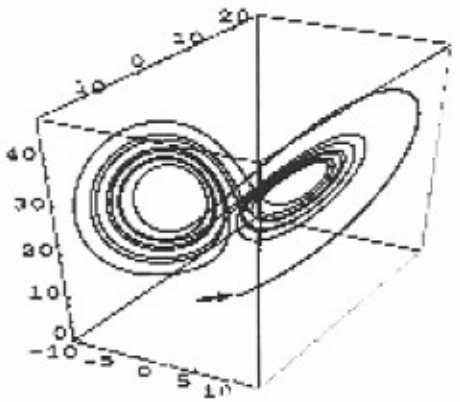


FIGURE 4.21 The Lorenz attractor. This system is an example of a deterministic chaotic system; the behavior never repeats but is not random.

system displays patterned movement through the space of the three variables in the equations which define it, but never returns to exactly the same position in state space as it was previously.[2]

There is a final point we want to make with regard to the behavior illustrated in Figure 4.20. We find that the same equation yields qualitatively different dynamical behavior as a function of a change in one of the constant parameters of the equation (in this case, we changed r from 3.45 to 4). Let's explore the effect of changing this parameter in a more systematic fashion. We'll begin by setting r to 2, feed in an initial value for x (we'll use 0.5 in this example), and then examine the behavior of the system when we iterate the equation. The question we're interested in is how many attractors the

2. Here the variations are continuous in time, as opposed to the examples we have given so far, which involve discrete time steps. Although there are differences between discrete and continuous dynamical systems, all the attractors we have discussed with discrete systems have analogues in continuous time systems.

equation has.

If we carry this experiment out with Equation 4.38, we discover that when r is 2 and x is 0.5, the system converges immediately on a single fixed point (0.5). If we run this experiment again many times, each time increasing the value of r in very small increments, we discover that all versions of this equation have a single fixed point attractor until we get to the case where r is 3.0. Suddenly the behavior of the system shifts and we find a limit cycle with two attractor fixed points. Increasing the value of r even more reveals successive changes in behavior as the system switches abruptly at r=3.44949 from a 2-attractor limit cycle to a 4-attractor cycle, then with r=3.54409 to an 8-cycle, and so on. These shifts always result in increasing the number of cycles by a power of two. Then we find that when r=3.569945 is reached we enter a chaotic regime in which the dynamical behavior never repeats. But if r is increased even more, we revert to a limit cycle regime for a while, then return to chaos. And so on.

The results of this experiment have been graphed in Figure 4.22. The vertical axis shows us the different output values yielded by the equation for the different values of r, shown along the horizontal axes. The changes in behavior are called *bifurcations* (and the graph in Figure 4.22 is therefore called a bifurcation plot). This is because the behavior of the system splits (bifurcates) at the change points.

The relevance of bifurcations for our purposes should be obvious. Children often go through distinct stages in development, exhibiting qualitatively different behavior in each stage. It is tempting to assume that these stages reflect equally dramatic changes and reorganization within the child. Bifurcations illustrate that this need not be the case; the same physical system may act in different ways over time merely as a result of small changes in a single parameter. Later in this chapter we shall present a neural network model of this phenomenon. We'll see how a network which appears to appears to go through different stages of learning and abstraction does so as a result of bifurcations.

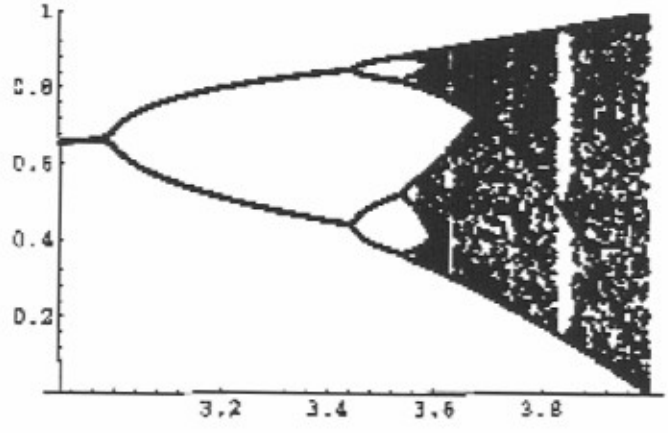


FIGURE 4.22 Bifurcation diagram.

Dynamics and nonlinearity in neural networks

The central topic of this chapter is change, and that means that we are ultimately interested in the dynamical properties of children and networks. One of the things which makes network dynamics so interesting is the presence of nonlinearities in learning and processing. We therefore begin by discussing where nonlinearity arises and what its effects are. Then we shall see how this affects the shape of change.

Nonlinearity in networks

It is possible to have a connectionist network that is fully linear, and some very interesting things can be done with such networks. But today most people are interested in networks that have nonlinear

characteristics, as discussed in Chapter 2. The major locus of the nonlinearity appears in the way in which a unit responds to its input.

