Reprinted from
THE BRITISH JOURNAL OF MATHEMATICAL AND STATISTICAL PSYCHOLOGY,
Vol. 21, Part 2, p. 141, November 1968

SLOW LEARNING

By M. FRANK NORMAN
University of Pennsylvania

Learning will be slow if the quantity that represents the state of learning changes by small steps or with small probability on each experimental trial. Approximations to the distribution of this quantity for both types of slow learning are obtained. These approximations are applicable to a wide variety of models.

1. INTRODUCTION

There are at least two ways that learning can be slow. An experimental trial may produce small changes in a subject's habits, or it may produce no change at all with large probability. In the former case we say that learning occurs *by small steps*; in the latter we say it occurs *with small probability*. For example, consider a two-choice experiment in which a subject's probability of making response A after n trials (i.e. on the $(n+1)$st trial) is determined by a real quantity X_n, A's ' habit strength ', $n \geqslant 0$. On trial $n+1$ any one of k events can occur. The jth of these has probability $\phi_j(X_n)$ and, if it occurs, then $X_{n+1} - X_n = \Delta X_n = U_j(X_n)$. Learning by small steps occurs if the increments have the form $U_j(x) = \theta V_j(x)$ and θ is positive but small. Learning occurs with small probability if one of the events, say the first, has no effect on habit strength ($U_1(x) \equiv 0$), and the event probabilities have the form $\phi_j(x) = c\gamma_j(x)$ ($j \geqslant 2$), and $\phi_1(x) = 1 - c \sum_{j \geqslant 2} \gamma_j(x)$, where c is positive but small.

We will consider the behaviour of the Markov process $\{X_n\}$ as parameters like θ and c become small. To see what we should expect, consider the above example of learning by small steps under the added assumption that none of the V_j or ϕ_j depend on x and $\text{var}(X_0) = 0$. Then

$$X_n - E[X_n] = \sum_{m=0}^{n-1} (\Delta X_m - E[\Delta X_m]),$$

so that, letting $Y_m = \Delta X_m / \theta$,

$$Z_n = (X_n - E[X_n])/\theta^{1/2} = (\theta n)^{1/2} \sum_{m=0}^{n-1} (Y_m - E[Y_m])/n^{1/2}.$$

Now the random variables Y_m are independent and identically distributed ($P(Y_m = V_j) = \phi_j$). Hence the central limit theorem implies that as $n \to \infty$ and $\theta \to 0$ in such a way that $n\theta \to t > 0$, the distribution of Z_n converges to the normal distribution with mean 0 and variance $t\sigma^2$, where $\sigma^2 = \text{var}(Y_m)$. Similarly, if $k = 2$, $U_2(x) \equiv 1$, $\gamma_2(x) \equiv 1$ and $P(X_0 = 0) = 1$ in the example of learning with small probability, then the ΔX_m are independent with $P(\Delta X_m = 1) = c$ and $P(\Delta X_m = 0) = 1 - c$. Thus the Poisson convergence theorem implies that as $n \to \infty$ and $c \to 0$ in such a way that $nc \to t$, the distribution of $X_n = \sum_{j=0}^{n-1} \Delta X_j$ converges to the Poisson distribution with mean t.

In this paper we will obtain extensions of the central limit and Poisson theorems sufficiently general to cover most of the instances of learning by small steps and with small probability that arise in learning models. We have made a point of considering multidimensional as well as unidimensional X_n since such processes, which arise, for instance, in models for choices among several alternatives or in multi-process models, are already important in learning theory and seem destined to become even more so. In all cases we will find the asymptotic distribution P_t of X_n (suitably normalized) as $n \to \infty$, $r(=\theta$ or $c) \to 0$ and $nr \to t$. In applications P_{nr} can then be considered an approximation to the distribution $F_n{}^r$ of X_n when the learning rate parameter r is small. Indeed, it follows from our results that, for any $T > 0$,

$$\max_{n:\, nr \leqslant T} d(F_n{}^r, P_{nr}) \to 0$$

as $r \to 0$, where $d(G, H)$ is a suitable measure of the distance between the distributions G and H. The restriction $nr \leqslant T$ on the natural time variable nr limits the approximation to the transient phase of a learning experiment.

Applications of the general theorems of Sections 2, 3 and 6 to some standard learning models will be given in Sections 4 and 7.

