JOURNAL OF MATHEMATICAL PSYCHOLOGY: 5, 61–101 (1968)

Some Convergence Theorems for Stochastic Learning Models with Distance Diminishing Operators[1]

M. FRANK NORMAN

University of Pennsylvania, Philadelphia, Pennsylvania 19104

A broad mathematical framework is considered that includes stochastic learning models with distance diminishing operators for experiments with finite numbers of responses and simple contingent reinforcement. Convergence theorems are presented that subsume most previous results about such models, and extend them in a variety of ways. These theorems permit, for example, the first treatment of the asymptotic behavior of the general linear model with experimenter-subject controlled events and no absorbing barriers. Some new results are also given for certain two-process discrimination learning models and for models with finite state spaces.

1. INTRODUCTION

Suppose that a subject is repeatedly exposed to an experimental situation in which various responses are possible, and suppose that each such exposure or trial can alter the subject's response tendencies in the situation. It is assumed that the subject's response tendencies on trial n are determined by his *state* S_n at that time. The set of possible states is denoted S and called the *state space*. The effect of the nth trial is represented by the occurrence of a certain *event* E_n. The set of possible events is denoted E and referred to as the *event space*. The quantities S_n and E_n are to be considered random variables. The corresponding small letters s_n and e_n are used to indicate particular values of these variables, and, in general, s and e denote elements of the state and event spaces, respectively.

To represent the fact that the occurrence of an event effects a change of state, with each event e is associated a mapping $f_e(\cdot)$ of S into S such that, if $E_n = e$ and $S_n = s$, then $S_{n+1} = f_e(s)$. Thus

H1. $$S_{n+1} = f_{E_n}(S_n)$$

for $n \geqslant 1$. The function $f_e(\cdot)$ will be called the *operator for the event e* or simply an *event operator*. Throughout the paper it is assumed that

H2. *E is a finite set.*

[1] This research was supported in part by National Science Foundation grant NSF GU 1923-3.

It is further supposed that the learning situation is memory-less and temporally homogeneous, in the sense that the probabilities of the various possible events on trial n depend only on the state on trial n, and not on earlier states or events, or on the trial number. That is, there is a real valued function $\varphi_{.}(\cdot)$ on $E \times S$ such that

H3. $$P_s(E_1 = e_1) = \varphi_{e_1}(s),$$

and

$$P_s(E_{n+1} = e_{n+1} \mid E_j = e_j\, , 1 \leqslant j \leqslant n) = \varphi_{e_{n+1}}(f_{e_1 \ldots e_n}(s)),$$

for $n \geqslant 1$, where

$$f_{e_1 \ldots e_n}(s) = f_{e_n}(f_{e_{n-1}}(\cdots(f_{e_1}(s)))). \tag{1.1}$$

Throughout the paper, state subscripts on probabilities and expectations are initial states, that is, values of S_1.

Two examples will be discussed in Sec. 3: a linear model for ordinary two-choice learning, and a two-stage linear discrimination learning model. In the first linear model, the state is the probability of one of the responses, so $S = [0, 1]$. In the linear discrimination learning model the state is a pair of probabilities that determine, respectively, the "response" probabilities at the two stages. Thus, $S = [0, 1] \times [0, 1]$. In these examples each event involves the subject's overt response (suitably coded), the observable outcome of that response (i.e., the experimenter's response), and, sometimes, a hypothetical occurrence that is not directly observable (e.g., the state of attention on a trial). The force of assumption H3 for the experimenter is to limit reinforcement schedules to those in which the outcome probabilities depend only on the immediately preceding response, that is, to simple contingent schedules.

The research reported in this paper is directed toward understanding the asymptotic behavior of the stochastic processes $\{S_n\}$ and $\{E_n\}$ for a class of models with distance diminishing event operators defined below by imposing additional restrictions on the functions f and φ. This class generalizes the familiar linear models, and the latter provide much of the motivation for the axioms for the former.

To discuss "distance diminishing" event operators, it is necessary to assume that S is a metric space with respect to some metric d. A formulation in terms of Euclidean space and root-sum-square distance would yield sufficient generality to cover the linear models of Sec. 3. Such a formulation would, however, restrict generality without any redeeming simplification. Moreover, a treatment in terms of general metric spaces highlights those aspects that are crucial to the theory. For these reasons it is assumed only that

H4. *(S, d) is a metric space.*

The reader who prefers a Euclidean setting can easily specialize most of what follows to suit his preferences.

The next assumption is suggested by the linear examples of Sec. 3:

H5. *(S, d) is compact.*

The remaining hypotheses are most easily stated in terms of the following notations. If ψ and g are mappings of S into the real numbers and into S, respectively, their maximum "difference quotients" $m(\psi)$ and $\mu(g)$ are defined by

$$m(\psi) = \sup_{s \neq s'} \frac{|\psi(s) - \psi(s')|}{d(s, s')}, \tag{1.2}$$

and

$$\mu(g) = \sup_{s \neq s'} \frac{d(g(s), g(s'))}{d(s, s')}, \tag{1.3}$$

whether or not these are finite. If, for instance, S is a real interval (with $d(s, s') = | s - s' |$) and ψ is differentiable throughout S, $m(\psi)$ is the supremum of $| \psi'(s)|$. The hypothesis

H6. $m(\varphi_e) < \infty$ *for all* $e \in E$

is a mere regularity condition. The next two assumptions, however, are genuinely restrictive:

H7. $\mu(f_e) \leqslant 1$ *for all* $e \in E$,

and

H8. *for any $s \in S$ there is a positive integer k and there are k events $e_1, \ldots, e_k$ such that*

$$\mu(f_{e_1 \ldots e_k}) < 1 \quad \textit{and} \quad \varphi_{e_1 \ldots e_k}(s) > 0,$$

where

$$\varphi_{e_1 \ldots e_n}(s) = P_s(E_j = e_j, 1 \leqslant j \leqslant n). \tag{1.4}$$

In H8 it is understood that the integers and events associated with different states may be different.

The inequality

$$d(g(s), g(s')) \leqslant \mu(g)\, d(s, s') \tag{1.5}$$

for mappings g of S into S suggests that such a function be called *distance diminishing* if $\mu(g) \leqslant 1$ and *strictly distance diminishing* if $\mu(g) < 1$. Hypothesis H7 then says that all event operators are distance diminishing, while H8 says that, whatever the present state, some finite sequence of events with strictly distance diminishing cumulative effect can occur on subsequent trials. Both H7 and H8 (with $k = 1$ for all states), are satisfied, for example, if all event operators are strictly distance diminishing.

It is now possible to introduce the following precise and convenient terminology.

DEFINITION 1.1. *A system* $((S, d), E, f, \varphi)$ *of sets and functions is a* distance diminishing model (*or simply a* model) *if* $f_.(\cdot)$ maps $E \times S$ *into* S, $\varphi_.(\cdot)$ *maps* $E \times S$ *into the nonnegative real numbers*, $\sum_{e \in E} \varphi_e(s) \equiv 1$, *and* H2, H4, H5, H6, H7, *and* H8 *are satisfied.*

DEFINITION 1.2. *Stochastic processes* $\{S_n\}$ *and* $\{E_n\}$ *in the spaces* S *and* E, *respectively, are* associated *with the model if they satisfy* H1 *and* H3.

2. SURVEY OF RESULTS

Most remarks on earlier work by other authors will be deferred until Sec. 4.

A. THEOREMS CONCERNING STATES

The process $\{S_n\}$ associated with any distance diminishing model is a Markov process with stationary transition probabilities given by

$$K(s, A) = \sum_{e : f_e(s) \in A} \varphi_e(s) = P_s(S_2 \in A) \tag{2.1}$$

for Borel subsets A of S. The n step transition probabilities for the process are given by

$$K^{(n)}(s, A) = \sum_{\substack{e_1 \ldots e_n : \\ f_{e_1 \ldots e_n}(s) \in A}} \varphi_{e_1 \ldots e_n}(s) = P_s(S_{n+1} \in A), \tag{2.2}$$

for $n \geqslant 1$. It is convenient to let $K^{(0)}(s, A)$ be 1 if $s \in A$ and 0 otherwise. Functions like K and $K^{(n)}$, probability measures in their second variable for each value of their first, and measurable in their first variable for each value of their second, will be called *stochastic kernels.*

A basic problem is the asymptotic behavior of $K^{(n)}(s, \cdot)$ as $n \to \infty$. Before considering this question, it is necessary to specify what is meant by "convergence" of a sequence $\{\mu_n\}$ of probability measures on S to a probability measure μ on S. The appropriate notion is this: μ_n converges to μ if for any Borel subset A of S

$$\mu(\mathring{A}) \leqslant \liminf_{n \to \infty} \mu_n(A) \quad \text{and} \quad \limsup_{n \to \infty} \mu_n(A) \leqslant \mu(\bar{A})$$

where $\mathring{A}$ is the interior and $\bar{A}$ is the closure of A. If, for instance, S is a real interval, such convergence is equivalent to convergence of distribution functions at all points of continuity of the limit—the usual notion of convergence for distribution functions. The extension of this notion to stochastic kernels that will be used below is as follows.

DEFINITION 2.1. *A sequence* $\{K_n\}$ *of stochastic kernels* converges uniformly *to a stochastic kernel* K_∞ if, *for any Borel subset* A *of* S *and any* $\epsilon > 0$, *there is an integer* N *such that*

$$K_\infty(s, \mathring{A}) - \epsilon \leqslant K_n(s, A) \leqslant K_\infty(s, \bar{A}) + \epsilon$$

for all $n \geqslant N$ *and* s *in* S.

If a limiting stochastic kernel $K_\infty(s, A)$ is independent of s for all A, it is sometimes natural to write $K_\infty(A)$ instead of $K_\infty(s, A)$. Aside from this change of notation Def. 2.1 is unaffected.

A closely related problem is the asymptotic behavior of functions $E_.[\psi(S_n)]$ (moments, for instance) where ψ is a real-valued function on S. Two notions of convergence for sequences of real-valued functions on S are important in what follows. For any such function γ, define $|\gamma|$ and $\|\gamma\|$ by

$$|\gamma| = \sup_{s \in S} |\gamma(s)| \tag{2.3}$$

and

$$\|\gamma\| = |\gamma| + m(\gamma). \tag{2.4}$$

The class of continuous real-valued functions on S is denoted $C(S)$ (note that $|\gamma| < \infty$ if $\gamma \in C(S)$), and the subclass on which $m(\gamma) < \infty$ (and thus $\|\gamma\| < \infty$) is denoted CL. A sequence $\{\gamma_n\}$ of functions in $C(S)$ converges uniformly to $\gamma \in C(S)$ if $|\gamma_n - \gamma| \to 0$ as $n \to \infty$. A stronger notion of convergence, applicable to functions in CL, is $\|\gamma_n - \gamma\| \to 0$ as $n \to \infty$. If S is a real interval then the collection D of functions with a bounded derivative is a closed subset of CL in the sense that, if $\gamma_n \in D$, $\gamma \in CL$, and $\lim_{n\to\infty} \|\gamma_n - \gamma\| = 0$, then $\gamma \in D$. Since $\|\psi\| = |\psi| + |\psi'|$ for any $\psi \in D$, it follows that $|\gamma_n - \gamma| \to 0$ and $|\gamma_n' - \gamma'| \to 0$ as $n \to \infty$. If f_e and $\varphi_e \in D$ for all $e \in E$ and if $\psi \in D$, then $E_.[\psi(S_n)] \in D$ for all $n \geqslant 1$. Thus these observations are applicable to $\gamma_n(\cdot) = E_.[\psi(S_n)]$ and $\gamma_n(\cdot) = (1/n) \sum_{j=1}^n E_.[\psi(S_j)]$.

Theorem 2.1 gives some information about the asymptotic behavior of $\{S_n\}$ for distance diminishing models with no further assumptions.

THEOREM 2.1. *For any distance diminishing model, the stochastic kernel* $(1/n) \sum_{j=0}^{n-1} K^{(j)}$ *converges uniformly as* $n \to \infty$ *to a stochastic kernel* K^∞. *For any Borel subset* A *of* S, $K^\infty(\cdot, A) \in CL$. *There is a constant* $C < \infty$ *such that*

$$\left\| \frac{1}{n} \sum_{j=1}^n E_.[\psi(S_j)] - E_.[\psi(S_\infty)] \right\| \leqslant C \frac{\|\psi\|}{n} \tag{2.5}$$

for all $n \geqslant 1$ *and* $\psi \in CL$, *where*

$$E_s[\psi(S_\infty)] = \int_S \psi(s')\, K^\infty(s, ds'). \tag{2.6}$$

The notation $E_s[\psi(S_\infty)]$ for the expectation of ψ with respect to the asymptotic distribution $K^\infty(s, \cdot)$ is not meant to suggest that there is a random variable S_∞ to which S_n converges with probability 1. Though such convergence occurs, for example, under the hypotheses of Theorem 2.3, it does not occur in general.

Two situations will now be discussed in which the conclusions of Theorem 2.1 can be substantially strengthened. The first is characterized by the loss, asymptotically, of all information about the initial state; the second, by the convergence of S_n to absorbing states with probability 1. Both occur frequently in mathematical learning theory. To describe hypotheses that lead to these situations, it is convenient to have a notation for the set of values that S_{n+1} takes on with positive probability when $S_1 = s$. This set is denoted $T_n(s)$:

$$T_n(s) = \{s' : K^{(n)}(s, \{s'\}) > 0\}. \tag{2.7}$$

An absorbing state is, of course, one that, once entered, cannot be left; that is, $K(s, \{s\}) = 1$ or $T_1(s) = \{s\}$. Another convenient notation is $d(A, B)$ for the (minimum) distance between two subsets A and B of S:

$$d(A, B) = \inf_{s \in A, s' \in B} d(s, s'). \tag{2.8}$$

If B is the unit set $\{b\}$, then $d(A, B)$ is written $d(A, b)$.

Theorem 2.2 shows that, to obtain asymptotic independence of the initial state, it suffices to assume that

H9. $$\lim_{n\to\infty} d(T_n(s), T_n(s')) = 0 \textit{ for all } s, s' \in S.$$

Theorem 2.3 shows that, to obtain convergence to absorbing states, it suffices to assume that:

H10. *There are a finite number of absorbing states* $a_1, \ldots, a_N$, *such that, for any* $s \in S$, *there is some* $a_{j(s)}$ *for which*

$$\lim_{n\to\infty} d(T_n(s), a_{j(s)}) = 0.$$

It is easy to see that H9 and H10 are inconsistent except when there is exactly one absorbing state, in which case they are equivalent.

THEOREM 2.2. *If a distance diminishing model satisfies* H9, *then the asymptotic distribution* $K^\infty(s, \cdot) = K^\infty(\cdot)$ *does not depend on the initial state* s, *and* $K^{(n)}$ *converges uniformly to* K^∞. *There are constants* $\alpha < 1$ *and* $C < \infty$ *such that*

$$\| E_{.}[\psi(S_n)] - E[\psi(S_\infty)] \| \leq C\alpha^n \| \psi \| \tag{2.9}$$

for $n \geq 1$ *and* $\psi \in CL$, *where* $E[\psi(S_\infty)] = \int_S \psi(s)\, K^\infty(ds)$.