The basic job carried out by the units (or nodes) in a network is simple. A node collects input from other nodes (excitatory or inhibitory) and also sends output to other nodes. Viewed this way, a node merely serves as a way-station shuttling activation values around the network. But each node does one small thing in addition, and this has serious consequences. Before passing its input on to other nodes, a node performs a relatively simple transformation on it, so that what the node puts out is not exactly what it took in. This input/output function (what is called the activation function) is typically nonlinear. One common function is graphed in Figure 4.23.

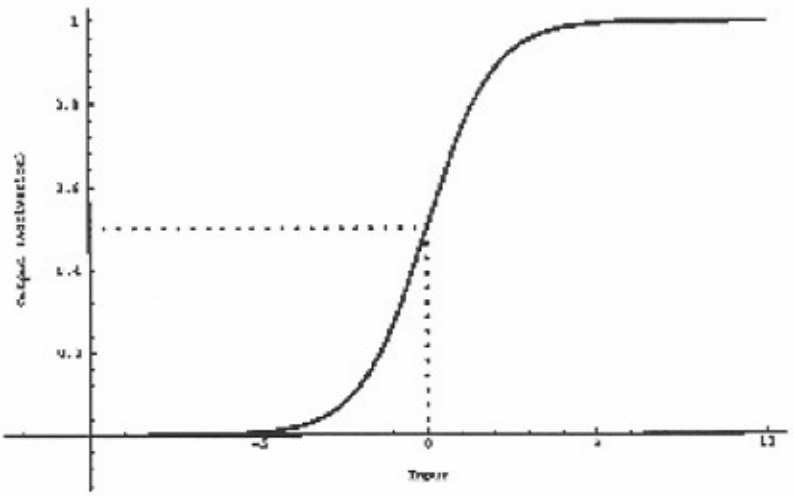


FIGURE 4.23 The activation function used most often in neural networks. The function is $1/1 + \exp^{(-net)}$, where *net* is the net input to a node and *exp* is the exponential function. The net input is potentially unbounded, whereas the resulting output is limited to (in this case) a range of activation values from 0.0 to 1.0. When net input is 0.0, the output is 0.5, which is in the middle of the unit's range of possible activations.

The shape of this function shows how the node's output (measured along the y-axis) changes in relation to the node's total input (shown along the x-axis). Notice that the range of possible values along the two axes differs. The range of possible inputs is unrestricted, since a node might receive input a very large number of other nodes. The range of outputs, on the other hand, is restricted (with this particular activation function) to the interval {0, 1}. The activation function thus squashes the nodes net input first before sending it off to the other nodes it excites or inhibits.

Why might it be useful for the activation function to take such a form? One of the obvious benefits is that the sigmoid function keeps a node's activation under control. Should the input grow too large, the node will simply saturate rather than growing out of bounds. Unconstrained systems have a habit of "blowing up," which can be nasty if one wants stability.

There are other important benefits to this nonlinearity. When a node receives a great deal of inhibitory (negative) or excitatory (positive) input, its response will be binary; the resulting activation will be asymptotically close to 0 or 1, respectively. Furthermore, once the magnitude of the input has reached a certain value, it doesn't matter how much more input is received. The output will still be either 0 or 1. So when it has to, a node can generate an all-or-nothing decision.

At the same time, there is a large gray area of input conditions within which a node is capable of graded responses. When the net input is close to 0 (the dashed line in Figure 4.23), the output is 0.50, which is in the middle of the node's activation range. Notice that small differences in this region of input produce significant differences in output. An input of 2 gives an output of 0.88, whereas inputs of 10 and 12 both give outputs of 0.99. Thus, nodes can make finely tuned responses which vary subtly as a function of input. Importantly, the same unit can make both types of decisions along the same continuum. In certain regions along an input dimension (i.e., near the ends of the continuum) a node may be relatively insensitive to stimulus differences and output nearly the same response to inputs which may be very different. In other regions along the same dimension, the same node may be exquisitely sensi-

tive and respond in very different ways to inputs which are very similar.

Such responses are often found in human psychophysics. The perception of many speech distinctions, for instance, is characterized by an apparent insensitivity to acoustic differences which fall within a phonetic category, accompanied by great sensitivity to small differences which straddle a category boundary. Humans may perceive a difference of 5/1000 of a second as signalling a category distinction (e.g., the sound [b] vs. [p]) in some cases; whereas in other regions of the input dimension a 5/1000 goes undetected.

Finally, the nonlinearity extends the computational power of networks in an important way. Recall from Chapter 2 that there are problems which require that a network have additional layers (e.g., those which are nonlinearly separable). Why not use linear units? The reason is that it is the nature of linear networks that for any multi-layer network composed of linear units, one can devise a simpler linear network composed of just two layers. So the additional layers do no work for us. Since we need additional layers ("real" layers), the only way to get them is to have the units be nonlinear. That way we can have multi-layer networks which cannot be reduced to two-layer systems.

Interactions: A case study in nonlinearity

The simple nonlinearity introduced in node activation functions turns out to have even more interesting and useful properties when we consider how it manifests itself in networks which are learning complicated tasks. What we find is that behaviors which at first glance might appear to arise because of multiple mechanisms can actually be accounted for as well (or better) by single but nonlinear networks.