2. Small Steps

2.1. *Assumptions*

Our assumptions in this section are abstracted from the example of learning by small steps given above. Let J be a bounded set of positive real numbers having 0 as a limit point. For every $\theta \in J$, let $\{X_n{}^\theta\}_{n \geqslant 0}$ be a Markov process with stationary transition probabilities in a closed (but not necessarily bounded) real interval I. We assume (a) that the conditional distribution of $Y_n{}^\theta = \Delta X_n{}^\theta / \theta$ given $X_n{}^\theta = x$ does not depend on θ. Let

$$w(x) = E[Y_n{}^\theta | X_n{}^\theta = x], \quad z(x) = E[(Y_n{}^\theta)^2 | X_n{}^\theta = x],$$
$$s(x) = z(x) - w^2(x) \quad \text{and} \quad r(x) = E[|Y_n{}^\theta|^3 | X_n{}^\theta = x].$$

We also assume (b) that $w(x)$ has two bounded derivatives and $s(x)$ has one, and (c) that $r(x)$ is bounded.

Under these assumptions we will determine functions $f(t, x)$ and $g(t, x)$ such that, if the initial state of $\{X_n{}^\theta\}$ is x $(P(X_0{}^\theta = x) = 1)$, then $X_n{}^\theta$ is asymptotically normally distributed with mean $f(t, x)$ and variance $\theta g(t, x)$ as $n \to \infty$, $\theta \to 0$, and $n\theta \to t$ (Theorem 2.3). Since the variance of $X_n{}^\theta$ is small when θ is small (Corollary 2.2.2) the 'mean learning curve' $E[X_n{}^\theta]$ is especially significant. Our approximation $f(n\theta, x)$ to this quantity is relatively easy to compute. $f(\cdot, x)$ is the solution of the differential equation $f'(t) = w(f(t))$ satisfying the initial condition $f(0) = x$ (Lemma 2.2.3).

The picture we derive of a fundamental theoretical variable X_n^θ with a tightly clustered nearly normal distribution is reminiscent of the situation postulated by Hull (1943, p. 319).

2.2. *Two Lemmas*

Since $|w(x)| \leqslant z^{1/2}(x) \leqslant r^{1/3}(x)$, (c) implies that w, z and s are bounded. Let

$$a = \sup_{x\in I} s(x), \quad b = \sup_{x\in I}|w(x)|, \quad c = \sup_{x\in I}|w''(x)|,$$
$$d = \sup_{x\in I} w'(x) \quad \text{and} \quad g = \sup_{x\in I}|w'(x)|.$$

When x is the initial state of the process $\{X_n^\theta\}$, we sometimes denote the corresponding probabilities, expectations and variances by P_x, E_x and var_x, though, more often, especially in proofs, our notation will reflect neither x nor θ. Let

$$\mu_n = \mu_n(\theta, x) = E_x[X_n^\theta]$$

and

$$\omega_n = \omega_n(\theta, x) = \text{var}_x(X_n^\theta).$$

LEMMA 2.2.1. For any $x \in I$, $n \geqslant 0$ and $\theta \in J$ sufficiently small that $\theta(2d + \theta g^2) \geqslant -1$,

$$\omega_n \leqslant \theta a A(n, \theta, 2d + \theta g^2),$$

where

$$A(n, \theta, h) = \theta \sum_{j=0}^{n-1} (1 + \theta h)^j.$$

Proof.

$$\begin{aligned}\omega_{n+1} &= \text{var}(X_n + \Delta X_n)\\ &= \omega_n + \text{var}(\Delta X_n) + 2\text{cov}(X_n, \Delta X_n).\end{aligned} \tag{2.2.1}$$

Now

$$\begin{aligned}\text{cov}(X_n, \Delta X_n) &= E[E[(X_n - \mu_n)(\Delta X_n - \Delta\mu_n)|X_n]]\\ &= E[(X_n - \mu_n)E[\Delta X_n - \Delta\mu_n|X_n]],\end{aligned}$$

and

$$\begin{aligned}E[\Delta X_n - \Delta\mu_n|X_n] &= \theta(w(X_n) - E[w(X_n)])\\ &= \theta((w(X_n) - w(\mu_n)) + (w(\mu_n) - E[w(X_n)])),\end{aligned}$$