THEOREM 2.3. *If a distance diminishing model satisfies* H10, *then the stochastic process* $\{S_n\}$ *converges with probability* 1 *to a random absorbing state* S_∞. *For any* $1 \leqslant i \leqslant N$, *the function* $\gamma_i(s) = P_s(S_\infty = a_i)$ *belongs to CL. If* $b_1, ..., b_N$ *are real numbers, the function* $\gamma(s) = \sum_{i=1}^N b_i\gamma_i(s)$ *is the only continuous solution of the equation* $E_s[\gamma(S_2)] = \gamma(s)$ *that has the boundary values* $\gamma(a_j) = b_j$. *The stochastic kernels* $K^{(n)}$ *converge uniformly to* K^∞, *and* $K^\infty(s, \cdot)$ *assigns weight* $\gamma_i(s)$ *to* a_i, *so that* $E_s[\psi(S_\infty)] = \sum_{i=1}^N \gamma_i(s)\,\psi(a_i)$. *There are* $\alpha < 1$ *and* $C < \infty$ *such that*

$$\| E_.[\psi(S_n)] - E_.[\psi(S_\infty)]\| \leqslant C\alpha^n \| \psi \| \tag{2.10}$$

for all $n \geqslant 1$ *and* $\psi \in CL$.

These theorems suggest the following terminology:

DEFINITION 2.2. *A distance diminishing model is* ergodic *if it satisfies* H9, *and* absorbing *if it satisfies* H10.

Note that, whereas in Theorem 2.1 only the convergence of Cesaro averages is asserted, in Theorems 2.2 and 2.3 the sequences $\{K^{(n)}\}$ and $\{E_.[\psi(S_n)]\}$ themselves converge. It is also worth pointing out that, although it is of little importance that (2.9) and (2.10) imply $\| E_.[\psi(S_n)] - E_.[\psi(S_\infty)]\| \to 0$ instead of simply $| E_.[\psi(S_n)] - E_.[\psi(S_\infty)]| \to 0$, it is of considerable importance that these formulas give a geometric rate of convergence, independent of ψ as long as $\| \psi \|$ is less than some fixed constant.

Proofs of Theorems 2.1–2.3 are given in Sec. 5. The main tool used is the uniform ergodic theorem of Ionescu Tulcea and Marinescu (1950). The results given above do not exhaust the implications of this theorem, even for distance diminishing models, as will be seen in Sec. 5.

B. THEOREMS CONCERNING EVENTS

Consider some characteristic C^ℓ that pertains to ℓ consecutive events, $\ell \geqslant 1$; e.g., "response R occurs on trial n" ($\ell = 1$), "the responses on trials n and $n + 1$ differ" ($\ell = 2$), or "outcome O occurs on trial n and response R on trial $n + 1$" ($\ell = 2$). It is often of interest to know the asymptotic behavior of the probability that $(E_n, ..., E_{n+\ell-1})$ has the property C^ℓ. Let E^ℓ be the set of ℓ-tuples of events, and let A^ℓ be the subset of E^ℓ that corresponds to C^ℓ; that is, $A^\ell = \{(e_1, ..., e_\ell): (e_1, ..., e_\ell)$ has the property $C^\ell\}$. Then it is the asymptotic behavior of

$$P_s^{(n)}(A^\ell) = P_s((E_n, ..., E_{n+\ell-1}) \in A^\ell) \tag{2.11}$$

that is in question. Theorem 2.4, which applies to both ergodic and absorbing models, gives much information.

THEOREM 2.4. *For any ergodic or absorbing model there is an* $L < \infty$ *such that, for any* $\ell \geqslant 1$ *and* $A^\ell \subset E^\ell$,

$$\| P_{\cdot}^{(n)}(A^\ell) - P_{\cdot}^{\infty}(A^\ell) \| \leqslant L\alpha^n \tag{2.12}$$

for all $n \geqslant 1$, *where*

$$P_s^{\infty}(A^\ell) = \int_S P_{s'}^{(1)}(A^\ell)\, K^{\infty}(s, ds'), \tag{2.13}$$

and α *is as in* (2.9) *and* (2.10).

In the ergodic case the subscript s on $P_s^{\infty}(A^\ell)$ can, of course, be dropped.

The following corollary for absorbing models is very useful.

COROLLARY 2.5. *If an absorbing model and an* $A^\ell \subset E^\ell$ *have the property that* $P_{a_i}^{(1)}(A^\ell) = 0$ *for* $i = 1,..., N$, *then the total number* X *of positive integers* n *for which* $(E_n ,..., E_{n+\ell-1}) \in A^\ell$ *is finite with probability* 1, *and*

$$\| E_{\cdot}[X] \| \leqslant L\alpha/(1 - \alpha). \tag{2.14}$$

The function $\chi(s) = E_s[X]$ *is the unique continuous solution of the equation*

$$\chi(s) = P_s^{(1)}(A^\ell) + E_s[\chi(S_2)],$$

for which $\chi(a_i) = 0$, $i = 1,..., N$.

The next theorem concerns ergodic models, and requires some additional notation for its statement. Let h be a real-valued function on E. Then the asymptotic expectations of $h(E_n)$ and $h(E_n)\, h(E_{n+j})$ are denoted $E[h(E_\infty)]$ and $E[h(E_\infty)\, h(E_{\infty+j})]$, respectively. Thus

$$E[h(E_\infty)] = \sum_{e \in E} h(e)\, P^{\infty}(\{e\}), \tag{2.15}$$

and

$$E[h(E_\infty)\, h(E_{\infty+j})] = \sum_{e_1 \cdots e_{j+1}} h(e_1)\, h(e_{j+1})\, P^{\infty}(\{(e_1 ,..., e_{j+1})\}). \tag{2.16}$$

In typical applications h will be the indicator function of some $A \subset E$, so that $\sum_{j=m}^{m+n-1} h(E_j)$ is the number of occurrences of events in A during the block of n trials beginning on trial m. In this case,

$$E[h(E_\infty)] = P^{\infty}(A),$$

$$E[h(E_\infty)\, h(E_{\infty+1})] = P^{\infty}(A \times A),$$

and

$$E[h(E_\infty)\, h(E_{\infty+j})] = P^{\infty}(A \times E^{j-1} \times A) \quad \text{for} \quad j \geqslant 2.$$

THEOREM 2.6. (i) *For any ergodic model, and any real valued function h on E, the series*

$$E[h^2(E_\infty)] - E^2[h(E_\infty)] + 2 \sum_{j=1}^{\infty} (E[h(E_\infty)\, h(E_{\infty+j})] - E^2[h(E_\infty)]) \tag{2.17}$$

converges to a nonnegative constant σ_h^2.

(ii) *For some* $C_h < \infty$ *and all* $m, n \geqslant 1$

$$\left| \frac{1}{n} E_{\cdot} \left[\left(\sum_{j=m}^{m+n-1} h(E_j) - nE[h(E_\infty)] \right)^2 \right] - \sigma_h^2 \right| \leqslant C_h n^{-1/2}. \tag{2.18}$$

Consequently, the law of large numbers

$$\lim_{n\to\infty} E_s \left[\left(\frac{1}{n} \sum_{j=m}^{m+n-1} h(E_j) - E[h(E_\infty)] \right)^2 \right] = 0, \tag{2.19}$$

holds uniformly in s.

(iii) *If* $\sigma_h^2 > 0$, *the* central limit theorem

$$\lim_{n\to\infty} P_s \left(\frac{\sum_{j=m}^{m+n-1} h(E_j) - nE[h(E_\infty)]}{(n)^{1/2}\sigma_h} < x \right) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{x} \exp(-t^2/2)\, dt \tag{2.20}$$

is valid for all $s \in S$.

A distance diminishing model can be regarded as an example of what Iosifescu (1963) calls a *homogeneous random system with complete connections*. Theorem 2.6 is a consequence of Theorem 2.4 and a theorem of Iosifescu on such systems. Results in this subsection will be proved in Sec. 6.

3. EXAMPLES

The examples to be discussed have been selected so as to illustrate a variety of ramifications of the theory developed in Secs. 1 and 2.

A. LINEAR MODELS WITH EXPERIMENTER-SUBJECT CONTROLLED EVENTS

Suppose that the subject in a learning experiment has response alternatives A_1 and A_2 on each trial, and that, following response A_i, one of two observable outcomes O_{i1} and O_{i2} occurs. It is assumed that O_{1j} and O_{2j} positively reinforce A_j, in the weak sense that they do not decrease the probability of A_j. The outcome probabilities are supposed to depend at most on the most recent response. Let $A_{i,n}$ and $O_{ij,n}$

denote, respectively, the occurrence of A_i and O_{ij} on trial n, and denote the probability $P(O_{ij,n} \mid A_{i,n})$ by π_{ij}.

Linear models with experimenter-subject controlled events (Bush and Mosteller, 1955) for this situation can be described within the framework of Sec. 1 by identifying p_n, the (conditional) probability of $A_{1,n}$, with the state S_n, by identifying the response-outcome pair that occurs on trial n with the event E_n, and by making the following stipulations:

$$S = [0, 1], \quad d(p, p') = | p - p' |, \tag{3.1}$$

$$(i, j) = (A_i, O_{ij}) \quad \text{and} \quad E = \{(i, j): 1 \leqslant i, j \leqslant 2\}, \tag{3.2}$$

$$f_{ij}(p) = (1 - \Theta_{ij})\, p + \Theta_{ij}\, \delta_{j1}, \tag{3.3}$$

$$\varphi_{ij}(p) = (p\, \delta_{i1} + (1 - p)\, \delta_{i2})\, \pi_{ij}, \tag{3.4}$$

$$\pi_{i1} + \pi_{i2} = 1, \quad \text{and} \quad 0 \leqslant \Theta_{ij}, \pi_{ij} \leqslant 1 \quad \text{for} \quad 1 \leqslant i, j \leqslant 2. \tag{3.5}$$

In (3.3) and (3.4), δ_{ij} is the Kronecker δ. For convenience, any system $((S, d), E, f, \varphi)$ of sets and functions satisfying (3.1–3.5) will be referred to as a *four-operator model.* In this terminology, (3.1–3.5) define a six-parameter family of four-operator models, one for each choice of Θ_{11}, Θ_{12}, Θ_{21}, Θ_{22}, π_{11}, and π_{22} consistent with (3.5). Since $m(\varphi_{ij}) = \pi_{ij}$ and $\mu(f_{ij}) = (1 - \Theta_{ij}) \leqslant 1$, it is clear that any four-operator model satisfies all of the conditions of Def. 1.1 except perhaps H8.

The asymptotic behavior of the process $\{p_n\}$ associated with a four-operator model depends critically on the number of absorbing states. Lemma 3.1 catalogues the absorbing states for a four-operator model.

Lemma 3.1. *The state 1 is absorbing if and only if $\pi_{12} = 0$ or $\Theta_{12} = 0$. The state 0 is absorbing if and only if $\pi_{21} = 0$ or $\Theta_{21} = 0$. A state $p \in (0, 1)$ is absorbing if and only if for each $(i, j) \in E$, $\Theta_{ij} = 0$ or $\pi_{ij} = 0$. In this case all states are absorbing, and the model is said to be trivial.*

Proof. A state $p \in (0, 1)$ is absorbing if and only if, for any $(i, j) \in E$, either $f_{ij}(p) = p$ (in which case $\Theta_{ij} = 0$ and $f_{ij}(x) \equiv x$) or $\varphi_{ij}(p) = 0$ (in which case $\pi_{ij} = 0$ and $\varphi_{ij}(x) \equiv 0$).

The state 1 is absorbing if and only if $1 - \Theta_{12} = f_{12}(1) = 1$ or $\pi_{12} = \varphi_{12}(1) = 0$. The assertion concerning the state 0 is proved similarly. Q.E.D.

The next lemma tells which four-operator models satisfy H8.

Lemma 3.2. *A four-operator model is distance diminishing if and only if, for each $i \in \{1, 2\}$, there is some $j_i \in \{1, 2\}$ such that $\Theta_{ij_i} > 0$ and $\pi_{ij_i} > 0$.*

Proof. Suppose that the condition given by the lemma is met. If $p > 0$ then $\varphi_{1j_1}(p) = p\pi_{1j_1} > 0$ and $\mu(f_{1j_1}) = 1 - \Theta_{1j_1} < 1$. Similarly, if $p < 1$ then $\varphi_{2j_2}(p) > 0$ and $\mu(f_{2j_2}) < 1$. Thus H8 is satisfied with $k = 1$ for all states.

Suppose that the condition fails. Then for some $i \in \{1, 2\}$ and all $j \in \{1, 2\}$, $\Theta_{ij} = 0$, or $\pi_{ij} = 0$. Since the cases $i = 1$ and $i = 2$ can be treated similarly, only $i = 1$ will be considered. It follows from Lemma 3.1 on taking $j = 2$ that 1 is an absorbing state. Thus $\varphi_{m_1 n_1, \ldots, m_k n_k}(1) > 0$ implies $m_\ell = 1$ and $\pi_{1n_\ell} > 0$, $1 \leqslant \ell \leqslant k$. But then $\Theta_{1n_\ell} = 0$ for $1 \leqslant \ell \leqslant k$ and $\mu(f_{m_1 n_1, \ldots, m_k n_k}) = 1$. So H8 is not satisfied. Q.E.D.

Clearly a distance diminishing four-operator model is nontrivial.

With one inconsequential exception, distance diminishing four-operator models are either ergodic or absorbing. Theorems 3.1 and 3.2 show slightly more.

THEOREM 3.1. *If neither* 0 *nor* 1 *is absorbing for a four-operator model, then* $\Theta_{ij} > 0$ *and* $\pi_{ij} > 0$ *for* $i \neq j$, *and the model is distance diminishing. Either* (i) $\Theta_{ij} = 1$ *and* $\pi_{ij} = 1$ *if* $i \neq j$; *or* (ii) *the model is ergodic.*

THEOREM 3.2. *If a distance diminishing four-operator model has an absorbing state, then it is an absorbing model.*

The behavior of the process $\{p_n\}$ when $\Theta_{ij} = 1$ and $\pi_{ij} = 1$ for $i \neq j$ is completely transparent. Starting at p the process moves on its first step to 1 with probability $1 - p$ and to 0 with probability p, and thereafter alternates between these two extreme states. This *cyclic* model is of no psychological interest and will be discussed no further.

Proof of Theorem 3.1. By Lemma 3.1 if neither 0 nor 1 is absorbing then $\Theta_{ij} > 0$ and $\pi_{ij} > 0$ for $i \neq j$, and the model is distance diminishing by Lemma 3.2.