Consider the question of stimulus frequency and consistency. To make this example we turn to a model developed by Plaut, McClelland, Seidenberg, and Patterson (1994). Plaut and his colleagues have been interested in the question of what mechanisms underlie reading. One proposal that has been made is that skilled readers use two independent processing routes when going from

written form to pronunciation. One route involves the application of grapheme-phoneme correspondence rules (e.g., "c followed by i or e is soft"). A second route is also hypothesized. This second pathway allows readers to look words up in their mental lexicons and retrieve pronunciations directly. Both routes seem to be needed. The first route is useful in reading novel words which can't be found in the reader's mental dictionary; the second route is needed to read words with exceptional pronunciations which do not accord with the rules.

Certain empirical facts would seem to lend support to this analysis of reading as involving two routes. One important fact is that the speed with which subjects can name words (presented briefly on a visual display) is highly correlated with word frequency. Overall, frequent words are read more rapidly than low frequency words. However, this effect interacts with the regularity of a word's pronunciation. The pronunciation of regular word is more or less independent of frequency, whereas pronunciation latencies are much faster for frequent irregular words than infrequent irregular words.

The apparent insensitivity of regular forms to frequency has been hypothesized to occur because the regulars are produced by rule, whereas the irregulars are stored directly and their retrieval time is facilitated by frequent access. (A similar conclusion has been drawn in the case of English verbs, since a frequency/regularity similar to that found in reading has been observed: subjects are faster at producing the past tense of regular English verbs ("walked" given the prompt "walk") compared with irregular verbs ("went", given "go").) Thus, there are two reasons why one might believe that reading involves two distinct mechanisms: (1) the ability to read non-words (which cannot be stored in the lexicon), and (2) the effect of word frequency on the reading of irregulars, coupled with the lack of such an effect on regulars.

In fact, such interactions can also be produced by a single mechanism which has nonlinear characteristics. Plaut et al., building on earlier work by Seidenberg and McClelland (1989) trained several networks to read. (That is, the networks produced the correct phonological output for graphemic input.) Their network, shown in

Figure 4.24, was trained on 2998 monosyllabic words, with the training of each word weighted by its frequency of occurrence. After training, the network was tested on several sets of nonwords. Performance closely resembled that reported for human subjects reading the same words.

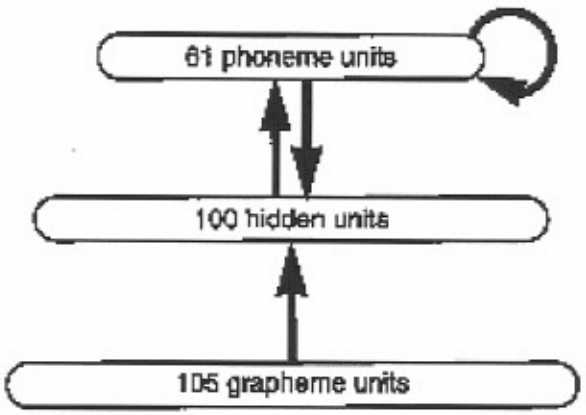


FIGURE 4.24 The architecture of the Plaut et al. (1994) reading network. (After Plaut et al., Figure 12.)

The network also showed systematic differences in the speed with which it processed words, depending on the frequency of a word's occurrence. However, this frequency advantage disappeared in the case of regular words. More interestingly, there was also an effect for consistency. As Glushko (1979) first noted, a word can be regular in its pronunciation but resemble another word which is irregular. "have" and "mint" are pronounced in a way which follows the normal rules, so they are regular. However, there are orthographically similar words "gave" and "pint" which are irregular. The presence of these irregulars in the neighborhood of the "save" and "mint" mean that the latter are regular but inconsistent. "slap," on the other hand, is not only regular but has no irregular neighbors. So it is regular and consistent.

The network, like skilled readers, also shows differences in processing latency for these two classes of words. Regular but inconsistent words are processed more slowly than regular consistent words. Furthermore, word frequency has no effect on regular consistent words, but regular inconsistent words show a pattern which is an attenuated version of what happens with irregulars.

As Plaut et al. point out, all three effects (frequency, regularity, and consistency) have a common source in the network. The weight changes which occur when a network is trained are in proportion to the sources of error. Because more frequent items are by definition encountered more often during training, these items contribute a larger source of error. So the network's weights are altered to accommodate them more than infrequent words. Furthermore, regular words all induce similar changes. This means that the frequency of a given regular form is of less consequence to regulars. The network will learn a rare regular word anyway since it resembles all the other regulars and benefits from the learning they induce. Consistency is simply a milder version of this. A word may be regular, but if there are not many other regular forms which resemble it (or, worse, there are high frequency irregulars which do), then will be learned less well, and learning will depend more on its individual frequency.

These factors are additive, but because the network is nonlinear, the cumulative consequence of their contribution asymptotically diminishes. This follows directly from the sigmoidal shape of the activation function of nodes. A schematic representation of this is shown in Figure 4.25.

