Chapter 9
Stochastic Learning Models

THE BASIC ASSUMPTION underlying this class of models, developed by Estes (1950), Bush and Mosteller (1955), is that the probability of a response changes by an increment (positive or negative) which is a linear function of the response probability, namely

$$p(t+1) = \alpha p(t) + (1-\alpha)\lambda, \quad 1 \geq \alpha, \quad \gamma \geq 0. \qquad (83)$$

The right-hand side of (83) is viewed as an operator acting upon $p(t)$. Clearly, this operator is determined by two parameters, α and λ.[27] The magnitude of these parameters depends on the outcome of the response. In particular, if the outcome is "rewarding" to the subjects we may have $\lambda = 1$, in which case $p(t)$ will tend toward 1 (certainly) as the response is repeated. That is to say, the response will be ultimately fixated. If the outcome is "punishing," we may have $\lambda = 0$, in which case the response will become extinguished. Intermediate values of λ are also possible, in which case the probability of the response in question will tend to some limiting value ($0 \leq p \leq 1$). The rate of learning is reflected in α. The smaller the magnitude of this parameter, the more quickly $p(t)$ will approach its limiting value. In the extreme cases, if $\alpha = 0$, the limiting value of λ is reached at once; if $\alpha = 1$, $p(t)$ remains constant.

If there are two choices of response, we need to consider only one probability, for the other is clearly the complement. However, if each choice leads to a different outcome, there will be generally two operators, each associated with one of the outcomes. Both operators

may be affecting $p(t)$ in the same direction. For example, both the success associated with the "correct" choice and the failure associated with the "wrong" choice may cause $p(t)$, the probability of the correct response to increase, but the rates of increase may be different. For example, the effect of rewarding the correct response may be greater than the effect of punishing the wrong response or vice versa.

The outcome of the subject's response may not be unique. Suppose for example that each of his responses is sometimes rewarded and sometimes punished, the probabilities of the rewards and punishments being fixed for each response. If the rewards can be reduced to a common utility scale, then the "rational solution" of the problem (essentially a decision under risk) is always to make the choice associated with the greatest expected gain. However, the subject (especially if he is not human) may be ignorant of rational decision theory or the rewards may not be reducible to a common utility scale. In this case, it is conceivable that neither response will be either fixated or extinguished.

A subject playing Prisoner's Dilemma has two choices of response (C or D). Each of these choices may result in two outcomes. For player 1, C may result in R (if CC occurs) or in S (if CD occurs). For player 2, C may result in R (if CC occurs) or in S (if DC occurs), and similarly for D. Associated with each player, therefore, we have four operators acting on the probability of cooperative response. Thus, assuming $i = 1$ or 2,

$$C_i(t+1) = \alpha_i^{(1)} C_i(t) + (1 - \alpha_i^{(1)})\lambda_i^{(1)}, \quad \text{if } CC \text{ occurs;} \tag{84}$$

$$C_i(t+1) = \alpha_i^{(2)} C_i(t) + (1 - \alpha_i^{(2)})\lambda_i^{(2)}, \quad \text{if } CD \text{ occurs, when } i = 1 \text{ or if } DC \text{ occurs, when } i = 2; \tag{85}$$

$$C_i(t+1) = \alpha_i^{(3)} C_i(t) + (1 - \alpha_i^{(3)})\lambda_i^{(3)}, \qquad (86)$$

if DC occurs, when $i = 1$ or
if CD occurs, when $i = 2$;

$$C_i(t+1) = \alpha_i^{(4)} C_i(t) + (1 - \alpha_i^{(4)})\lambda_i^{(4)}, \qquad (87)$$

if DD occurs.

Note that C is affected whatever the outcome. This is in consequence of the fact that whenever D is reinforced or inhibited, C is also necessarily inhibited or reinforced, since $C = 1 - D$.

This model, then, contains eight parameters associated with each player or sixteen associated with each pair.

The model can be simplified if we make some commonsense assumptions based on observations. For example, we suspect that the CC response is fixated if it is repeated sufficiently many times (at least in the Matrix conditions). Thus we can set $\lambda_i^{(1)} = 1$. Similarly, in view of the virtual extinction of the unilateral cooperative responses, we can set $\lambda_i^{(2)} = 0$.

Successful defection also inhibits the cooperative response. Hence we can set $\lambda_i^{(3)} = 0$. With respect to $\lambda_i^{(4)}$, we are not sure. On the one hand DD is a punishing state, so that one might suppose that it would inhibit D (reinforce C) whenever it occurs. On the other hand in the special context of Prisoner's Dilemma, DD may be self-enhancing, since it indicates to each player that the other is not to be trusted. We therefore should leave the magnitude of $\lambda_i^{(4)}$ undetermined a priori.

This simplification leaves us with ten parameters, namely $\alpha_i^{(j)}$ $(i = 1, 2; j = 1, \ldots 4)$, $\lambda_1^{(4)}$ and $\lambda_2^{(4)}$.