Suppose $\pi_{21} < 1$. Then by considering first the case $p = 0$, then $p > 0$ and $\Theta_{12} = 1$, and finally $p > 0$ and $\Theta_{12} < 1$, it is seen that $(1 - \Theta_{12})^n p \in T_n(p)$ for all $n \geqslant 1$. Thus $d(T_n(p), T_n(q)) \leqslant (1 - \Theta_{12})^{n-1} \mid p - q \mid \to 0$ as $n \to \infty$, and the model is ergodic according to Def. 2.1. By symmetry the same conclusion obtains if $\pi_{12} < 1$. Suppose that $\Theta_{12} < 1$. Then $(1 - \Theta_{12})^n p \in T_n(p)$ for all $p > 0$ and $n \geqslant 1$, and $(1 - \Theta_{12})^{n-1} \Theta_{21} \in T_n(0)$ for all $n \geqslant 1$. Since both sequences tend to 0, ergodicity follows. The same conclusion follows by symmetry when $\Theta_{21} < 1$. Thus if (i) does not hold the model is ergodic. Q.E.D.

Proof of Theorem 3.2. The condition given by Lemma 3.2 for a four-operator model to be distance diminishing allows four possibilities. These are distinguished by the values of j_i, $i = 1, 2$: A: $j_1 = 1, j_2 = 1$; B: $j_1 = 2, j_2 = 2$; C: $j_1 = 1$, $j_2 = 2$; and D: $j_1 = 2, j_2 = 1$. Lemma 3.1 shows that D is inconsistent with the existence of absorbing states. Thus it remains to show that a model is absorbing under A, B, or C if there are absorbing states.

Under A, $1 - (1 - \Theta_{21})^n (1 - p) \in T_n(p)$ for all $n \geqslant 1$ and $0 \leqslant p \leqslant 1$, so $d(T_n(p), 1) \leqslant (1 - \Theta_{21})^n \to 0$ as $n \to \infty$. This implies that 0 is not an absorbing state. By assumption, however, there is at least one absorbing state, so 1 is absorbing. But then $\lim_{n\to\infty} d(T_n(p), 1) = 0$ for all $0 \leqslant p \leqslant 1$ implies that the model is absorbing. By symmetry the model is also absorbing under B.

If 0 is not absorbing $\pi_{21} > 0$ and $\Theta_{21} > 0$ by Lemma 3.1. Thus, if C holds, A does also, and the model is absorbing. If C holds, and 1 is not absorbing, the same conclusion follows by symmetry. Condition C implies that $(1 - \Theta_{22})^n p \in T_n(p)$ for $p < 1$, $1 - (1 - \Theta_{11})^n (1 - p) \in T_n(p)$ for $p > 0$, and $\Theta_{11}, \Theta_{22} > 0$. Thus if both 0 and 1 are absorbing, H10 is satisfied with $j(1) = 1, j(0) = 0$ and $j(p) = 1$ or 0 for $0 < p < 1$. Q.E.D.

As a consequence of Theorems 3.1 and 3.2 all of the theorems of Sec. 2 for ergodic models are valid for noncyclic four-operator models without absorbing states, and all theorems of Sec. 2 for absorbing models are valid for distance diminishing four-operator models with absorbing states. A few illustrative specializations of the theorems of Sec. 2 to the case at hand will now be given. The first concerns convergence of the moments $E_p[p_n^{\nu}]$ of the process $\{p_n\}$.

THEOREM 3.3. *For any noncyclic distance diminishing four-operator model there are constants $C < \infty$ and $\alpha < 1$ such that*

$$\| E_{.}[p_n^{\nu}] - E_{.}[p_{\infty}^{\nu}]\| \leqslant C(\nu + 1)\alpha^n \tag{3.6}$$

for all real $\nu \geqslant 1$ and positive integers n. The function $E_{.}[p_{\infty}^{\nu}]$ has a bounded derivative.

This is obtained from (2.9) and (2.10) by noting that the function $\psi(p) = p^{\nu}$ belongs to D with $| \psi | = 1$ and $m(\psi) = | \psi' | = \nu$, so that $\| \psi \| = \nu + 1$.

If 0 is the only absorbing state of a distance diminishing four-operator model, Theorem 2.3 implies that $\lim_{n\to\infty} p_n = 0$ with probability 1, whatever the value of p_1. It is conceivable, however, that the convergence is sufficiently slow that the total number X of A_1 responses is infinite with positive probability. Furthermore, even if X is finite with probability 1, it might have an infinite mean. Similarly, even though p_n converges to 0 or 1 in the case of two absorbing states, *a priori* considerations do not rule out the possibility that the total number Y of alternations between responses is infinite, or barring that, that its mean is infinite. Theorem 3.4 excludes these possibilities.

THEOREM 3.4. *If 0 is the only absorbing state of a distance diminishing four-operator model, then X, the total number of A_1 responses, is finite with probability 1, and $\| E_{.}[X]\| < \infty$. If both 0 and 1 are absorbing states of a distance diminishing four-operator model, then Y, the total number of alternations between responses, is finite with probability 1 and $\| E_{.}[Y]\| < \infty$.*

Naturally the first assertion is still true if 1 replaces 0 as the only absorbing state and X is the total number of A_2 responses.

Proof. Let B and D be the subsets of E and E^2, respectively, defined by

$$B = \{(1, 1), (1, 2)\},$$

and

$$D = \{((i, j), (k, \ell)) : i \neq k\}. \tag{3.7}$$

Then $E_n \in B$ if and only if A_1 occurs on trial n, and $(E_n, E_{n+1}) \in D$ if and only if there is a response alternation between trials n and $n + 1$. Since $P_0^{(1)}(B) = 0$, and, if both 0 and 1 are absorbing, $P_0^{(1)}(D) = P_1^{(1)}(D) = 0$, the conclusions of the theorem follow directly from Corollary 2.5. Q.E.D.

If h is the indicator function of the subset B of E given by (3.7), then $A_n = h(E_n)$ is the indicator random variable of $A_{1,n}$ and $\sum_{j=m}^{m+n-1} A_j$ is the frequency of A_1 in the block of n trials beginning on trial m. Theorems 3.1 and 2.6 yield a law of large numbers and, perhaps, a central limit theorem for this quantity for any noncyclic four-operator model with no absorbing states. The full power of this result comes into play when the quantities $E[A_\infty] = \lim_{n\to\infty} P_p(A_{1,n})$ and $\sigma_h{}^2$ can be computed explicitly in terms of the parameters of the model. This is the case, for instance, when all Θ_{ij} are equal.

THEOREM 3.5. *A four-operator model with $\Theta_{ij} = \Theta > 0$ for $1 \leqslant i, j \leqslant 2$, and $\pi_{ij} > 0$ for $i \neq j$, but not $\Theta = \pi_{12} = \pi_{21} = 1$, is ergodic. The law of large numbers*

$$\lim_{n\to\infty} E_p[((1/n)\, S_{m,n} - \ell)^2] = 0, \tag{3.8}$$

and central limit theorem

$$\lim_{n\to\infty} P_p\left(\frac{S_{m,n} - n\ell}{(n)^{1/2}\sigma} < x\right) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{x} \exp(-t^2/2)\, dt, \tag{3.9}$$

hold, where

$$\ell = \pi_{21}/(\pi_{21} + \pi_{12}), \tag{3.10}$$

$$\sigma^2 = \frac{\ell(1-\ell)}{\pi_{21} + \pi_{12}}\left[\pi_{11} + \pi_{22} + \frac{2(1-\Theta)}{(2-\Theta) + 2(1 - \pi_{11} - \pi_{22})(1-\Theta)}\right], \tag{3.11}$$

and $S_{m,n}$ is the total number of A_1 responses in the n trial block beginning on trial m.

Outline of proof. Ergodicity follows from Theorem 3.1, so Theorem 2.6 is applicable. Straightforward computation yields

$$E[p_\infty] = E[A_\infty] = \ell, \tag{3.12}$$

$$E[p_\infty^2] = \ell^2 + \Theta\ell(1-\ell)/[(2-\Theta) + 2(1-\pi_{11}-\pi_{22})(1-\Theta)], \tag{3.13}$$

$$E[A_\infty A_{\infty+1}] = (1-\Theta)\,E[p_\infty^2] + \pi_{11}\Theta E[p_\infty], \tag{3.14}$$

and

$$E[A_\infty A_{\infty+k}] - E^2[A_\infty] = (1-\Theta(\pi_{12}+\pi_{21}))^{k-1}(E[A_{\infty+1}A_\infty] - E^2[A_\infty]) \tag{3.15}$$

for $k \geqslant 1$. These formulas permit computation of $\sigma^2 = \sigma_h^2$, the series in (2.17). The result, recorded in (3.11), is positive, since $0 < \ell < 1$, and either $\pi_{11} + \pi_{22} > 0$ or $(1-\Theta) > 0$. Q.E.D.

The equality

$$M(p) - p = (\pi_{12} + \pi_{21})(\ell - p),$$

where

$$M(p) = P(O_{11,n} \quad \text{or} \quad O_{21,n} \mid p_n = p),$$

shows that the asymptote ℓ of A_1 response probability for the linear model with equal Θ's is associated with the asymptotic equality of the probability of A_1 and the probability of reinforcement of A_1:

$$\lim_{n\to\infty} P(O_{11,n} \text{ or } O_{21,n}) = \lim_{n\to\infty} P(A_{1,n}).$$

Such *probability matching* is a well known prediction of the linear model with equal Θ's. Theorem 3.5 contains a much stronger prediction. The law of large numbers (3.8) asserts that the proportion $(1/n)\,S_{m,n}$ of A_1 responses *for a single subject* in a long block of trials is close to ℓ with high probability. The terms "close" and "high" are further quantified by the central limit theorem (3.9). To illustrate, if reinforcement is noncontingent with $\pi_{11} = \pi_{21} = .75$ and Θ is small (that is, learning is slow), then $\sigma^2 \doteq 2\pi(1-\pi) = .375$ so that, in a block of 400 trials commencing on trial 100, the probability is approximately .01 that $(1/400)\,S_{100,400}$ will depart from .75 by as much as $(2.58)(.612)/20 = .079$.

There is one modification of the four-operator model, examples of which have occurred sufficiently frequently in the literature (see Estes and Suppes, 1959, Norman, 1964, and Yellott, 1965) to warrant comment here. If, following any of the outcomes O_{ij}, conditioning is assumed to be effective (i.e., $p_{n+1} = f_{ij}(p_n)$) with probability c_{ij} and, otherwise, ineffective (i.e., $p_{n+1} = p_n$), a *five-operator model* is obtained. It is easy to amend (3.2–3.5) to obtain a formal description within the framework of Sec. 1. Such an addition of an identity operator does not affect the validity of any of the results preceding Theorem 3.5 (or their proofs) provided that π_{ij} is everywhere replaced by $\pi_{ij}c_{ij}$. The first sentence of the amended Theorem 3.5 should read: *A five-operator model with* $\Theta_{ij} = \Theta > 0$ *and* $c_{ij} = c > 0$ *for* $1 \leqslant i, j \leqslant 2$, *and* $\pi_{ij} > 0$ *for* $i \neq j$, *but not* $\Theta = c = \pi_{12} = \pi_{21} = 1$, *is ergodic*. Also (3.11) should be replaced by

$$\sigma^2 = \frac{\ell(1-\ell)}{\pi_{21}+\pi_{12}}\left[\pi_{11}+\pi_{22}+\frac{2(1-c\Theta)}{c((2-\Theta)+2(1-\pi_{11}-\pi_{22})(1-\Theta))}\right], \quad (3.16)$$

and Θ should be replaced by $c\Theta$ in (3.14) and (3.15). An interesting implication of (3.16) is that $\lim_{\Theta\to 0}\sigma^2 < \infty$, whereas, if $c\Theta < 1$, $\lim_{c\to 0}\sigma^2 = \infty$. Thus the variance of the total number of A_1 responses in a long block of trials may be useful in deciding whether a given instance of "slow learning" is due to small Θ or small c.

B. Lovejoy's Model I

Lovejoy's (1966) Model I is a simple model for simultaneous discrimination learning. Let the relevant stimulus dimension be brightness, and let white (W) be positive and black (B) be negative. On each trial the subject is supposed either to attend to brightness (A) or not ($\bar{A}$), which events have probabilities $P_n(A)$ and $1 - P_n(A)$. Given A the probability of the response appropriate to white is $P_n(W \mid A)$, while given $\bar{A}$ the probability of this response is $1/2$. The subject's state on trial n is then described by the vector $(P_n(A), P_n(W \mid A))$, and the state space is

$$S = \{(p, p'): 0 \leqslant p, p' \leqslant 1\}.$$

This is a compact metric space with respect to the ordinary Euclidean metric d. The events are the elements of

$$E = \{(A, W), (A, B), (\bar{A}, W), (\bar{A}, B)\},$$

the corresponding transformations are

$$f_{AW}(p, p') = (\alpha_1 p + (1-\alpha_1), \alpha_3 p' + (1-\alpha_3)),$$

$$f_{AB}(p, p') = (\alpha_2 p, \alpha_4 p' + (1-\alpha_4)),$$

$$f_{\bar{A}W}(p, p') = (\alpha_1 p, p'),$$

and

$$f_{\bar{A}B}(p, p') = (\alpha_2 p + (1-\alpha_2), p'),$$

where $0 < \alpha_1, \alpha_2, \alpha_3, \alpha_4 < 1$, and their probabilities are

$$\varphi_{AW}(p, p') = pp',$$

$$\varphi_{AB}(p, p') = p(1-p'),$$

$$\varphi_{\bar{A}W}(p, p') = (1-p)/2,$$

and

$$\varphi_{\bar{A}B}(p, p') = (1-p)/2.$$

Any system $((S, d), E, f, \varphi)$ of sets and functions satisfying the above stipulations will be called a *discrimination model of type* I below.

THEOREM 3.6. *Any discrimination model of type* 1 *is distance diminishing and absorbing with single absorbing state* (1, 1).

Proof. Axiom H6 is satisfied because of the continuous differentiability of the φ's, and H7 follows from

$$\mu(f_{AW}) = \max(\alpha_1, \alpha_3) < 1,$$

$$\mu(f_{AB}) = \max(\alpha_2, \alpha_4) < 1,$$

and

$$\mu(f_{\bar{A}W}) = \mu(f_{\bar{A}B}) = 1.$$

Thus it remains only to verify H8 and H10.