Imagine that a word receives 1 or 3 units of activation, depending on its frequency of occurrence, and that regularity contributes an additional 1 units of activation. So we might imagine that High Frequency Regulars would receive the most (4 units of input), Low Frequency Exceptions the least (1 unit of input), and High Frequency Exceptions and Low Frequency Regulars intermediate amounts (3 and 2 units of input respectively). We see in the top panel of Figure 4.25 that although the contribution of frequency and regularity is additive, because the output of nodes in the network is a nonlinear function of their input, the difference between high and

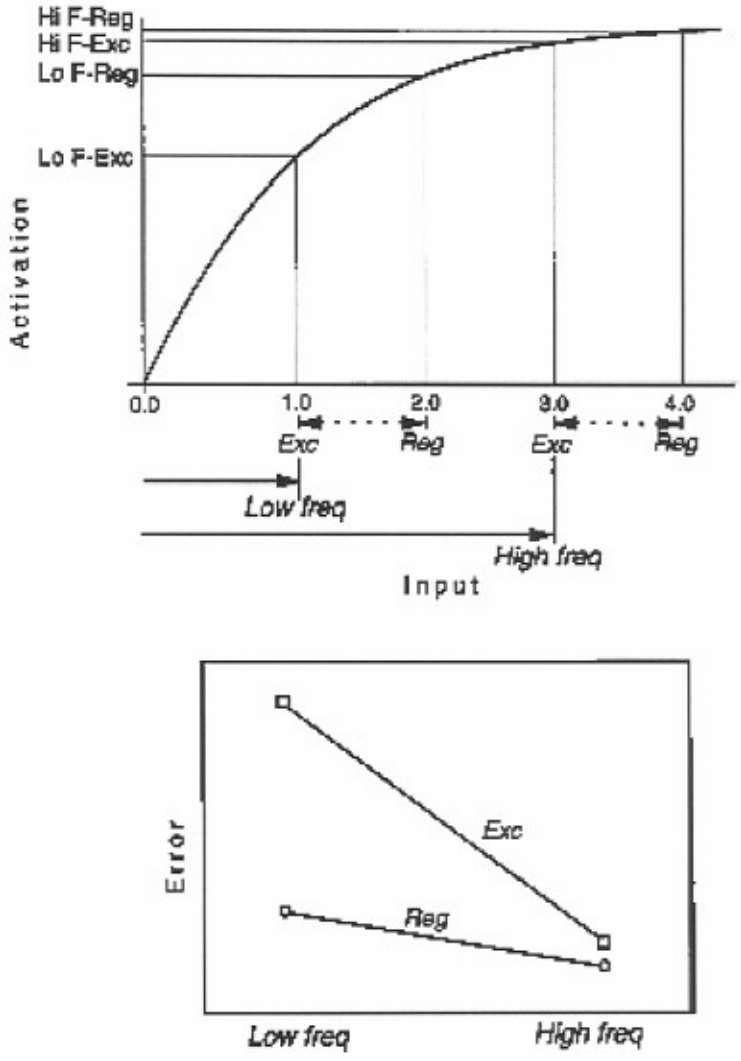


FIGURE 4.25 The nonlinear outcome of additive effects of word frequency and regularity of pronunciation. Top panel: Frequency and regularity each contribute the same amount to processing of both regular and irregular words, but the effect of frequency is less in the case of regular words because of the nonlinear activation function. Bottom panel: The differences in activation translate into error differences in performance. (After Plaut et al., 1994, Figures 7 and 9.)

low frequency regulars is much less than between high and low frequency irregulars. This translates into performance differences shown in the bottom panel of the figure; word frequency results in relatively minor differences in performance in the case of regulars, but large differences in the case of exceptions.

Dynamics

The previous example shows how a single mechanism with nonlinear characteristics can produce behaviors which closely resemble those expected from multiple mechanisms. What about change? This is, after all, the critical issue of this chapter. Let us now turn to dynamics.

There are two contexts in which one finds dynamics in networks. The first has to do with the pattern of change which occurs as a network learns. This comes closest to the sense in which we talked about dynamics in the first part of this chapter. Secondly, in certain kinds of networks (those with recurrent connections), we can also find behaviors (after learning) which exhibit dynamics.

Dynamics in learning

At first it might seem odd to talk about dynamics in feedforward networks. These are networks in which the output is a direct function of the input. When we look at their instantaneous behavior, these are "one-shot" networks. Indeed, in this sense, they have no dynamics.

The dynamics arise when these systems are changed in response to training. The dynamics now has to do with the changes in weight space, which are (in a back-propagation network) governed by the equation given in Chapter 2 as Equation 2.5, reproduced below.

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}} \qquad \text{(EQ 4.39)}$$

This equation tells us how to change the weights in order to minimize error. It says that the change in the weight between units i and j should be changed in proportion to the product of a small learning rate (η) times the derivative of the network's error with respect to the weight connecting units i and j. The derivative tells us how changes in the error are related to changes in the weights; if we know this then we know how to learn: Change the weights in such a way as to reduce error. We also see from this equation that because the change in weight is depending on the existing weight, this is a system which has the potential for interesting dynamics.