so that

$$\begin{aligned}\text{cov}(X_n, \Delta X_n) &= \theta E[(X_n - \mu_n)(w(X_n) - w(\mu_n))]\\ &= \theta E\left[(X_n - \mu_n)^2 \frac{(w(X_n) - w(\mu_n))}{X_n - \mu_n}\right]\\ &\leqslant \theta d E[(X_n - \mu_n)^2] = \theta d \omega_n\end{aligned} \tag{2.2.2}$$

by the mean value theorem. Also,

$$\begin{aligned}\operatorname{var}(\Delta X_n) &\leqslant E[(\Delta X_n - \theta w(\mu_n))^2] \\ &= \theta^2 E[E[(Y_n - w(\mu_n))^2 | X_n]],\end{aligned}$$

and

$$\begin{aligned}E[(Y_n - w(\mu_n))^2 | X_n] &= s(X_n) + (w(X_n) - w(\mu_n))^2, \\ &\leqslant a + g^2(X_n - \mu_n)^2,\end{aligned}$$

so

$$\operatorname{var}(\Delta X_n) \leqslant \theta^2 a + \theta^2 g^2 \omega_n. \tag{2.2.3}$$

Combining eqns. (2.2.1) to (2.2.3) we obtain

$$\omega_{n+1} \leqslant (1 + \theta(2d + \theta g^2))\omega_n + \theta^2 a. \tag{2.2.4}$$

The conclusion of the lemma is obtained by iterating eqn. (2.2.4) and noting that $\omega_0 = 0$. Q.E.D.

If $h < 0$,

$$A(n, \theta, h) \leqslant 1/|h|,$$

whereas, for any h,

$$\begin{aligned}A(n, \theta, h) &\leqslant A(n, \theta, |h|) \\ &\leqslant [(1 + \theta|h|)^n - 1]/|h| \\ &\leqslant [\exp(n\theta|h|) - 1]/|h|.\end{aligned}$$

Thus we have the following corollary.

COROLLARY 2.2.2. (i) $\omega_n = O(\theta)$ uniformly in $n\theta \leqslant L$ for any $L < \infty$. (ii) If $d < 0$, $\omega_n = O(\theta)$ uniformly in $n \geqslant 0$.

By (i) we mean that there is a $\theta_0 > 0$ (in this case sup J is suitable) such that for any L there is a $B_L < \infty$ such that $\omega_n \leqslant B_L \theta$ whenever $\theta \leqslant \theta_0$ and $n\theta \leqslant L$. Similarly (ii) means that there is a $\theta_0 > 0$ (e.g. $|d|/g^2$) and a $B < \infty$ such that $\omega_n \leqslant B\theta$ for all $\theta \leqslant \theta_0$ and $n \geqslant 0$.

LEMMA 2.2.3. For any $x \in I$ there is a unique differentiable function $f(t) = f(t, x)$ such that

$$f'(t) = w(f(t)) \tag{2.2.5}$$

and

$$f(0) = x. \tag{2.2.6}$$

(i) $\mu_n - f(n\theta) = O(\theta)$ uniformly in $n\theta \leqslant L$ for any L.

(ii) If $d < 0$ then $\mu_n - f(n\theta) = O(\theta)$ uniformly in $n \geqslant 0$.

Proof.

$$\begin{aligned}\mu_{n+1} &= E[X_n+\Delta X_n] = \mu_n + E[\Delta X_n] \\ &= \mu_n + \theta E[w(X_n)] \\ &= \mu_n + \theta w(\mu_n) + \theta E[w(x_n) - w(\mu_n)]. \end{aligned} \tag{2.2.7}$$

By Taylor's theorem

$$w(X_n) - w(\mu_n) = w'(\mu_n)(X_n - \mu_n) + \epsilon_n c(X_n - \mu_n)^2/2,$$

where here and in what follows ϵ_n denotes a quantity such that $|\epsilon_n| \leqslant 1$. Thus

$$E[w(X_n) - w(\mu_n)] = c\epsilon_n \omega_n/2.$$

Substituting this into eqn. (2.2.7) we obtain

$$\mu_{n+1} = \mu_n + \theta w(\mu_n) + \theta c \epsilon_n \omega_n/2. \tag{2.2.8}$$

Now if μ_n is to have an approximation of the form $f(n\theta)$, then $f(0) \doteq \mu_0 = x$ and, by Corollary 2.2.2 and eqn. (2.2.8), $f((n+1)\theta) \doteq f(n\theta) + \theta w(f(n\theta))$, or $f'(t) \doteq w(f(t))$. Assume for the time being that I is the entire real line. Then standard theorems on ordinary differential equations (Birkhoff & Rota, 1962, Theorem 1, p. 103; Theorem 6, pp. 112–113) ensure that there is one and only one solution to eqns. (2.2.5) and (2.2.6). We will now prove that $\mu_n - f(n\theta) = O(\theta)$.