Note that, as a consequence of (1.5), for any mappings f and g of S into S such that $\mu(f) < \infty$ and $\mu(g) < \infty$, the inequality

$$\mu(f \circ g) \leqslant \mu(f)\,\mu(g)$$

obtains, where $f \circ g(s) = f(g(s))$. This implies that $\mu(f_{e_1 \dots e_k}) < 1$ if $e_k = (A, W)$. Now $\varphi_{AW}(p, p') > 0$ throughout

$$S' = \{(p, p'): p > 0, p' > 0\},$$

so to complete the verification of H8 it suffices to show that if $(p, p') \in' S'$ there is a $k \geqslant 2$ and there are events $e_1, \dots, e_{k-1}$ such that $f_{e_1 \dots e_{k-1}}(p, p') \in S'$ and $\varphi_{e_1 \dots e_{k-1}}(p, p') > 0$. If $0 < p' \leqslant 1$, $\varphi_{\bar{A}B}(0, p') > 0$ and $f_{\bar{A}B}(0, p') \in S'$, while if $0 < p \leqslant 1$, $\varphi_{AB}(p, 0) > 0$ and $f_{AB}(p, 0) \in S'$. Finally $f_{\bar{A}B}(0, 0)$ has positive first and null second coordinate, so $f_{\bar{A}B,AB}(0, 0) \in S'$ and $\varphi_{AB}(f_{\bar{A}B}(0, 0)) > 0$. Since $\varphi_{\bar{A}B}(0, 0) > 0$ the latter inequality implies $\varphi_{\bar{A}B,AB}(0, 0) > 0$.

The above argument shows that for any $(p, p') \in S$ there is a point

$$s^{pp'} \in T_{k-1}(p, p') \cap S'.$$

Since f_{AW} maps S' into S' it follows that $f_{AW}^{(n)}(s^{pp'}) \in T_{n+k-1}(p, p')$ for $n \geqslant 0$, where $f_{AW}^{(n)}$ is the nth iterate of f_{AW}, i.e., $f_{AW}^{(0)}(s) \equiv s$ and $f_{AW}^{(j+1)} = f_{AW} \circ f_{AW}^{(j)}$, $j \geqslant 0$. But for any $(q, q') \in S$ and $n \geqslant 0$,

$$f_{AW}^{(n)}(q, q') = (1 - (1 - q)\,\alpha_1^{\,n}, 1 - (1 - q')\,\alpha_3^{\,n}),$$

so

$$d(T_{n+k-1}(p, p'), (1, 1)) \leqslant d(f_{AW}^{(n)}(s^{pp'}), (1, 1))$$
$$\leqslant (\alpha_1^{2n} + \alpha_3^{2n})^{1/2} \to 0 \quad \text{as} \quad n \to \infty.$$

Since (1, 1) is obviously an absorbing state, the verification of H10 is complete.

Q.E.D.

Here is a sample of what can be concluded about Lovejoy's Model I on the basis of Theorem 3.6, Theorem 2.3, and Corollary 2.5.

THEOREM 3.7. *For any discrimination model of type* I, $\lim_{n\to\infty} P_n(A) = 1$ *and* $\lim_{n\to\infty} P_n(W \mid A) = 1$ *with probability* 1. *There are constants* $C < \infty$ *and* $\alpha < 1$ *such that*

$$\| E_{.}[P_n^{\nu}(A)\, P_n^{\omega}(W \mid A)] - 1 \| \leqslant C((\nu^2 + \omega^2)^{1/2} + 1)\, \alpha^n, \tag{3.17}$$

for all real ν, $\omega \geqslant 1$ *and positive integers* n. *The total number* Z *of* B *responses is finite with probability* 1 *and* $\| E_{.}[Z]\| < \infty$. *If* $\alpha_1 = \alpha_2 = 1 - \Theta$ *and* $\alpha_3 = \alpha_4 = 1 - \Theta'$, *then*

$$E_{p,p'}[Z] = (1 - p)/\Theta + 2(1 - p')/\Theta'. \tag{3.18}$$

Proof. The first statement follows directly from Theorem 2.3. The second follows from (2.10) on taking $\psi(p, p') = p^{\nu} p'^{\omega}$ and noting that $m(\psi) \leqslant (\nu^2 + \omega^2)^{1/2}$ as a consequence of the mean value theorem and the Schwartz inequality. The third statement follows from Corollary 2.5 on taking $A^{(1)} = \{(A, B), (\tilde{A}, B)\}$ so that $P_{p,p'}(A^{(1)}) = p(1 - p') + (1 - p)/2$. Since the function $\chi(p, p') = (1 - p)/\Theta + 2(1 - p')/\Theta'$ is obviously continuous with $\chi(1, 1) = 0$, (3.18) is proved by verifying that this function satisfies the functional equation given in the statement of Corollary 2.5 when $\alpha_1 = \alpha_2 = 1 - \Theta$ and $\alpha_3 = \alpha_4 = 1 - \Theta'$. Q.E.D.

Of the two learning rate parameters appearing in (3.18), Θ' is associated with the response learning process $\{P_n(W \mid A)\}$, while Θ is associated with the perceptual learning process $\{P_n(A)\}$. Suppose that the discrimination problem under consideration (with $P_1(A) = p$ and $P_1(W \mid A) = p'$) has been preceded by j trials of a previous (reversed) problem with black the positive stimulus. Then p will tend to increase and p' to decrease as j increases. Thus overtraining tends to decrease $(1 - p)/\Theta$ and to increase $2(1 - p')/\Theta'$. Which effect predominates and determines the effect of overtraining on $E_{p,p'}[Z]$ will depend on the magnitudes of Θ and Θ', large Θ and small Θ' leading to an increase in errors with overtraining, and small Θ and large Θ' leading to a decrease in errors with overtraining—the "overlearning reversal effect." This oversimplified argument ignores the effect of the magnitudes of Θ and Θ' on p and p', but it none the less suggests the power of (3.18).

In concluding this subsection it is worth remarking that the theory of Sec. 2 is also applicable to Bush's (1965, pp. 172–175) linear operator version of Wyckoff's (1952) discrimination model when $P = 1/2$. In that case, (x_n, y_n, u_n) can be taken to be the state on trial n, and this triple determines the error probability on trial n. When there are only two learning rate parameters, $\Theta' > 0$ for $\{x_n\}$ and $\{y_n\}$, and $\Theta > 0$ for $\{u_n\}$, the expected total errors is given by

$$E_{x,y,u}[Z] = (1 - u)/\Theta + 2(1 - x)/\Theta' + 2y/\Theta'.$$

4. SOME REMARKS ON PREVIOUS WORK

Theorem 2.2 includes the main convergence theorem of Onicescu and Mihoc (1935, Sec. 5), and an ergodic theorem of Ionescu Tulcea (1959, Sec. 8). It includes many of Karlin's (1953) results, and has points of contact with the work of Lamperti and Suppes (1959) and Iosifescu and Theodorescu (1965). None of this previous work covers the general noncyclic four-operator (linear) model without absorbing states. Karlin's results are concerned, for the most part, with two-operator models, that is, four-operator models with $\Theta_{11} = \Theta_{21} = \Theta_1$ and $\Theta_{12} = \Theta_{22} = \Theta_2$. The main ergodic theorem of Iosifescu and Theodorescu (1965, Theorem 2) is not applicable to any four-operator model, since one of its assumptions is that there is some positive integer k, positive real number α, and response A_{i_0}, such that response A_{i_0} has probability at least α on trial $k + 1$, regardless of the initial probability of A_{i_0} and the responses and outcomes on trials 1 through k. Such a hypothesis would be more appropriate if some of the operators in (3.3) had fixed points other than 0 or 1.

The method of Lamperti and Suppes is somewhat different from that of the present paper, and has a certain shortcoming. Consider a two-operator model with Θ_1, $\Theta_2 > 0$ and $0 < \pi_{12}$, $\pi_{21} < 1$. Such a model satisfies the hypotheses of Lamperti and Suppes' Theor. 4.1 (with $m^* = 1$, $k^* = 1$ or 2, and $m_0 = 0$) if their event "$E_n = j$" is identified with "$O_{1j,n}$ or $O_{2j,n}$" in the notation of Subsec. 3a. One of the conclusions of that theorem is that, for all positive integers ν, $\alpha_{1,n}^{\nu} = E_p[p_n^{\nu}]$ converges, as $n \to \infty$, to a quantity $\alpha_1^{\nu} = E[p_{\infty}^{\nu}]$ which does not depend on the initial A_1 response probability p. The α notation is theirs. This conclusion follows, of course, from Theor. 2.2 of the present paper (along with an estimate of the rate of convergence that their method does not yield). But the author has found no arguments in their paper that bear directly on the lack of dependence of the limit on p. (Their notation, $\alpha_{1,n}^{\nu}$, does not even refer to p.) The only kind of conclusion that can be drawn from the arguments given by Lamperti and Suppes is that (in the notation of the present paper), for any p,

$$E_p[p_{n+k}^{\nu} \mid O_{i_k j_k, k} A_{i_k, k} \cdots O_{i_1 j_1, 1} A_{i_1, 1}] = E_{f_{i_1 j_1, \ldots, i_k j_k}(p)}[p_n^{\nu}]$$

converges as $n \to \infty$ to a quantity that does not depend on k or $i_1, j_1, \ldots, i_k, j_k$. This is not quite what is required. The recent corrections (Lamperti and Suppes, 1965) of the Lamperti and Suppes paper do not affect this observation. The method of Lamperti and Suppes is an extension of the method used by Doeblin and Fortet (1937) to study what they call *chaînes* (B). It appears that Doeblin and Fortet's treatment of Onicescu and Mihoc's *chaînes à liaisons complètes* (*chaînes* ($O - M$)) by means of their theory of *chaînes* (A) has the same shortcoming.

A distance diminishing four-operator model with two absorbing states necessarily has π_{ii}, $\Theta_{ii} > 0$ for $i = 1, 2$ and either $\pi_{ij} = 0$ or $\Theta_{ij} = 0$ for $i \neq j$. Thus it has two effective operators and, perhaps, an identity operator. Such models were studied by Karlin (1953), and the implications of Theorem 2.3 for these models do not add much to his results. The generality of Theorem 2.3 is roughly comparable to that of Kennedy's (1957) theorems, though Kennedy's assumptions exclude Lovejoy's Model I.

The ergodic theorem of Ionescu Tulcea and Marinescu (1950) used in the proof of Theorems 2.1–2.3 extends earlier work by Doeblin and Fortet (1937, see section titled *Note sur une équation fonctionnelle*). The condition H9 was used by Jamison (1964).

Let Y be the total number of response alternations for a distance diminishing four-operator model with two absorbing barriers. That Y is finite with probability 1 (see Theorem 3.4) follows, in the special case $\pi_{ii} = 1$, $\Theta_{ii} = \Theta > 0$, $1 \leqslant i \leqslant 2$, from a result of Rose (1964, Corollary 2 of Theor. 5).

Theorems 3 and 4 of Iosifescu and Theodorescu (1965) give results like those of Theorem 2.6 of the present paper for a subclass of the class of models to which their Theorem 2 is applicable. This class of models is disjoint from the class of four-operator models, as was pointed out above. However, once Theorem 2.4 has been proved, a theorem of Iosifescu (1963) leads to Theorem 2.6. To the results in Theorem 3.5 and its five-operator generalization could be added

$$\frac{1}{n} \operatorname{var}_p(S_{m,n}) = \sigma^2 + O(n^{-1/2}),$$

which also follows from Theorem 2.6. In the special case of noncontingent reinforcement and $c = 1$, the result is a consequence of Theorem 8.10 of Estes and Suppes (1959). A similar result for $\lim_{m\to\infty} \operatorname{var}_p (S_{m,n})$ when reinforcement is noncontingent and $0 < c \leqslant 1$ follows from formula (2.16) of Yellott (1965).

5. PROOFS OF THEOREMS CONCERNING STATES

A. The Basic Ergodic Theorem

In this section only, $C(S)$ is the set of *complex valued* continuous functions on S, and $m(\cdot)$, $|\cdot|$, $\|\cdot\|$, and CL are redefined accordingly [see (1.2), (2.3), (2.4) and the sentence following (2.4)]. The spaces $C(S)$ and CL are Banach spaces with respect to the norms $|\cdot|$ and $\|\cdot\|$ respectively. The space CL is also a normed linear space with respect to $|\cdot|$. The norm of a bounded linear operator on $C(S)$ or CL is denoted in the same way as the norm of an element of these spaces. Thus if U is a bounded linear operator on $C(S)$ its norm is

$$| U | = \sup_{\substack{\psi \in C(S): \\ |\psi| \leqslant 1}} | U\psi |,$$

while if U is a bounded linear operator on CL its norm is

$$\| U \| = \sup_{\substack{\psi \in CL: \\ \|\psi\| \leqslant 1}} \| U\psi \|.$$

Finally, if U is an operator on CL, bounded with respect to $| \cdot |$, its norm is denoted $| U |_{CL}$, thus

$$| U |_{CL} = \sup_{\substack{\psi \in CL: \\ |\psi| \leqslant 1}} | U\psi |.$$

If U is a linear operator on a linear space W over the complex numbers, and if λ is a complex number, $D(\lambda)$ denotes the set of all $x \in W$ such that $Ux = \lambda x$. Obviously $D(\lambda)$ is a linear subspace of W and always contains 0. If $D(\lambda)$ contains an element $x \neq 0$, λ is an *eigenvalue* of U.

One of the mathematical cornerstones of this paper is the following lemma, which is a specialization of a uniform ergodic theorem of Ionescu Tulcea and Marinescu (1950, Sec. 9) along lines suggested by these authors (Ionescu Tulcea and Marinescu, 1950, Sec. 10).

LEMMA 5.1. *Let U be a linear operator on CL such that* (i) $| U |_{CL} \leqslant 1$,

(ii) *U is bounded with respect to the norm* $\| \cdot \|$, *and*

(iii) *for some positive integer k and real numbers* $0 \leqslant r < 1$ *and* $R < \infty$

$$m(U^k\psi) \leqslant rm(\psi) + R | \psi |,$$

for all $\psi \in CL$. Then

(a) *there are at most a finite number of eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ of U for which* $| \lambda_i | = 1$,

(b) *for all positive integers n*

$$U^n = \sum_{i=1}^{p} \lambda_i^n U_i + V^n,$$

where V and the U_i are linear operators on CL, bounded with respect to $\| \cdot \|$,

(c) $U_i^2 = U_i$, $U_iU_j = 0$ *for* $i \neq j$, $U_iV = VU_i = 0$,

(d) $D(\lambda_i) = U_i(CL)$ *is finite dimensional,* $i = 1, \ldots, p$, *and*

(e) *for some* $M < \infty$ *and* $h > 0$,

$$\| V^n \| \leqslant M/(1 + h)^n,$$

for all positive integers n.

This lemma will be applied to the restriction to CL of the bounded linear operator

$$U\psi(s) = E_s[\psi(S_2)] = \sum_{e \in E} \psi(f_e(s))\, \varphi_e(s) \tag{5.1}$$

on $C(S)$ associated with any distance diminishing model. This operator is of interest because its $(n-1)$st iterate, applied to a function $\psi \in C(S)$, gives the expectation of $\psi(S_n)$ as a function of the initial state; that is,

$$E_s[\psi(S_n)] = U^{n-1}\psi(s), \tag{5.2}$$

$n \geqslant 1$. This formula is easily proved by induction. It holds by definition for $n = 1$ and $n = 2$, and, if it holds for an $n \geqslant 1$, then

$$\begin{aligned} E_s[\psi(S_{n+1})] &= E_s[E[\psi(S_{n+1}) \mid S_2]] \\ &= E_s[U^{n-1}\psi(S_2)] = U^n\psi(s). \end{aligned}$$

THEOREM 5.1. *The conclusions* (a)–(e) *of Lemma* 5.1 *hold for the operator* $U\psi(s) = E_s[\psi(S_2)]$ *associated with any distance diminishing model. In addition,* 1 *is an eigenvalue of* U *and* $D(1)$ *contains all constant functions on* S.