We might have supposed, for example, that because network learning is incremental (in the sense that we employ the same learning rule at all points in time, making small adjustments in accord with current experience) the change in performance will be incremental in some straightforward manner. But consider what happens when a network learns the logical XOR function. If we measure the performance error over time, sampling the mean error every 100 patterns, we see the expected monotonic drop in performance. This is illustrated with the heavy line in Figure 4.26.

However, if we examine error more precisely, looking at the performance on each of the four patterns which comprise XOR (the lighter lines in Figure 4.26), we find that some of the patterns exhibit nonmonotonic changes in performance. The most dramatic instance of this is the error associated with the pattern 11. Initially, error on this pattern actually increases while the error on the other patterns decreases. During this phase, the network has found a temporary solution which assumes the function underlying the data is logical OR. Because this solution works for three of the four patterns, the overall drop in error is sufficient that the network tolerates increasing error on the one pattern for which OR does not fit (an input of 11 produces an output of 1 for OR rather than the output of 0 which is correct for XOR). It is only when the price paid by this outlier is sufficiently great that the network backs off from this solution and finds another. There is a small trade-off during the transition period; as the error for 11 drops, the error for another pattern (01) temporarily increases. But eventually the network finds a setting of weights which accommodates all the patterns.

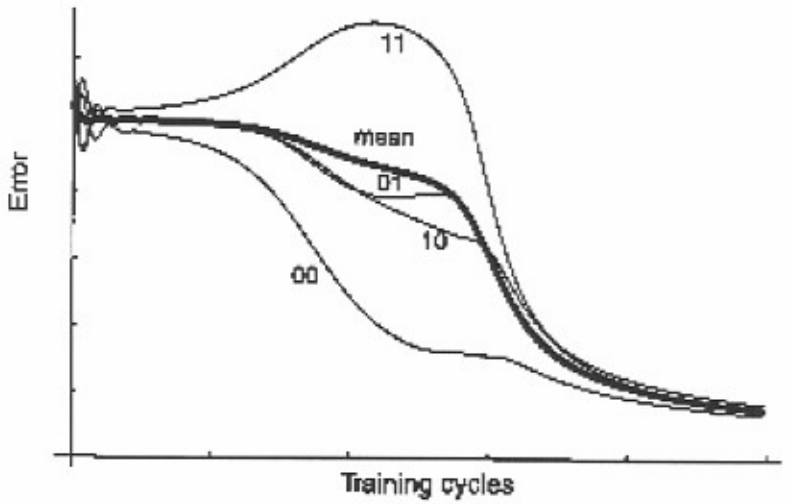


FIGURE 4.26 Error plot during learning the XOR function. The heavy line shows the mean error calculated over successive batches of 100 training cycles. The lighter lines show the error produced by each of the four patterns which comprise XOR. Two of the patterns (00 and 10) have monotonically decreasing error; the other two (01 and 11) temporarily increase error before a solution is found which accommodates all four patterns.

This sort of interaction, in which performance on different components of a complex task varies over time, has obvious parallels with phenomena in child development. For example, many children display nonmonotonic changes in performance in the course of learning the morphology of the English past tense. The classic story, as detailed in Chapter 3 (Berko, 1958; Ervin, 1964) is that children begin by producing both regular ("talked") and irregular forms ("came") correctly. As time passes, children continue to produce the regulars correctly but begin to make errors on the irregulars, regularizing forms which they used to produce correctly as irregulars (e.g., "comed"). A reasonable interpretation is that the first phase reflects the rote memorization of verbs, whereas the second phase occurs when children begin the process of rule abstraction. This rule should apply only to the regulars (by definition); over-extension

occurs because the children have not yet fine-tuned the conditions of the rule's application.

As we pointed out in Chapter 3, both the data and the interpretation have been called into question (Bybee & Slobin, 1982; Marchman, 1988; Marcus et al., 1992; Plunkett & Marchman, 1991, 1993). The U-shape is in fact composed of many "micro-Us", and correct irregulars often co-exist in free variation with the incorrect regularized form. Furthermore, regulars themselves are sometimes "irregularized." Plunkett and Marchman's simulation, which we described in Chapter 3, shows similar patterns of nonmonotonic change and irregularizations. This nonmonotonicity arises for essentially the same reasons it does in the XOR simulation, not because the network has suddenly moved from a strategy of memorization to rule abstraction.

Readiness, stages of learning, and bifurcations

Another phenomenon which is often observed with children is the sudden transition from what appears to be insensitivity to input a stage where the child seems to be extraordinarily sensitive to new data. The abruptness with which this occurs suggests that some internal event has occurred which has moved the child from a prior stage in which she was not able to make use of information, to a new stage in which the information can now have an effect. Until the child is ready for it, the information is simply ignored.

One example of stages of learning, discussed in Chapter 3, is the sequence of phases children go through in the course of learning to balance weights on a beam. As the McClelland and Jenkins simulation demonstrates, such stages of learning can be readily reproduced in a network which uses an incremental learning strategy. But of course simulating one black box with another does us little good. What we really want is to understand exactly what the underlying mechanism is in the network which gives rise to such behavior. We outlined the rudiments of an explanation in Chapter 3. We are now in a position to flesh the explanation out in greater deal.