Let $\nu_n = \nu_n(\theta, x) = f(n\theta, x)$ and $n\theta = t$. Then

$$\nu_{n+1} - \nu_n - \theta w(\nu_n) = \int_0^\theta (f'(t+\tau) - f'(t))d\tau,$$

and

$$\begin{aligned}|f'(t+\tau) - f'(t)| &\leqslant g|f(t+\tau) - f(t)| \\ &= g|f'(t+\tau^*)|\tau \leqslant gb\tau,\end{aligned}$$

where $0 < \tau^* < \tau$. Thus

$$|\nu_{n+1} - \nu_n - \theta w(\nu_n)| \leqslant gb\theta^2/2,$$

so that

$$\nu_{n+1} = \nu_n + \theta w(\nu_n) + \epsilon_n gb\theta^2/2. \tag{2.2.9}$$

Let $\Delta_n = \nu_n - \mu_n$. Subtracting eqn. (2.2.8) from eqn. (2.2.9) yields

$$\Delta_{n+1} = \Delta_n + \theta(w(\nu_n) - w(\mu_n)) + \theta^2 r_n, \tag{2.2.10}$$

where

$$|r_n| \leqslant (gb + c\omega_n/\theta)/2.$$

Thus

$$\Delta_{n+1} = \left[1 + \theta \frac{w(\nu_n) - w(\mu_n)}{\nu_n - \mu_n}\right] \Delta_n + \theta^2 r_n.$$

If $\theta g < 1$, the coefficient of Δ_n is positive, so that

$$|\Delta_{n+1}| \leqslant \left[1+\theta \frac{w(\nu_n)-w(\mu_n)}{\nu_n-\mu_n}\right]|\Delta_n|+\theta^2|r_n|$$
$$\leqslant (1+\theta d)|\Delta_n|+\theta^2|r_n|.$$

Iterating this equation, and noting that $\Delta_0 = 0$, we obtain

$$|\Delta_n| \leqslant \theta^2 \sum_{j=0}^{n-1}(1+\theta d)^j |r_{n-j}|$$
$$\leqslant \frac{\theta}{2}\left[gbA(n, \theta, d)+c\theta \sum_{j=0}^{n-1}(1+\theta d)^j \omega_{n-j}/\theta\right]. \tag{2.2.11}$$

The conclusions of the lemma then follow from eqn. (2.2.11) and Corollary 2.2.2.

It remains only to consider the case $I \neq (-\infty, \infty)$. Extend w to the entire line without increasing its supremum and in such a way that the Lipschitz condition $|w(x)-w(y)| \leqslant g|x-y|$ is satisfied for all real x and y. For example, if $B = \inf I > -\infty$ ($C = \sup I < \infty$) we can take $w(x) = w(B)$ ($w(x) = w(C)$) for $x \leqslant B$ ($x \geqslant C$). Eqns. (2.2.5) and (2.2.6) have a unique solution for the extended function w, and all of the work in the preceding paragraph is valid if d is replaced by g. Corollary 2.2.2(i) and eqn. (2.2.11) (with d replaced by g) imply that $\mu_n \to f(t)$ as $n\theta \to t$. But $\mu_n \in I$, so $f(t) \in I$ for all $t \geqslant 0$. Thus f is the unique solution of eqns. (2.2.5) and (2.2.6) for the unextended function w, and the work of the preceding paragraph is applicable as it stands. Q.E.D.

2.3. *The Main Theorem on Small Steps*

Let

$$g(t, x) = \int_0^t s(f(u, x)) \exp\left[2\int_u^t w'(f(v, x))dv\right] du. \tag{2.3.1}$$

THEOREM 2.3. For any $x \in I$, if x is the initial state of $\{X_n^\theta\}$ for all $\theta \in J$, then the distribution of $Z_n = (X_n^\theta - \mu_n(\theta, x))/\theta^{1/2}$ converges to the normal distribution $P_{t,x}$ with mean 0 and variance $g(t, x)$ as $n \to \infty$, $\theta \to 0$ and $n\theta \to t \geqslant 0$.