Throughout the rest of this paper it will be assumed, without loss of generality, that $\lambda_1 = 1$ and $\lambda_j \neq 1, j = 2,..., p$, where the λ_i's are the eigenvalues of U of modulus 1.

Proof. The last statement of the theorem is obvious. It thus remains only to verify the hypotheses of Lemma 5.1. For any $\psi \in CL$

$$U\psi(s) - U\psi(s') = \sum_e (\psi(f_e(s)) - \psi(f_e(s')))\, \varphi_e(s) + \sum_e \psi(f_e(s'))(\varphi_e(s) - \varphi_e(s')),$$

so

$$\begin{aligned} | U\psi(s) - U\psi(s')| &\leqslant \sum_e | \psi(f_e(s)) - \psi(f_e(s'))|\, \varphi_e(s) + \sum_e | \psi(f_e(s'))|\, | \varphi_e(s) - \varphi_e(s')| \\ &\leqslant m(\psi) \sum_e d(f_e(s), f_e(s'))\, \varphi_e(s) + | \psi | \left(\sum_e m(\varphi_e)\right) d(s, s') \\ &\leqslant \left[m(\psi) \sum_e \varphi_e(s) + | \psi | \left(\sum_e m(\varphi_e)\right)\right] d(s, s'), \end{aligned}$$

by H7. Thus $U\psi \in CL$ as a consequence of H6, and

$$m(U\psi) \leqslant m(\psi) + | \psi | \sum_e m(\varphi_e). \tag{5.3}$$

Clearly

$$| U\psi(s)| \leqslant \sum_e | \psi(f_e(s))|\, \varphi_e(s) \leqslant | \psi |,$$

so

$$| U\psi | \leqslant | \psi |.$$

Therefore (i) is satisfied, and

$$\| U\psi \| \leqslant m(\psi) + |\psi| \Big(1 + \sum_e m(\varphi_e)\Big)$$

$$\leqslant \Big(1 + \sum_e m(\varphi_e)\Big) \| \psi \|,$$

so (ii) is satisfied also.

Hypothesis (iii) of Lemma 5.1 will now be verified. Let $k(s)$ be the integer and $e_{s,1}, \ldots, e_{s,k(s)}$ the events in the statement of H8. For each $s \in S$ let

$$Z(s) = \{t \in S: \varphi_{e_{s,1} \ldots e_{s,k(s)}}(t) > 0\}.$$

Since

$$\varphi_{e_1 \ldots e_n}(t) = \varphi_{e_1}(t)\, \varphi_{e_2}(f_{e_1}(t)) \cdots \varphi_{e_n}(f_{e_1 \ldots e_{n-1}}(t))$$

is continuous for any events $e_1, \ldots, e_n$, $Z(s)$ is open. Furthermore $s \in Z(s)$. Since S is compact the open covering $\{Z(s): s \in S\}$ has a finite subcovering $Z(s_1), \ldots, Z(s_N)$. Let $k_i = k(s_i)$ and let $K = \max_{1 \leqslant i \leqslant N} k_i$. If $t \in Z(s_i)$, let $e'_{t,j} = e_{s_i,j}$, $j = 1, \ldots, k_i$. Clearly it is possible to choose $e'_{t,k_i+1}, \ldots, e'_{t,K}$ in such a way that

$$\varphi_{e'_{t,1} \ldots e'_{t,K}}(t) > 0.$$

Hypothesis H7 implies $\mu(f_{e_1 \ldots e_j}) \leqslant 1$ for any events $e_1, \ldots, e_j$, so the inequality

$$\mu(f_{e'_{t,1} \ldots e'_{t,K}}) < 1$$

is obtained from H8. Thus the integer K, which does not depend on t, and the system $e'_{t,1}, \ldots, e'_{t,K}$ of events satisfy H8. Therefore it can be assumed without loss of generality that the integer k in H8 does not depend on the state.

If $\psi \in CL$ and $s, s' \in S$ then

$$U^k\psi(s) - U^k\psi(s') = \sum_{e_1 \cdots e_k} (\psi(f_{e_1 \ldots e_k}(s)) - \psi(f_{e_1 \ldots e_k}(s')))\, \varphi_{e_1 \ldots e_k}(s)$$
$$+ \sum_{e_1 \cdots e_k} \psi(f_{e_1 \ldots e_k}(s'))(\varphi_{e_1 \ldots e_k}(s) - \varphi_{e_1 \ldots e_k}(s')).$$

Therefore

$$| U^k\psi(s) - U^k\psi(s')| \leqslant \Big[m(\psi) \sum_{e_1 \cdots e_k} \mu(f_{e_1 \ldots e_k})\, \varphi_{e_1 \ldots e_k}(s) + |\psi| \sum_{e_1 \cdots e_k} m(\varphi_{e_1 \ldots e_k})\Big]\, d(s, s').$$

Now CL is closed under multiplication and under composition with mappings of S into S for which $\mu(f) < \infty$. Thus $\varphi_{e_1 \ldots e_j} \in CL$ for any events $e_1, \ldots, e_j$ and

$$m_j = \sum_{e_1 \cdots e_j} m(\varphi_{e_1 \ldots e_j}) < \infty. \tag{5.4}$$

It follows that for $s \neq s'$

$$| U^k\psi(s) - U^k\psi(s')|/d(s, s')$$

$$\leqslant m(\psi)\Big[\sum_{\substack{e_1\ldots e_k:\\ \mu(f_{e_1\ldots e_k})<1}} + \sum_{\substack{e_1\ldots e_k:\\ \mu(f_{e_1\ldots e_k})=1}}\Big] \mu(f_{e_1\ldots e_k})\,\varphi_{e_1\ldots e_k}(s) + |\psi|\, m_k$$

$$\leqslant m(\psi)\Big[\lambda \sum_{\substack{e_1\ldots e_k:\\ \mu(f_{e_1\ldots e_k})<1}} + \sum_{\substack{e_1\ldots e_k:\\ \mu(f_{e_1\ldots e_k})=1}}\Big] \varphi_{e_1\ldots e_k}(s) + |\psi|\, m_k$$

$$\leqslant m(\psi)(\lambda\Delta + (1 - \Delta)) + |\psi|\, m_k\,,$$

where

$$\lambda = \max_{\substack{e_1\ldots e_k:\\ \mu(f_{e_1\ldots e_k})<1}} \mu(f_{e_1\ldots e_k}) < 1 \tag{5.5}$$

and

$$\Delta = \min_{s\in S} \sum_{\substack{e_1\ldots e_k:\\ \mu(f_{e_1\ldots e_k})<1}} \varphi_{e_1\ldots e_k}(s) > 0. \tag{5.6}$$

The latter inequality is a consequence of H8, H5, and the continuity of the $\varphi_{e_1\ldots e_k}$. Therefore

$$m(U^k\psi) \leqslant rm(\psi) + m_k\, |\psi| \tag{5.7}$$

where $r = \lambda\Delta + (1 - \Delta) < 1$, and (iii) of Lemma 5.1 is satisfied. Q.E.D.

Though it constitutes a slight digression and will not be used in the sequel, an additional consequence of (5.3) and (5.7) is worth pointing out. From these inequalities

$$m(U^n\psi) \leqslant m(\psi) + n\, |\psi|\, m_1 \tag{5.8}$$

and

$$m(U^{nk}\psi) \leqslant m(\psi) + |\psi|\, m_k/(1 - r), \tag{5.9}$$

$n \geqslant 0$, are easily obtained by induction. These yield, on combination,

$$m(U^j\psi) \leqslant m(\psi) + |\psi|((k - 1)\, m_1 + m_k/(1 - r)), \tag{5.10}$$

valid for all $\psi \in CL$ and $j \geqslant 0$. It follows that $\{U^j\psi\}$ is equicontinuous. This, together with the fact that CL is dense in $C(S)$ (as a consequence of the Stone–Weierstrass Theorem) and $|U| = 1$, implies that, for any $\psi \in C(S)$, $\{U^j\psi\}$ is equicontinuous. In the terminology of Jamison (1964, 1965) the operator U on $C(S)$ associated with any distance diminishing model is *uniformly stable*. In the terminology of Feller (1966) the corresponding stochastic kernel K is *regular*.

B. Proof of Theorem 2.1

The following lemma includes most of the assertions of Theorem 2.1.

Lemma 5.2. *For any distance diminishing model there is a stochastic kernel* K^∞ *such that* (2.5) *holds, where* $E_s[\psi(S_\infty)]$ *is given by* (2.6). *For any Borel subset* A *of* S, $K^\infty(\cdot, A) \in CL$.

Proof of Lemma 5.2. Theorem 5.1 implies that

$$\| U^n \| \leqslant \sum_{i=1}^{p} \| U_i \| + \| V^n \| \to \sum_{i=1}^{p} \| U_i \|$$

as $n \to \infty$. Therefore there is a constant $W < \infty$ such that

$$\| U^n \| < W \tag{5.11}$$

for all $n \geqslant 0$.

Let $\bar{U}_n = (1/n) \sum_{j=0}^{n-1} U^j$. Then, by Theorem 5.1,

$$\begin{aligned} \bar{U}_n &= (1/n)(I - U^n) + (1/n) \sum_{j=1}^{n} U^j \\ &= (1/n)(I - U^n) + (1/n) \sum_{i=1}^{p} \left(\sum_{j=1}^{n} \lambda^j \right) U_i + (1/n) \sum_{j=1}^{n} V^j. \end{aligned}$$

Therefore

$$\bar{U}_n - U_1 = (1/n)(I - U^n) + (1/n) \sum_{i=2}^{p} \lambda_i[(1 - \lambda_i^n)/(1 - \lambda_i)]\, U_i + (1/n) \sum_{j=1}^{n} V^j,$$

so that

$$\| \bar{U}_n - U_1 \| \leqslant C/n$$

where

$$C = (1 + W) + 2 \sum_{i=2}^{p} \| U_i \| / | 1 - \lambda_i | + M/h.$$

Thus, for any $\psi \in CL$,

$$\| \bar{U}_n\psi - U_1\psi \| \leqslant \| \bar{U}_n - U_1 \| \| \psi \| \leqslant C \| \psi \| / n, \tag{5.12}$$

and, *a fortiori*,

$$\lim_{n\to\infty} | \bar{U}_n\psi - U_1\psi | = 0 \tag{5.13}$$

for all $\psi \in CL$.

Since CL is dense in $C(S)$ and $| \bar{U}_n | = 1$ for $n \geqslant 1$, it follows that (5.13) holds for all $\psi \in C(S)$, where U_1 has been extended (uniquely) to a bounded linear operator on $C(S)$. Since the operators $\bar{U}_n$ on $C(S)$ are all positive and preserve constants, (5.13) implies that the same is true of U_1. Thus, for any $s \in S$, $U_1\psi(s)$ is a positive linear functional on $C(S)$ with $U\Gamma(s) = 1$ where $\Gamma(s) \equiv 1$. Hence, by the Riesz representation theorem, there is a (unique) Borel probability measure $K^\infty(s, \cdot)$ on S such that

$$U_1\psi(s) = \int_S \psi(s')\, K^\infty(s, ds') \tag{5.14}$$

for all $\psi \in C(S)$. In view of (5.14), (5.12) reduces to (2.5).

That K^∞ is a stochastic kernel follows from the fact, now to be proved, that $K^\infty(\cdot, A) \in CL$ for every Borel set A. This is obviously true if $A = S$. Suppose that A is an open set such that its complement $\tilde{A}$ is not empty. For $j \geqslant 1$ define $\eta_j \in CL$ by

$$\eta_j(s) = \begin{cases} 1, \text{ if } d(s, \tilde{A}) \geqslant 1/j, \\ j\, d(s, \tilde{A}), \text{ if } d(s, \tilde{A}) \leqslant 1/j. \end{cases} \tag{5.15}$$

Then

$$\lim_{j\to\infty} \eta_j(s) = \begin{cases} 1, \text{ if } s \in A \\ 0, \text{ if } s \in \tilde{A} \end{cases} = I_A(s), \tag{5.16}$$

where I_A is the indicator function of the set A, and the convergence is monotonic. Therefore

$$\lim_{j\to\infty} U_1\eta_j(s) = \int_S I_A(s')\, K^\infty(s, ds') = K^\infty(s, A) \tag{5.17}$$

for all $s \in S$. By Theorem 5.1, $D(1) = U_1(CL)$ is a finite dimensional subspace of CL. Hence there exists a constant $J < \infty$ such that $\| \psi \| \leqslant J \,|\, \psi \,|$ for all $\psi \in D(1)$. Therefore

$$| U_1\eta_j(s_1) - U_1\eta_j(s_2)| \leqslant m(U_1\eta_j)\, d(s_1, s_2) \leqslant J \,|\, U_1\eta_j \,|\, d(s_1, s_2) \leqslant J\, d(s_1, s_2),$$

for all $j \geqslant 1$ and $s_1, s_2 \in S$. Equation 5.17 then yields, on letting j approach ∞,

$$| K^\infty(s_1, A) - K^\infty(s_2, A)| \leqslant J\, d(s_1, s_2).$$

If A is an arbitrary Borel set, $s_1, s_2 \in S$, and $\epsilon > 0$, the regularity of $K^\infty(s_i, \cdot)$ insures the existence of an open set $A_{i,\epsilon}$ such that $A_{i,\epsilon} \supset A$ and

$$K^\infty(s_i, A_{i,\epsilon}) - K^\infty(s_i, A) \leqslant \epsilon,$$

for $i = 1, 2$. Thus $A_\epsilon = A_{1,\epsilon} \cap A_{2,\epsilon}$ is open and

$$0 \leqslant K^\infty(s_i, A_\epsilon) - K^\infty(s_i, A) \leqslant \epsilon,$$

$i = 1, 2$. Combination of these inequalities with the result of the last paragraph yields

$$| K^\infty(s_1, A) - K^\infty(s_2, A)| \leqslant J\, d(s_1, s_2) + 2\epsilon,$$

or, since ϵ is arbitrary,

$$| K^{\infty}(s_1, A) - K^{\infty}(s_2, A)| \leqslant J\, d(s_1, s_2).$$

Thus $K^{\infty}(\cdot, A) \in CL$ with $m(K^{\infty}(\cdot, A)) \leqslant J$ for all Borel subsets A of S. Q.E.D.

Actually, this proof gives (2.5) for the complex as well as real valued functions ψ, though this is not important here.

To complete the proof of Theorem 2.1 it remains only to prove

LEMMA 5.3. *The stochastic kernel* $(1/n) \sum_{j=0}^{n-1} K^{(j)}$ *converges uniformly to* K^{∞}.