To do this we shall use an example of a behavior in child development which in some ways resembles the balance beam experi-

ment. When children first learn the principle of commutativity (i.e., the principle that $x+y=y+x$), they demonstrate a limited ability to generalize beyond the sums that were used to teach them the principle. Children thus can reliably report that 11+12 and 12+11 yield the same sum, even though they may never have seen this specific example. The ability to generalize is fairly limited, however. Children at this stage will fail to recognize that sums such as 645,239,219+14,345 (which differ greatly in magnitude from those they encounter as examples of the principle) are the same as 14,345+645,239,219. In Karmiloff-Smith's Representational Redescription model (Karmiloff-Smith, 1992a), children are in the Implicit phase. At a later stage, children do achieve complete mastery of the principle and are able to apply it to any arbitrary sum, no matter what the magnitude. In Karmiloff-Smith's terms, children have now moved to the stage of Explicit-1 (and subsequently move to Explicit-2 and then to Explicit-3, when they can verbally articulate their knowledge).

The question is, what sort of mechanisms might be responsible for what seems to be qualitatively different sorts of knowledge, and how can we move from one phase to the next? The character of the knowledge seems quite different in the two phases, and it is not clear how a network might simulate such drastic transformations. What we shall see, however, is that there is a far simpler explanation for the change in network terms than we might have imagined.

We shall build a model of a somewhat different mathematical task. Instead of commutativity, we chose the problem of detecting whether a sequence of 1s and 0s is odd or even (counting the number of 1s in the string). If the sequence is odd, we shall ask a network to output a 1. If the sequence is even, the correct output is 0. Furthermore, we shall ask the network to give us a running report, so that if the network sees a sequence of five inputs, it should tell us after each input whether the sequence to that point is odd or even. (This is a temporal version of the *parity* task.) Given the input string

1 1 0 1 0

the correct outputs would be

1 0 0 1 1 (*odd, even, even, odd, odd*).

Since this task involves a temporal sequence, we shall use the

simple recurrent network architecture shown in Figure 4.27

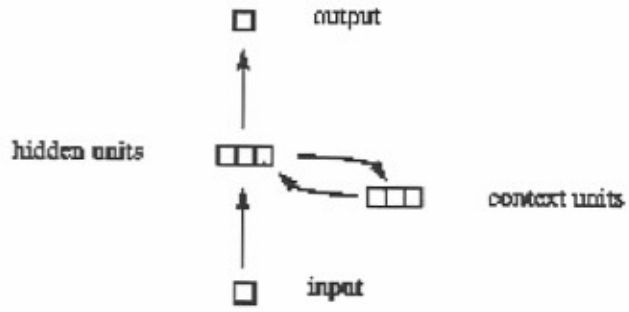


FIGURE 4.27 Simple recurrent network architecture used to learn the odd/even task. Context units store the state of the hidden units from the prior time step.

We train the network on a series of short sequences, two to five bits in length. At each stage during a sequence the network outputs a 1 just in case the sequence up to that point contains an odd number of 1s; otherwise, it should output a 0. After a sequence is done, the network context units are reset to 0.5 and the next sequence is input.

To evaluate the network's performance during the course of training, we ask it to do the odd/even task on a string of 100 1s. The length of this string is an order of magnitude greater than any of the sequences encountered in training. What we are interested in is just how far into the sequence the network can go before it fails. If the network generalizes perfectly, then the output will be an alternating sequence of 1010... (that is, *odd, even, odd, even*...). Graphically, such output would be a zig-zag pattern with 100 peaks and 100 valleys.

In Figure 4.28 we see the results of tests performed at several stages in learning. The top panel shows the network's outputs to the test string after 15,000 inputs. The network shows the correct pattern of alternating *odd, even* responses for four successive inputs. Since the training data contained sequences which varied randomly