Stochastic Learning Model Superimposed upon the Four-State Markov Chain

As has already been pointed out, the basic idea of the stochastic learning model is that probabilities of

responses become modified in a specific way as a result of the responses themselves, whereby the outcome resulting from the response determines the operator acting on the probability in question. In the previous section we have assumed that the probabilities being modified are the unconditional probabilities of cooperative response. We can, however, assume another point of view, namely, that the probabilities being modified are the conditional propensities. If we suppose these to be the state-conditioned propensities x, y, z, and w, then our stochastic learning model expands into the following system of equations:

$$x_i(t+1) = \alpha_{i1}^{(jk)} x_i(t) + (1 - \alpha_{i1}^{(jk)})\lambda_{i1}^{(jk)}, \tag{88}$$

$$y_i(t+1) = \alpha_{i2}^{(jk)} y_i(t) + (1 - \alpha_{i2}^{(jk)})\lambda_{i2}^{(jk)}, \tag{89}$$

$$z_i(t+1) = \alpha_{i3}^{(jk)} z_i(t) + (1 - \alpha_{i3}^{(jk)})\lambda_{i3}^{(jk)}, \tag{90}$$

$$w_i(t+1) = \alpha_{i4}^{(jk)} w_i(t) + (1 - \alpha_{i4}^{(jk)})\lambda_{i4}^{(jk)}, \tag{91}$$

$(i = 1, 2; j, k = 1, \ldots 4)$.

We recall that the operator is determined by the outcome resulting from the last response. In this model the outcome is associated not with a state (like *CC*) but with a transition from one state to another (as $CC \rightarrow CD$). There are sixteen such transitions, and they are designated in our equations by the double superscripts (jk), where both indices range over the four states.

Thus the system (88)-(91) involves 256 parameters. However, as in the preceding case, we can make some simplifying assumptions. Suppose, for example, the propensity x is affected by experience only if the corresponding transition has just occurred, i.e., only if the transition has been from *CC* to one of the four states. This is tantamount to setting $\alpha_{ij}^{(jk)} = 1$ for all $j \neq 1$.

We can therefore write

$$x_1(t+1) = \alpha_{11}^{(k)} x_1(t) + (1 - \alpha_{11}^{(k)})\lambda_{11}^{(k)} \quad (k = 1 \ldots 4), \tag{92}$$

$$x_2(t+1) = \alpha_{21}^{(k)} x_2(t) + (1 - \alpha_{21}^{(k)})\lambda_{21}^{(k)} \quad (k = 1 \ldots 4). \tag{93}$$

We can make a similar assumption with regard to the other propensities. This simplification reduces the number of parameters from 256 to 64. We can further reduce the number to 32 if we assume as we did previously that the propensities tend to be either fixated or extinguished, i.e., the parameters λ are either equal to 1 or to 0.

For example, if *CC* were followed only by *CC*, we might expect this transition to be fixated. According to this assumption, we would set $\lambda_{11}^{(1)} = 1$.

On the other hand, if *CC* were always followed by *CD*, we would expect the *C* response following *CC* to be extinguished, both because it would be punishing for the cooperator and because the *D* response would be rewarding to the defector. (Note that the "higher order effects" are being ignored, for example the effect of retaliation for defecting from the *CC* response.) If *CC* is followed by *DD*, this might tend to fixate x (the cooperative response following *CC*), since *DD* is punishing for both players.

If we are guided only by these considerations, i.e., whether the immediate response is punished or rewarded, we may reduce all the λ's to either 1 or 0 on the basis of the following assumptions:

(1) If the payoff was positive and remained the same in the succeeding play, the corresponding propensity tends toward fixation.

(2) If the payoff was improved, the corresponding propensity tends to fixation.

(3) If the payoff was worsened, the corresponding propensity tends to extinction.

(4) If the payoff was negative and remained the same, the corresponding propensity tends to extinction.

All of these seem to us to be reasonable hypotheses with two exceptions. According to (4) above, the *D* responses following *DD* when transition $DD \rightarrow DD$

has occurred should be inhibited, since the payoff is negative and remains constant. However, this assumption seems to be contradicted by some of our data. In many cases prolonged sequences of *DD* seem to enhance the fixation of that response (cf. Chapter 11). The other exception is the transition from *DC* to *CC* from the standpoint of player 1. In this transition, player 1's payoff is reduced from T to R. However, assuming some insight into the nature of the game, the shift from D to C by player 1 might have been motivated by "repentance," i.e., by a decision to cooperate in response to the other's unilateral cooperation. If the outcome is *CC*, it may well be interpreted as rewarding, not punishing, by player 1, in spite of the reduced payoff. If we leave both of these cases open, then the parameters $\lambda_{i3}^{(1)}$ and $\lambda_{i4}^{(4)}$ cannot be set equal to zero as our assumption (3) above implies and must be left as free parameters. There are thus thirty-six parameters remaining in the simplified stochastic learning model, in which the state-conditioned propensities x, y, z, and w are subject to learning.

The test of this model would involve prodigious calculations. It would be ill-advised to undertake this work in the early stages of constructing a theory. A more promising strategy is to test some models which purport to describe gross features of the situation instead of attempting to determine the fine structure of the process.