Proof. Denote $(1/n) \sum_{j=0}^{n-1} K^{(j)}$ by $\bar{K}_n$.
Since

$$\bar{K}_n(s, \mathring{G}) \leqslant \bar{K}_n(s, G) \leqslant \bar{K}_n(s, \bar{G}) \tag{5.18}$$

it suffices to show that if A is open,

$$K^{\infty}(s, A) - \epsilon \leqslant \bar{K}_n(s, A) \tag{5.19}$$

for all s if n is sufficiently large, while if B is closed,

$$\bar{K}_n(s, B) \leqslant K^{\infty}(s, B) + \epsilon$$

for all s if n is sufficiently large. The statement concerning closed sets follows from that concerning open sets by taking complements, so only open sets need be considered. There is no loss in generality in assuming $A \neq S$. By (5.15), $I_A(t) \geqslant \eta_j(t)$ for all $t \in S$ so

$$\begin{aligned} \bar{K}_n(s, A) &\geqslant \bar{U}_n \eta_j(s) \\ &= K^{\infty}(s, A) + [U_1 \eta_j(s) - K^{\infty}(s, A)] + [\bar{U}_n \eta_j(s) - U_1 \eta_j(s)] \\ &\geqslant K^{\infty}(s, A) - | U_1 \eta_j(\cdot) - K^{\infty}(\cdot, A)| - | \bar{U}_n \eta_j - U_1 \eta_j |. \end{aligned}$$

Since the convergence in (5.17) is monotonic and the limit is continuous, convergence is uniform by Dini's theorem. Choose j so large that

$$| U_1 \eta_j(\cdot) - K^{\infty}(\cdot, A)| < \epsilon/2.$$

Then (5.13) applied to $\psi = \eta_j$ implies (5.19) for all $s \in S$ if n is sufficiently large. Q.E.D.

Theorem 5.1 asserts that $D(1)$, the linear space of $\psi \in CL$ such that $U\psi = \psi$, contains all constant functions. If, in addition, it is known that $D(1)$ is one dimensional; that is, the only $\psi \in CL$ such that $U\psi = \psi$ are constants, then it can be concluded that the probability measure $K^{\infty}(s, \cdot) = K^{\infty}(\cdot)$ does not depend on s. For, by (d) of

Lemma 5.1, $U_1\psi$ is a constant function for any $\psi \in CL$, and thus for any $\psi \in C(S)$. Therefore, in view of (5.14), for any $s, s' \in S$,

$$\int_S \psi(t)\, K^\infty(s, dt) = \int_S \psi(t)\, K^\infty(s', dt),$$

for all $\psi \in C(S)$. This implies that $K^\infty(s, \cdot) = K^\infty(s', \cdot)$, as claimed. It is, incidentally, easy to show that $K^\infty(\cdot)$ is the unique stationary probability distribution of $\{S_n\}$, from which it follows (Breiman, 1960) that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \psi(S_j) = E[\psi(S_\infty)]$$

with probability 1, for any $\psi \in C(S)$ and any initial state.

C. Proof of Theorem 2.2

Suppose that a distance diminishing model has the property that the associated operator U on CL has no eigenvalues of modulus 1 other than 1. Then Theorem 5.1 implies that

$$\| U^n - U_1 \| = \| V^n \| \leqslant M/(1+h)^n.$$

Therefore, for any $\psi \in CL$

$$\| U^n\psi - U_1\psi \| \leqslant M \| \psi \| /(1+h)^n, \tag{5.20}$$

and, for any $\psi \in C(S)$,

$$\lim_{n\to\infty} | U^n\psi - U_1\psi | = 0. \tag{5.21}$$

From (5.21) it follows that $K^{(n)}$ converges uniformly to K^∞, just as uniform convergence of $\bar{K}_n$ to K^∞ followed from (5.13) in Subsec. 5B. That is to say, when (5.21) holds the proof of Lemma 5.3 remains valid if $\bar{K}_n$ and $\bar{U}_n$ are everywhere replaced by $K^{(n)}$ and U^n. If, in addition, the only $\psi \in CL$ for which $U\psi = \psi$ are constants, then $K^\infty(s, \cdot) = K^\infty(\cdot)$ does not depend on s, as was shown in the last paragraph of Subsec. 5B. Therefore (5.20) reduces to (2.9) (with $C = (1+h)M$ and $\alpha = 1/(1+h)$), and all of the conclusions of Theorem 2.2 hold. To complete the proof of Theorem 2.2 it thus suffices to prove the following two lemmas.

Lemma 5.4 *If a distance diminishing model satisfies* H9, *then* 1 *is the only eigenvalue of* U *of modulus* 1.

Lemma 5.5 *If a distance diminishing model satisfies* H9, *then the only continuous solutions of* $U\psi = \psi$ *are constants.*

The arguments given below follow similar arguments by Jamison (1964).

Proof of Lemma 5.4. Suppose $|\lambda| = 1$, $\lambda \neq 1$, $\psi \neq 0$, and $\psi \in D(\lambda)$ so that $U\psi = \lambda\psi$. $|\psi(\cdot)| \in C(S)$ so there is an $s_0 \in S$:

$$|\psi(s_0)| = \max_{s\in S} |\psi(s)| = |\psi|.$$

Clearly $\psi(s_0) \neq 0$. Now $U\psi' = \lambda\psi'$ where $\psi' = \psi/\psi(s_0)$, $|\psi'| = |\psi|/|\psi(s_0)| = 1$ and $\psi'(s_0) = 1$. $U^n\psi' = \lambda^n\psi'$ so $U^n\psi'(s_0) = \lambda^n$. For $n = 1, 2, \ldots$, let $B_n = \{s: \psi'(s) = \lambda^n\}$. Clearly $K^{(n)}(s_0, B_n) = 1$ for all $n \geqslant 1$. Since $K(s_0, B_1) = 1$, B_1 is not empty. Let $s_1 \in B_1$. Then $\psi'(s_1) = \lambda$, $U^n\psi'(s_1) = \lambda^{n+1}$, and $K^{(n)}(s_1, B_{n+1}) = 1$. But $|\lambda^{n+1} - \lambda^n| = |\lambda - 1| > 0$. Since ψ' is uniformly continuous there exists $\delta > 0$ such that $d(s', s'') < \delta$ implies $|\psi'(s') - \psi'(s'')| < |\lambda - 1|$. If $s' \in B_n$ and $s'' \in B_{n+1}$ then $|\psi'(s') - \psi'(s'')| = |\lambda - 1|$ so $d(s', s'') \geqslant \delta$. Therefore $d(B_n, B_{n+1}) \geqslant \delta$, $n = 1, 2, \ldots$. But $B_n \supset T_n(s_0)$ and $B_{n+1} \supset T_n(s_1)$ so $d(T_n(s_0), T_n(s_1)) \geqslant \delta$, $n = 1, 2, \ldots$. Thus H9 is violated. Q.E.D.

Proof of Lemma 5.5. Suppose that there exists a real-valued nonconstant function $\psi \in C(S)$ such that $U\psi = \psi$. Let $M = \max \psi = \psi(s_0)$ and $m = \min \psi = \psi(s_1)$. Then $M > m$. Let $C_m = \{s: \psi(s) = m\}$ and $C_M = \{s: \psi(s) = M\}$. $U^n\psi = \psi$, so $U^n\psi(s_0) = M$ and $U^n\psi(s_1) = m$. Therefore $K^{(n)}(s_0, C_M) = 1$ and $K^{(n)}(s_1, C_m) = 1$, so $C_M \supset T_n(s_0)$ and $C_m \supset T_n(s_1)$ for all $n \geqslant 1$. By the uniform continuity of ψ there exists a $\delta > 0$ such that $d(s, s') < \delta$ implies $|\psi(s) - \psi(s')| < M - m$. If $s \in C_M$ and $s' \in C_m$ then $|\psi(s) - \psi(s')| = M - m$, so $d(s, s') \geqslant \delta$. Therefore $d(T_n(s_0), T_n(s_1)) \geqslant \delta$ for all $n \geqslant 1$. Thus H9 is violated. So under the hypotheses of Lemma 5.5 there is no real-valued nonconstant $\psi \in C(S)$ for which $U\psi = \psi$.

Suppose $\psi' \in C(S)$, $U\psi' = \psi'$. Then

$$U \operatorname{Re} \psi' + iU \operatorname{Im} \psi' = \operatorname{Re} \psi' + i \operatorname{Im} \psi'.$$

Thus $U \operatorname{Re} \psi' = \operatorname{Re} \psi'$ and $U \operatorname{Im} \psi' = \operatorname{Im} \psi'$. But $\operatorname{Re} \psi'$ and $\operatorname{Im} \psi'$ are continuous and real valued, so $\operatorname{Re} \psi'$ and $\operatorname{Im} \psi'$ are constants. Thus ψ' is a (complex valued) constant function. Therefore all continuous solutions of $U\psi = \psi$ are constants. Q.E.D.

D. Proof of Theorem 2.3

The first paragraph of Subsec. 5C shows that, to obtain the uniform convergence of $K^{(n)}$ to the limiting kernel K^∞ of Theorem 2.1, and to obtain (2.10) with $E_s[\psi(S_\infty)]$ defined as in (2.6), it suffices to prove Lemma 5.6 below. All lemmas in this subsection refer to a distance diminishing model satisfying H10.

Lemma 5.6. *U has no eigenvalues of modulus* 1 *other than* 1.

Proof. Suppose $U\psi = \lambda\psi$ where $|\lambda| = 1$, $\lambda \neq 1$, and $\psi \in C(S)$. Let s_0 be a state for which $|\psi(s_0)| = |\psi|$, and let $C_n = \{s: \psi(s) = \lambda^n\psi(s_0)\}$, $n = 1, 2....$ Now $U^n\psi(s_0) = \lambda^n\psi(s_0)$, thus $K^{(n)}(s_0, C_n) = 1$, and $C_n \supset T_n(s_0)$. By H10 there exists a sequence $\{t_n\}$ such that $t_n \in T_n(s_0)$ and $\lim_{n\to\infty} d(t_n, a_{j(s_0)}) = 0$. Hence $\lim_{n\to\infty} \psi(t_n) = \psi(a_{j(s_0)})$. But $t_n \in C_n$, so $\psi(t_n) = \lambda^n\psi(s_0)$, which converges only if $\psi(s_0) = 0$. Hence $|\psi| = 0$ and $\psi(s) \equiv 0$. Thus λ is not an eigenvalue of U. Q.E.D.

The proof that S_n converges with probability 1 must be deferred until more information has been obtained about K^∞. The next two lemmas provide such information. In the work that follows, $A = \{a_i : 1 \leqslant i \leqslant N\}$ is the set of absorbing states.

LEMMA 5.7. *If $b_1, ..., b_N$ are any N scalars, there is one and only one $\psi \in C(S)$ such that $U\psi = \psi$ and $\psi(a_i) = b_i$, $i = 1, ..., N$. This function belongs to CL.*

Proof. (1) *Uniqueness.* First, the following maximum modulus principle will be proved: If $\psi \in C(S)$ and $U\psi = \psi$, then all maxima of $|\psi(\cdot)|$ occur on A (and possibly elsewhere). Let s_0 be a state such that $|\psi(s_0)| = |\psi|$, and let $C = \{s: \psi(s) = \psi(s_0)\}$. Since $U^n\psi(s_0) = \psi(s_0)$, $K^{(n)}(s_0, C) = 1$, so $C \supset T_n(s_0)$. By H10 there exists a sequence t_n such that $t_n \in T_n(s_0)$ and $\lim_{n\to\infty} d(t_n, a_{j(s_0)}) = 0$. Hence $\lim_{n\to\infty} \psi(t_n) = \psi(a_{j(s_0)})$. But $t_n \in C$, so $\psi(t_n) = \psi(s_0)$. Thus $\psi(s_0) = \psi(a_{j(s_0)})$, and $|\psi(a_{j(s_0)})| = |\psi|$.

Suppose now that $\psi, \psi' \in C(S)$, $U\psi = \psi$, $U\psi' = \psi'$, and $\psi(s) = \psi'(s)$ for all $s \in A$. Let $\psi'' = \psi - \psi'$. Then $\psi'' \in C(S)$, $U\psi'' = \psi''$, and $\psi''(s) = 0$ for all $s \in A$. Thus $|\psi''| = 0$, so $\psi(s) \equiv \psi'(s)$.

(2) *Existence.* Since $U^n\psi(s) = E_s[\psi(S_{n+1})]$, $U^n\psi(s) = \psi(s)$ for all $s \in A$ and $\psi \in C(S)$. Thus $U_1\psi(s) = \lim_{n\to\infty} U^n\psi(s) = \psi(s)$ for all $s \in A$.

Let $\omega_1, ..., \omega_N \in CL$ with $\omega_i(a_j) = \delta_{ij}$, e.g.,

$$\omega_i(s) = (1 - \epsilon^{-1} d(s, a_i))^+,$$

where $\epsilon = \min_{i \neq j} d(a_i, a_j)$ and x^+ is x or 0 depending on whether $x \geqslant 0$ or $\leqslant 0$. It will now be shown that

$$\gamma(s) = \sum_{i=1}^{N} b_i U_1 \omega_i(s)$$

is the function sought. Clearly $\gamma \in CL$, and

$$\gamma(a_j) = \sum_{i=1}^{N} b_i U_1 \omega_i(a_j) = \sum_{i=1}^{N} b_i \omega_i(a_j) = b_j .$$

Finally,

$$U\gamma = \sum_{i=1}^{N} b_i U U_1 \omega_i = \sum_{i=1}^{N} b_i U_1 \omega_i = \gamma. \qquad \text{Q.E.D.}$$

LEMMA 5.8. *For $i = 1, \ldots, N$, let γ_i be the continuous function such that $\gamma_i(a_j) = \delta_{ij}$ and $U\gamma_i = \gamma_i$. Then*

$$K^\infty(s, \cdot) = \sum_{i=1}^{N} \gamma_i(s)\, \delta_{a_i}(\cdot),$$

where $\delta_{s'}$ is the Borel probability measure on S concentrated at s', and, for any $\psi \in C(S)$ and all $s \in S$,

$$E_s[\psi(S_\infty)] = \sum_{i=1}^{N} \gamma_i(s)\, \psi(a_i).$$

Proof. For $\psi \in C(S)$ let $\bar{\psi}(s) = \sum_{i=1}^{N} \gamma_i(s)\, \psi(a_i)$ and $\psi' = U_1\psi$. Clearly $\bar{\psi}, \psi' \in C(S)$, and $\psi'(a_j) = \psi(a_j) = \bar{\psi}(a_j), j = 1, \ldots, N$. Also

$$U\bar{\psi} = \sum_{i=1}^{N} \psi(a_i)\, U\gamma_i = \sum_{i=1}^{N} \psi(a_i)\, \gamma_i = \bar{\psi},$$

and

$$U\psi' = UU_1\psi = U_1\psi = \psi'.$$

Thus, by Lemma 5.7, $\bar{\psi} = \psi'$, which is the second assertion of Lemma 5.8.