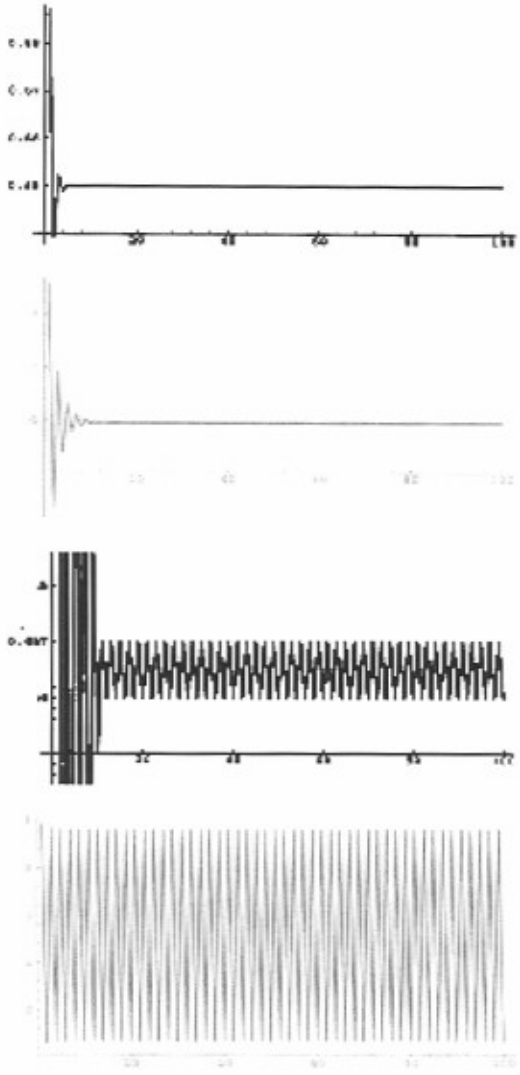


FIGURE 4.28 Performance of the simple recurrent network while learning the odd/even task at several points in time (measured in cycles). Generalization occurs at 17,999 cycles but is limited. One learning cycle later, however, the network is able to extend the generalization indefinitely.

between two and six inputs in length, the network has not yet mastered the task even for lengths it has seen.

After additional training (the panel marked "after 17,999 inputs"), we find that the network generalizes to the extent that it can give the correct answer for the first 13 inputs from the test sequence. Thus the network has generalized beyond its experience (Karmiloff-Smith's Implicit I phase). The network has not yet learned the principle of odd/even in a fully abstract manner, because the network is unable to give the appropriate response beyond 13 inputs.

At least, this is the state of affairs after 17,999 training cycles. What is remarkable is the change which occurs on the very next training cycle. After one additional input, the network's performance changes dramatically. The network now is able to discriminate odd from even. The magnitude of the output is not as great for longer sequences, so we might imagine that the network is not fully certain of its answer. But the answer is clearly correct and in fact can be produced for sequence of indefinite length. The network seems to have progressed to Karmiloff-Smith's Explicit-1—in a single trial! Subsequent training simply makes the outputs more robust.

We might say therefore that early in training, the network was not able to learn the odd/even distinction; and that at 17,999 cycles the network has moved into a stage of readiness such that all that was needed was one further example for it to learn the distinction fully. Yet we know that there are no abrupt endogenous maturational changes which occur in the network. And we know that the learning procedure is the same throughout, namely, gradient descent in error space through small incremental changes in weights. So why does performance change so dramatically?

To explain why this occurs, it will be useful to simplify matters even further. Let's consider the case where we have a single node, as shown in Figure 4.29. We will give it a constant input, labeled the bias, and a recurrent weight, labeled w_r. Finally, we inject some initial input to the node, and then we let the node activate itself for some number of iterations before we look at its final output.

The point of this example will be to look at what kinds of dynamics develop as the node recycles activation back to itself over

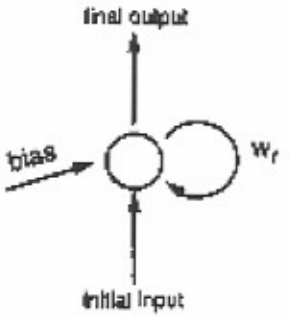


FIGURE 4.29 A one-node network which receives an initial input; the input is then removed, and processing consists of allowing the network to fold its activation back on itself through the recurrent weight. After some number of iterations, we examine the output.

the recurrent weight. We can see that for the odd/even task it is important for the network to remember what an initial input was, since this can make the difference between a string having an odd vs. an even number of 1s. The recurrence in the network allows nodes to "remember" their prior activations. When we look at the dynamics of this simple system, we are therefore looking at the network's memory.

One problem we notice immediately is that the sigmoidal activation function will continuously map the node's activation back to the interval (0.0, 1.0), squashing it in the process. For example, suppose the recurrent weight has a value of $w_r = 1.0$ and there is a constant bias of $b = -0.5$. Then if we start with the node having an initial activation of 1.0, on the next cycle the activation will be as given in Equation 4.40:

$$a(t+1) = \frac{1}{1+\exp(-(a(t)-0.5))} \qquad \text{(EQ 4.40)}$$

or 0.62. If this diminished value is then fed back a second time, the next activation will be 0.53. After 10 iterations, the value is 0.50—and it remains at that level forever. This is the mid-range of the node's activation. It would appear that the network has rapidly lost the initial information that a 1.0 was presented.

This behavior, in which a dynamical system settles into a resting state from which it cannot be moved (absent additional external input) is called a *fixed point*. In this example, we find a fixed point in the middle of the node's activation range. What happens if we change parameters in this one-node network? Does the fixed point go away? Do we have other fixed points?

Let's give the same network a recurrent weight $w_r = 10.0$ and a bias $b = -5.0$. Beginning again with an initial activation of 1.0, we find that now the activation stays close to 1.0, no matter how long we iterate. This makes sense, because we have much larger recurrent weight and so the input to the node is multiplied by a large enough number to counteract the damping of the sigmoidal activation function. This network has a fixed point at 1.0. Interestingly, if we begin with an initial activation of 0.0, we see that also is a fixed point. So too is an initial value of 0.5. If we start with initial node activations at any of these three values, the network will retain those values forever.