Now

$$\int_S \psi(s')\, K^\infty(s, ds') = E_s[\psi(S_\infty)] = \sum_{i=1}^{N} \gamma_i(s)\, \psi(a_i)$$

$$= \sum_{i=1}^{N} \gamma_i(s) \int_S \psi(t)\, \delta_{a_i}(dt)$$

$$= \int_S \psi(t) \left(\sum_{i=1}^{N} \gamma_i(s)\, \delta_{a_i}(dt) \right)$$

for all $\psi \in C(S)$. This yields the first assertion of the lemma. Q.E.D.

Now that it is known that $K^\infty(s, \cdot)$ is concentrated on A and the functions γ_i are available, probability 1 convergence of S_n can be proved.

LEMMA 5.9. *For any initial state s, $\{S_n\}$ converges with probability 1 to a random point S_∞ of A. For any Borel subset B of S, $K^\infty(s, B) = P_s(S_\infty \in B)$. In particular,*

$$\gamma_i(s) = P_s(S_\infty = a_i),$$

$i = 1, \ldots, N$.

Proof. It is a simple consequence of the triangle inequality that the function

$d(\cdot, A)$ on S belongs to CL, and clearly $d(a_i, A) = 0$, $i = 1,..., N$. Thus $E_s[d(S_\infty, A)] \equiv 0$, so that

$$\| E.[d(S_n, A)]\| \leqslant C\alpha^n \| d(\cdot, A)\|$$

for all $n \geqslant 1$. The initial state is regarded as fixed throughout the following discussion. Since

$$E_s[d(S_n, A)] \leqslant C\alpha^n \| d(\cdot, A)\|$$

it follows that

$$\sum_{n=1}^{\infty} E_s[d(S_n, A)] \leqslant C \| d(\cdot, A)\| \alpha/(1 - \alpha).$$

By the monotone convergence theorem the order of summation and expectation on the left can be interchanged to obtain

$$E_s\left[\sum_{n=1}^{\infty} d(S_n, A)\right] < \infty.$$

Therefore $\sum_{n=1}^{\infty} d(S_n, A) < \infty$, and, consequently, $\lim_{n\to\infty} d(S_n, A) = 0$ with probability 1.

For any $i = 1,..., N$,

$$E_s[\gamma_i(S_{n+1}) | S_n, ..., S_1] = E_s[\gamma_i(S_{n+1}) | S_n] = \gamma_i(S_n),$$

so $\{\gamma_i(S_n)\}$ is a martingale. Since it is bounded (by $| \gamma_i |$), it converges with probability 1.

Let G be the event "$\lim_{n\to\infty} d(S_n, A) = 0$ and $\lim_{n\to\infty} \gamma_i(S_n)$ exists, $i = 1,..., N$" in the underlying sample space. The above arguments show that $P_s(G) = 1$. Let $\omega \in G$. Since S is compact, every subsequence of $\{S_n(\omega)\}$ has a convergent subsequence, and, since $d(S_n(\omega), A) \to 0$ as $n \to \infty$, all subsequential limit points of $\{S_n(\omega)\}$ are in A. Suppose that a_i and $a_{i'}$, $i \neq i'$, are two distinct subsequential limit points—say $S_{n_j}(\omega) \to a_i$ and $S_{n_j'}(\omega) \to a_{i'}$, as $j \to \infty$. Then

$$\gamma_i(S_{n_j}(\omega)) \to \gamma_i(a_i) = 1,$$

and

$$\gamma_i(S_{n_j'}(\omega)) \to \gamma_i(a_{i'}) = 0,$$

which contradicts the convergence of $\{\gamma_i(S_n(\omega))\}$. Thus all convergent subsequences of $\{S_n(\omega)\}$ converge to the same point of A. Denote this point $S_\infty(\omega)$. It follows that $\lim_{n\to\infty} S_n(\omega) = S_\infty(\omega)$. Therefore $\lim_{n\to\infty} S_n = S_\infty$ with probability 1. This implies that the asymptotic distribution of S_n is the same as the distribution of S_∞, i.e., $K^\infty(s, B) = P_s(S_\infty \in B)$ for all Borel subsets B of S. Finally $\gamma_i(s) = P_s(S_\infty = a_i)$ follows by taking $B = \{a_i\}$. Q.E.D.

This completes the proof of Theorem 2.3.

6. PROOFS OF THEOREMS CONCERNING EVENTS

A. Proof of Theorem 2.4

The equality

$$P_s^{(n)}(A^\ell) = E_s[P((E_n, ..., E_{n+\ell-1}) \in A^\ell \mid S_n)]$$

can be rewritten in the form

$$P_s^{(n)}(A^\ell) = E_s[\psi(S_n)], \tag{6.1}$$

where

$$\psi(s) = P_s^{(1)}(A^\ell). \tag{6.2}$$

Thus (2.12), with $L = C(D + 1)$, follows from (2.9), (2.10), and the following lemma.

Lemma 6.1. *For any distance diminishing model there is a constant D such that*

$$m(P_{\cdot}^{(1)}(A^\ell)) \leqslant D \tag{6.3}$$

for all $\ell \geqslant 1$ *and* $A^\ell \subset E^\ell$.

Proof. For any $i, j \geqslant 1$, $s, s' \in S$, and $A^{i+j} \in E^{i+j}$

$$P_s^{(1)}(A^{i+j}) - P_{s'}^{(1)}(A^{i+j}) = \sum_{e_1 \cdots e_i} \varphi_{e_1 \ldots e_i}(s) \, P_{f_{e_1 \ldots e_i}(s)}^{(1)}(A_{e_1 \ldots e_i}^{i+j})$$
$$- \sum_{e_1 \cdots e_i} \varphi_{e_1 \ldots e_i}(s') \, P_{f_{e_1 \ldots e_i}(s')}^{(1)}(A_{e_1 \ldots e_i}^{i+j}),$$

where

$$A_{e_1 \ldots e_i}^{i+j} = \{(e_{i+1}, ..., e_{i+j}): (e_1, ..., e_{i+j}) \in A^{i+j}\},$$

$$= \sum_{e_1 \cdots e_i} \varphi_{e_1 \ldots e_i}(s) (P_{f_{e_1 \ldots e_i}(s)}^{(1)}(A_{e_1 \ldots e_i}^{i+j}) - P_{f_{e_1 \ldots e_i}(s')}^{(1)}(A_{e_1 \ldots e_i}^{i+j}))$$
$$+ \sum_{e_1 \cdots e_i} (\varphi_{e_1 \ldots e_i}(s) - \varphi_{e_1 \ldots e_i}(s')) \, P_{f_{e_1 \ldots e_i}(s')}^{(1)}(A_{e_1 \ldots e_i}^{i+j}).$$

Thus

$$| P_s^{(1)}(A^{i+j}) - P_{s'}^{(1)}(A^{i+j}) | \leqslant \sum_{e_1 \cdots e_i} \varphi_{e_1 \ldots e_i}(s) \mid P_{f_{e_1 \ldots e_i}(s)}^{(1)}(A_{e_1 \ldots e_i}^{i+j}) - P_{f_{e_1 \ldots e_i}(s')}^{(1)}(A_{e_1 \ldots e_i}^{i+j}) |$$
$$+ \sum_{e_1 \cdots e_i} | \varphi_{e_1 \ldots e_i}(s) - \varphi_{e_1 \ldots e_i}(s') | \, P_{f_{e_1 \ldots e_i}(s')}^{(1)}(A_{e_1 \ldots e_i}^{i+j})$$
$$\leqslant n_j \sum_{e_1 \cdots e_i} \varphi_{e_1 \ldots e_i}(s) \, \mu(f_{e_1 \ldots e_i}) \, d(s, s') + m_i \, d(s, s'),$$

where

$$n_j = \max_{A^j \subset E^j} m(P^{(1)}_{\cdot}(A^j)),$$

and m_i is given by (5.4). (Note that $n_1 \leqslant m_1$.)

Two cases are now distinguished.

CASE 1. $i = 1$. Then

$$| P^{(1)}_s(A^{1+j}) - P^{(1)}_{s'}(A^{1+j})| \leqslant (n_j + m_1)\, d(s, s'),$$

so $n_{j+1} \leqslant n_j + m_1$ or, by induction,

$$n_j \leqslant j m_1 . \tag{6.4}$$

CASE 2. $i = k$, where k is an integer that satisfies H8 for all s. It was shown in the proof of Theorem 5.1 that such an integer exists, and that there is a constant $0 \leqslant r < 1$ such that

$$\sum_{e_1 \ldots e_k} \varphi_{e_1 \ldots e_k}(s)\, \mu(f_{e_1 \ldots e_k}) \leqslant r$$

for all $s \in S$. Thus

$$| P^{(1)}_s(A^{k+j}) - P^{(1)}_{s'}(A^{k+j})| \leqslant (n_j r + m_k)\, d(s, s'),$$

so

$$n_{j+k} \leqslant n_j r + m_k . \tag{6.5}$$

This formula and a simple induction on ν imply

$$n_{j+\nu k} \leqslant n_j r^\nu + m_k \left(\sum_{\ell=0}^{\nu-1} r^\ell \right)$$

for $\nu \geqslant 0$. Thus

$$\begin{aligned} n_{j+\nu k} &\leqslant n_j r^\nu + m_k/(1 - r) \\ &\leqslant j m_1 r^\nu + m_k/(1 - r) \end{aligned}$$

by (6.4). But any positive integer ℓ can be represented as $\ell = \nu k + j$ for some $\nu \geqslant 0$ and $0 \leqslant j < k$. Thus

$$n_\ell \leqslant (k - 1)\, m_1 + m_k/(1 - r) = D$$

for all $\ell \geqslant 1$. Q.E.D.

B. Proof of Corollary 2.5

Under the hypotheses of the corollary,

$$P_s^{\infty}(A^{\ell}) = \sum_{i=1}^{N} P_{a_i}^{(1)}(A^{\ell})\, \gamma_i(s) = 0.$$

Thus (2.12) implies

$$\| P_{\cdot}^{(n)}(A^{\ell})\| \leqslant L\alpha^n,$$

for $n \geqslant 1$. Thus the series $\sum_{n=1}^{\infty} P_s^{(n)}(A^{\ell})$ converges in the norm $\| \cdot \|$ to an element of CL, and

$$\left\| \sum_{n=1}^{\infty} P_s^{(n)}(A^{\ell}) \right\| \leqslant \sum_{n=1}^{\infty} \| P_s^{(n)}(A^{\ell})\| \leqslant L\alpha/(1-\alpha). \tag{6.6}$$

Let

$$X_n = \begin{cases} 1 & \text{if } (E_n ,..., E_{n+\ell-1}) \in A^{\ell} \\ 0 & \text{if } (E_n ,..., E_{n+\ell-1}) \in' A^{\ell}. \end{cases}$$

Then

$$X = \sum_{n=1}^{\infty} X_n \quad \text{and} \quad E_s[X_n] = P_s^{(n)}(A^{\ell}),$$

so

$$E_s[X] = \sum_{n=1}^{\infty} P_s^{(n)}(A^{\ell}), \tag{6.7}$$

for all $s \in S$. This, in combination with (6.6), gives (2.14).

Clearly

$$\chi(s) = E_s[X_1] + E_s\left[\sum_{n=2}^{\infty} X_n\right]$$

$$= P_s^{(1)}(A^{\ell}) + E_s\left[E\left[\sum_{n=2}^{\infty} X_n \,\middle|\, S_2\right]\right]$$

$$= P_s^{(1)}(A^{\ell}) + E_s[\chi(S_2)],$$

and $\chi(a_i) = 0, i = 1,..., N$. If χ' is another continuous function satisfying these conditions then $\Delta = \chi - \chi' \in C(S)$ with $U\Delta = \Delta$ and $\Delta(a_i) = 0, i = 1,..., N$. Thus $\Delta(s) \equiv 0$ by Theorem 2.3. Q.E.D.

C. PROOF OF THEOREM 2.6

As was remarked in Sec. 2, a distance diminishing model can be regarded as an example of a homogeneous random system with complete connections. In the notational style of this paper (Iosifescu's is somewhat different), a homogeneous random system with complete connections is a system $((S, \mathscr{B}), (E, \mathscr{F}), f, \tilde{\varphi})$ such that $(S, \mathscr{B})$ and $(E, \mathscr{F})$ are measurable spaces, $f_{.}(\cdot)$ is a measurable mapping of $E \times S$ into S, $\tilde{\varphi}_{.}(s)$ is a probability measure on $\mathscr{F}$ for each $s \in S$, and $\tilde{\varphi}_A(\cdot)$ is a measurable real valued function on S for each $A \in \mathscr{F}$. An associated stochastic process $\{E_n\}$ for such a system satisfies

H3′

$$P_s(E_1 \in A) = \tilde{\varphi}_A(s)$$

and

$$P_s(E_{n+1} \in A \mid E_j = e_j\,,\, 1 \leqslant j \leqslant n) = \tilde{\varphi}_A(f_{e_1 \dots e_n}(s)),$$

for $n \geqslant 1$, $s \in S$, and $A \in \mathscr{F}$. Under H4, $\mathscr{B}$ can be taken to be the Borel subsets of S. Under H2, $\mathscr{F}$ can be taken to be all subsets of E, and $\tilde{\varphi}_{.}(s)$ is a probability measure on $\mathscr{F}$ if and only if there is a nonnegative real-valued function $\varphi_{.}(s)$ on E such that

$$\tilde{\varphi}_A(s) = \sum_{e \in A} \varphi_e(s)$$

and $\sum_{e \in E} \varphi_e(s) \equiv 1$. Then H3′ is equivalent to H3. Also, the measurability requirements on f and $\tilde{\varphi}_A$ are weaker than continuity of $f_e(\cdot)$ and $\varphi_e(\cdot)$, which are, in turn, weaker than H6 and H7.

The following lemma was proved (but not stated formally) by Iosifescu (1963, Chap. 3, Sec. 3).

LEMMA 6.2. *If a homogeneous random system with complete connections has the property that there is a sequence $\{\epsilon_n\}$ of positive numbers with $\sum_{n=1}^{\infty} n\epsilon_n < \infty$ and, for every $\ell \geqslant 1$, a probability measure $P^{\infty}(A^{\ell})$ on $\mathscr{F}^{\ell}$ such that*

$$| P_s^{(n)}(A^{\ell}) - P^{\infty}(A^{\ell})| < \epsilon_n$$

for all $s \in S$, $n, \ell \geqslant 1$, and $A^{\ell} \in \mathscr{F}^{\ell}$, then all of the conclusions of Theor. 2.6 *apply to the associated stochastic process $\{E_n\}$.* The quantity $P_s^{(n)}(A^{\ell})$ is defined in (2.11).

For an ergodic model the probability measure $P_s^{\infty}(A^{\ell}) = P^{\infty}(A^{\ell})$ defined in (2.13) does not depend on s. Therefore (2.12) implies the hypotheses of Lemma 6.2 with $\epsilon_n = L\alpha^n$, and the conclusions of Theorem 2.6 follow.

APPENDIX. MODELS WITH FINITE STATE SPACES

A. THEORY

The following definition is analogous to Def. 1.1.

DEFINITION. *A system (S, E, f, φ) is a finite state model if S and E are finite sets, $f.(\cdot)$ is a mapping of $E \times S$ into S, and $\varphi.(\cdot)$ is a mapping of $E \times S$ into the nonnegative real numbers such that $\sum_{e \in E} \varphi_e(s) = 1$.*

Definition 1.2 defines associated stochastic processes $\{S_n\}$ and $\{E_n\}$ for any such model. It is possible to develop, by the methods of Sects. 5 and 6, a theory of finite state models that completely parallels the theory of distance diminishing models surveyed in Sec. 2. This will not be done here, since the results concerning states obtained by these relatively complicated methods are, if anything, slightly inferior to those that can be obtained by applying the well known theory of finite Markov chains (see Feller, 1957; and Kemeny and Snell, 1960) to the process $\{S_n\}$. However, the results concerning events in the ergodic case are new and important. Therefore, a development will be presented that leads to the latter results as directly as possible. Applications to stimulus-sampling theory will be given in Subsec. B.

The natural analogue of H9 for finite state models is

H9′ *For any $s, s' \in S$, $T_n(s) \cap T_n(s')$ is not empty if n is sufficiently large.*

This is equivalent to

H9′ *The finite Markov chain $\{S_n\}$ has a single ergodic set, and this set is regular.*

The terminology for finite Markov chains used in this appendix follows Kemeny and Snell (1960). By analogy with Def. 2.2, a finite state model that satisfies H9′ will be called *ergodic*. The reader should note, however, that the associated process $\{S_n\}$ need not be an ergodic Markov chain, since it may have transient states. If there happen not to be any transient states, the chain is ergodic and regular.

Lemma 1 is analogous to Theor. 2.3.

LEMMA 1. *For any ergodic finite state model there are constants $C < \infty$ and $\alpha < 1$ and a probability distribution K^∞ on S such that*

$$| E.[\psi(S_n)] - E[\psi(S_\infty)]| \leqslant C\alpha^n \, | \psi | \tag{1}$$

for all real valued functions ψ on S and $n \geqslant 1$, where

$$E[\psi(S_\infty)] = \sum_{s \in S} \psi(s) \, K^\infty(\{s\}). \tag{2}$$

Proof. Let N be the number of states. To facilitate the use of matrix notation the states are denoted $s_1, s_2, \ldots, s_N$. The transition matrix P and the column vector ψ^* corresponding to ψ are then defined by

$$P_{ij} = K(s_i, \{s_j\}) \quad \text{and} \quad \psi_i^* = \psi(s_i). \tag{3}$$

Then

$$E_{s_i}[\psi(S_n)] = (P^{n-1}\psi^*)_i \tag{4}$$

for $n \geqslant 1$.

There is a stochastic matrix A, all of whose rows are the same, say $(a_1, ..., a_N)$, and there are constants $b < \infty$ and $\alpha < 1$ such that

$$|(P^{n-1})_{ij} - A_{ij}| \leqslant b\alpha^{n-1} \tag{5}$$

for all $n \geqslant 1$ and $1 \leqslant i, j \leqslant N$. When $\{S_n\}$ is regular, this assertion is Corollary 4.1.5 of Kemeny and Snell (1960). When $\{S_n\}$ has transient states, that corollary can be supplemented by Kemeny and Snell's Corollary 3.1.2 and a straightforward additional argument to obtain (5). Let K^∞ be the probability measure on S with $K^\infty(\{s_j\}) = a_j$, and let $E[\psi(S_\infty)]$ be any coordinate of $A\psi^*$. Then (2) holds and

$$\begin{aligned} | E_{s_i}[\psi(S_n)] - E[\psi(S_\infty)]| &= |(P^{n-1}\psi^*)_i - (A\psi^*)_i| \\ &= \left| \sum_{j=1}^{N} [(P^{n-1})_{ij} - A_{ij}]\, \psi_j^* \right| \\ &\leqslant \sum_{j=1}^{N} |(P^{n-1})_{ij} - A_{ij}|\, |\psi(s_j)| \\ &\leqslant Nb\alpha^{n-1} |\psi|. \end{aligned}$$

This gives (1) with $C = Nb/\alpha$. Q.E.D.

The next lemma parallels Theorem 2.4.

LEMMA 2. *For any ergodic finite state model*

$$| P_\cdot^{(n)}(A^\ell) - P^\infty(A^\ell)| \leqslant C\alpha^n \tag{6}$$

for all $n, \ell \geqslant 1$ *and* $A^\ell \subset E^\ell$, *where*

$$P^\infty(A^\ell) = \sum_{s \in S} P_s^{(1)}(A^\ell)\, K^\infty(\{s\}), \tag{7}$$

and C *and* α *are as in Lemma* 1.

Proof. Just as in the proof of Theorem 2.4, (6.1) and (6.2) hold. Thus (6) follows from Lemma 1. Q.E.D.

Theorem 1 is the main result of this subsection.

Theorem 1. *All of the conclusions of Theorem* 2.6 *hold for any ergodic finite state model.*

Proof. A finite state model can be regarded as a homogeneous random system with complete connections, in the same sense that a distance diminishing model can be so regarded (see the first paragraph of Subsec. 6C—if $\mathscr{B}$ and $\mathscr{F}$ are taken to be the collections of all subsets of S and E, respectively, the measurability conditions in the definition of such a system evaporate). Thus Lemma 6.2 is applicable, and Theorem 1 follows from Lemma 2. Q.E.D.

B. Application to Stimulus-Sampling Theory

Consider the general two-choice situation described in the first paragraph of Subsec. 3A. The state S_n at the beginning of trial n for the N element component model with fixed sample size ν (Estes, 1959) can be taken to be the number of stimulus elements conditioned to response A_1 at the beginning of the trial. Thus

$$S = \{0, 1,..., N\}, \tag{8}$$

a finite set. The event space E can be taken to be

$$E = \{(i, j, k, \ell): 0 \leqslant i \leqslant \nu,\ 1 \leqslant j, k \leqslant 2, 0 \leqslant \ell \leqslant 1\} \tag{9}$$

where i is the number of stimulus elements in the trial sample conditioned to A_2, A_j is the response and O_{jk} the trial outcome, and $\ell = 1$ or 0 depending on whether or not conditioning is effective. The corresponding event operators can be written

$$f_{ij11}(s) = s + \min(i, N - s), \tag{10}$$

$$f_{ij21}(s) = s - \min(\nu - i, s), \tag{11}$$

and

$$f_{ijk0}(s) = s. \tag{12}$$

Of course, i elements conditioned to A_2 cannot be drawn if $i > N - s$, so the definition of $f_{ij11}(s)$ is irrelevant in this case. The definition given in (10) makes $f_{ij11}(\cdot)$ monotonic throughout S. The same holds for (11). The operators given by (10) and (11) are, incidentally, analogous in form to the linear operators

$$f(p) = p + \Theta(1 - p) \quad \text{and} \quad g(p) = p - \Theta p,$$

if taking the minimum of two numbers is regarded as analogous to multiplying them. Finally, the corresponding operator application probabilities are

$$\varphi_{ijk\ell}(s) = \frac{\binom{N - s}{i}\binom{s}{\nu - i}}{\binom{N}{\nu}} \left(\frac{\nu - i}{\nu}\right)_j \pi_{jk}[c_{jk}]_\ell\,, \tag{13}$$

where $(p)_i = \delta_{i1}p + \delta_{i2}(1-p)$, $[p]_\ell = \delta_{\ell 1}p + \delta_{\ell 0}(1-p)$, c_{jk} is to be interpreted as the probability that conditioning is effective if outcome O_{jk} occurs, and $\binom{J}{m}$ is 0 unless $0 \leqslant m \leqslant J$.

For any choice of $N \geqslant \nu \geqslant 1$ and $0 \leqslant \pi_{ij}, c_{ij} \leqslant 1$, (8–13) define a finite state model that will be referred to below as a *fixed sample size model.*

THEOREM 2. *A fixed sample size model with π_{ij} and c_{ij} positive for all $1 \leqslant i, j \leqslant 2$ is ergodic.*

Proof. It is clear that if $S_n < N$ then the sample on trial n will contain elements conditioned to A_2 with positive probability. Given such a sample, $A_{2,n}$ will occur and be followed by $O_{21,n}$ with positive probability, and conditioning will be effective with positive probability. Thus $S_{n+1} > S_n$ with positive probability, and it follows that the state N can be reached from any state. Thus there is only one ergodic set, and it contains N. Furthermore, if $S_n = N$ then $A_{1,n}$ occurs with probability 1, and $O_{11,n}$ follows with positive probability. So $S_{n+1} = N$ with positive probability, and the ergodic set is regular. Q.E.D.

It follows from Theorems 1 and 2 that the conclusions of Theorem 2.6 are available for any fixed sample size model with $0 < \pi_{ij}, c_{ij}$ for all i and j. Letting D be the subset

$$D = \{(i, j, k, \ell): j = 1\}$$

of E, and $A_n = h(E_n)$, where h is the indicator function of D, the conclusions of Theorem 2.6 include a law of large numbers and, possibly, a central limit theorem for the number $S_{m,n} = \sum_{j=m}^{m+n-1} A_j$ of A_1 responses in the n trial block starting on trial m. A simple expression for $\sigma^2 = \sigma_h{}^2$ can be readily calculated for the *pattern model* ($\nu = 1$) with equal c_{ij} under noncontingent reinforcement.

THEOREM 3. *A fixed sample size model with $\nu = 1$, $0 < \pi_{11} = \pi_{21} = \pi_1 < 1$, and $c_{ij} = c > 0$ is ergodic. The law of large numbers*

$$\lim_{n\to\infty} E_s[((1/n)\, S_{m,n} - \pi_1)^2] = 0$$

and central limit theorem

$$\lim_{n\to\infty} P\left(\frac{S_{m,n} - n\pi_1}{(n)^{1/2}\sigma} < x\right) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{x} \exp(-t^2/2)\, dt$$

hold, where

$$\sigma^2 = \pi_1(1-\pi_1)(1 + 2(1-c)/c).$$

Proof. That

$$E[h(E_\infty)] = \lim_{n\to\infty} P(A_{1,n}) = \pi_1$$

follows from (37) in Atkinson and Estes (1963). The value of σ^2 can be obtained from (2.17) and Atkinson and Estes' formula (41). Q.E.D.

The methods of this subsection are equally applicable to the component model with fixed sampling probabilities (Estes, 1959).

REFERENCES

ATKINSON, R. C., AND ESTES, W. K. Stimulus sampling theory. In R. D. Luce, R. R. Bush, and E. Galanter (Eds.), *Handbook of mathematical psychology*, Vol. II. New York: Wiley, 1963. Pp. 121–268.

BREIMAN, L. A strong law of large numbers for a class of Markov chains. *Annals of Mathematical Statistics*, 1960, **31**, 801–803.

BUSH, R. R. Identification learning. In R. D. Luce, R. R. Bush, and E. Galanter (Eds.), *Handbook of mathematical psychology*, Vol. III. New York: Wiley, 1965. Pp. 161–203.

BUSH, R. R., AND MOSTELLER, F. *Stochastic models for learning*. New York: Wiley, 1955.

DOEBLIN, W., AND FORTET, R. Sur des chaînes à liaisons complètes. *Bulletin de la Société Mathématique de France*, 1937, **65**, 132–148.

ESTES, W. K. Component and pattern models with Markovian interpretations. In R. R. Bush and W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univ. Press, 1959. Pp. 9–52.

ESTES, W. K., AND SUPPES, P. Foundations of linear models. In R. R. Bush and W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford: Stanford Univer. Press, 1959. Pp. 137–179.

FELLER, W. *An introduction to probability theory and its applications*, Vol. I, (2nd ed.). New York: Wiley, 1957.

FELLER, W. *An introduction to probability theory and its applications*, Vol. II. New York: Wiley, 1966.

IONESCU TULCEA, C. On a class of operators occurring in the theory of chains of infinite order. *Canadian Journal of Mathematics*, 1959, **11**, 112–121.

IONESCU TULCEA, C., AND MARINESCU, G. Théorie ergodique pour des classes d'opérations non complètement continues. *Annals of Mathematics*, 1950, **52**, 140–147.

IOSIFESCU, M. Random systems with complete connections with an arbitrary set of states. *Revue de Mathématiques Pures et Appliquées, Académie de la République Populaire Roumaine*, 1963, **8**, 611–645. (A translation of this paper from the Russian can be obtained for $7.50 from Addis Translations, 129 Pope Street, Menlo Park, California 94025.)

IOSIFESCU, M., AND THEODORESCU, R. On Bush-Mosteller stochastic models for learning. *Journal of Mathematical Psychology*, 1965, **2**, 196–203.

JAMISON, B. Asymptotic behavior of successive iterates of continuous functions under a Markov operator. *Journal of Mathematical Analysis and Applications*, 1964, **9**, 203–214.

JAMISON, B. Ergodic decompositions induced by certain Markov operators. *Transactions of the American Mathematical Society*, 1965, **117**, 451–468.

KARLIN, S. Some random walks arising in learning models. *Pacific Journal of Mathematics*, 1953, **3**, 725–756.

KEMENY, J. G., AND SNELL, J. L. *Finite Markov chains*. Princeton, New Jersey: Van Nostrand, 1960.

KENNEDY, M. A convergence theorem for a certain class of Markov processes. *Pacific Journal of Mathematics*, 1957, **7**, 1107–1124.

LAMPERTI, J., AND SUPPES, P. Chains of infinite order and their application to learning theory. *Pacific Journal of Mathematics*, 1959, **9**, 739–754.

LAMPERTI, J., AND SUPPES, P. A correction to "Chains of infinite order and their application to learning theory." *Pacific Journal of Mathematics*, 1965, **15**, 1471–1472.

LOVEJOY, E. Analysis of the overlearning reversal effect. *Psychological Review*, 1966, **73**, 87–103.

NORMAN, M. F. Incremental learning on random trials. *Journal of Mathematical Psychology*, 1964, **1**, 336–350.

ONICESCU, O., AND MIHOC, G. Sur les chaînes de variables statistiques. *Bulletin de la Société Mathématique de France*, 1935, **59**, 174–192.

ROSE, R. M. Models for experiments with two complementary reinforcing events. Unpublished doctoral dissertation, University of Pennsylvania, 1964.

WYCKOFF, L. B. The role of observing responses in discrimination learning, part I. *Psychological Review*, 1952, **59**, 431–442.

YELLOTT, J. I. Some effects of noncontingent success in human probability learning. Technical Report No. 89. Institute for Mathematical Studies in the Social Sciences, Stanford University, 1965.

RECEIVED: September 21, 1966