What happens if we begin with activations at other values? As we see in Figure 4.30, starting with an initial value of 0.6 results over the next successive iterations in an increase in activation (it looks as if the node is "climbing" to its maximum activation value of 1.0). If we had started with a value of 0.4, we would have found successive decreases in activation until the node reached its fixed point close 0.0. Configured in this way, our simple one-node network has three stable fixed points which act as basins of attraction. No matter where the node begins in activation space, it will eventually converge on one of these three activation values.

The critical parameter in this scheme is the recurrent weight (actually, the bias plays a role here as well, although we shall not pursue that here). Weights which are too small will fail to preserve a desired value. Weights which are too large might cause the network to move too quickly toward a fixed point. What are good weights?

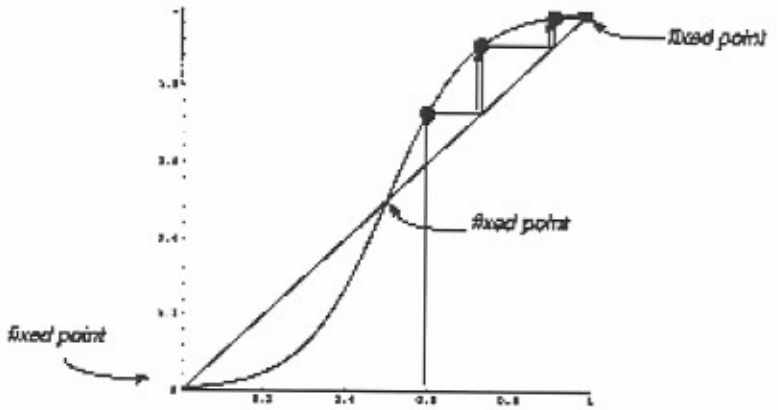


FIGURE 4.30 If a recurrent unit's initial activation value is set to 0.6, after successive iterations the activation will saturate close to 1.0. An initial value of 0.5 will remain constant; an initial value of less than 0.5 will tend to 0.0 (assumes a bias of -5.0 and recurrent weight of 10.0).

Working with a network similar to the one shown in Figure 4.29, we can systematically explore the effects of different recurrent weights. We will look to see what happens when a network begins with different initial activation states and is allowed to iterate for 21 cycles, and across a range of different recurrent weights. (This time we'll use negative weights to produce oscillation; but the principle is the same.) Figure 4.31 shows the result of our experiment.

Along the base of the plot we have a range of possible recurrent weights, from 0.0 to -10.0. Across the width of the plot we have different initial activations, ranging from 0.0 to 1.0. And along the vertical axis, we plot the final activation after 21 iterations.

This figure shows us that when we have small recurrent weights (below about -5.0), no matter what the initial activation is (along the width of the plot), we end up in the middle of the vertical axis with a resting activation of 0.5. With very large values of weights, however, when our initial activation is greater than 0.5 (the portion of the surface closer to the front), after 21 iterations the final

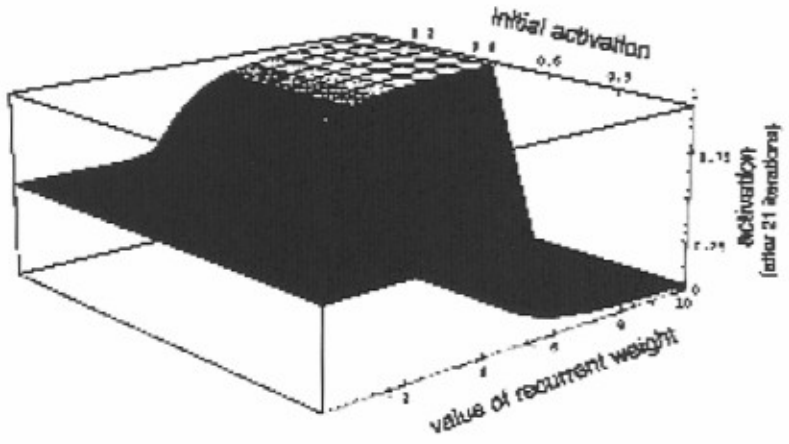


FIGURE 4.31 The surface shows the final activation of the node from the network shown in Figure 4.29 after 21 iterations. Final activations vary, depending on the initial activation (graphed along the width of the plot) and the value of the recurrent weight (graphed along the length of the plot). For weights smaller than approximately -5.0, the final activation is 0.5, regardless of what the initial activation is. For weights greater than -5.0, the final activation is close to 0.0 when the initial activation is above the node's mid-range (0.5); when the initial activation is below 0.5, the final activation is close to 1.0.

value is 0.0 (because the weight is negative and we iterate an odd number of times, the result is to be switched off; with 22 iterations we'd be back on again). If the initial activation is less than 0.5, after 21 iterations we've reached the 1.0 fixed point state. The important thing to note in the plot, however, is that the transition from the state of affairs where we have a weight which is too small to preserve information to the state where we hold on (and in fact amplify the initial starting activation) is relatively steep. Indeed, there is a very precise weight value which delineates these two regimes. The abruptness is the effect of taking the nonlinear activation function and folding it back on itself many times through network recurrence. This phenomenon (termed a bifurcation) occurs frequently in nonlinear dynamical systems.