Proof. In order to keep the length of the proof within bounds and the main ideas in view, we will omit most details. The order of magnitude computations omitted are similar to those in the preceding section.

First we obtain a difference–differential equation for the characteristic function $h_n(\lambda)$ of Z_n. Observe that

$$h_{n+1}(\lambda) = E[\exp(i\lambda Z_{n+1})] = E[\exp(i\lambda Z_n)\exp(i\lambda \Delta Z_n)]$$
$$= E[\exp(i\lambda Z_n)E[\exp(i\lambda\Delta Z_n)|X_n]]. \tag{2.3.2}$$

For real y

$$e^{iy} = 1+iy-y^2/2+O(|y|^3),$$

thus

$$E[\exp(i\lambda\Delta Z_n)|X_n] = 1 + i\lambda E[\Delta Z_n|X_n] - \lambda^2 E[(\Delta Z_n)^2|X_n]/2 + |\lambda|^3 O(E[|\Delta Z_n|^3|X_n]). \quad (2.3.3)$$

Using Corollary 2.2.2 (i) and Lemma 2.2.3(i) repeatedly, we obtain

$$E[\Delta Z_n|X_n] = \theta w'(v_n)Z_n + O(\theta^{3/2}(1+Z_n^2)) \quad (2.3.4)$$

and

$$E[(\Delta Z_n)^2|X_n] = \theta s(v_n) + O(\theta^{3/2}(1+Z_n^2)), \quad (2.3.5)$$

where $v_n = f(n\theta)$. Here and subsequently all O's are uniform when $n\theta$ is bounded. Also

$$E[|\Delta Z_n|^3|X_n] = \theta^{3/2}E[|Y_n - E[Y_n]|^3|X_n] \leqslant \theta^{3/2}(E^{1/3}[|Y_n|^3|X_n] + |E[Y_n]|)^3$$

by Minkowski's inequality. But $E[Y_n] = E[w(X_n)]$ is in absolute value no greater than $\sup_x|w(x)|$, and $E[|Y_n|^3|X_n] = r(X_n)$ is bounded by assumption (c). (Here we use the full force of (c) for the first time.) Thus

$$E[|\Delta Z_n|^3|X_n] = O(\theta^{3/2}). \quad (2.3.6)$$

Substituting eqns. (2.3.4) to (2.3.6) into eqn. (2.3.3), substituting the resulting expression for the conditional characteristic function of Z_n given X_n into eqn. (2.3.2), and using Corollary 2.2.2(i), we obtain the desired equation

$$h_{n+1}(\lambda) - h_n(\lambda) = \theta[w'(v_n)\lambda h_n'(\lambda) - s(v_n)(\lambda^2/2)h_n(\lambda)] + O(\theta^{3/2}). \quad (2.3.7)$$

The error term is of the order of magnitude of $\theta^{3/2}$ uniformly over n and λ when $n\theta$ and λ are bounded.

The difference–differential equation (2.3.7) suggests the partial differential equation

$$\frac{\partial H}{\partial t}(t, \lambda) = w'(f(t))\lambda\frac{\partial H}{\partial \lambda}(t, \lambda) - s(f(t))\frac{\lambda^2}{2}H(t, \lambda). \quad (2.3.8)$$

We will show that the solution of this equation for which $H(0, \lambda) = h_0(\lambda) = 1$ is the characteristic function

$$H(t, \lambda) = \exp[-g(t)\lambda^2/2] \quad (2.3.9)$$

of the normal distribution with mean 0 and variance $g(t)$, and that $h_n(\lambda) \to H(t, \lambda)$ as $\theta \to 0$, $n \to \infty$ and $n\theta \to t$. The key to both objectives is the reduction of eqn. (2.3.8) to a family of ordinary differential equations. Let $B(t)$ be the solution of the differential equation

$$B'(t) = -w'(f(t))B(t) \quad (2.3.10)$$

for which $B(0) = 1$. That is,

$$B(t) = \exp\left[-\int_0^t w'(f(v))dv\right].$$

Eqn. (2.3.8) implies that

$$\frac{dH}{dt}(t, \xi B(t)) = -s(f(t))\frac{(\xi B(t))^2}{2} H(t, \xi B(t)), \tag{2.3.11}$$

so that, since $H(0, \xi) = 1$,

$$H(t, \xi B(t)) = \exp\left[-\frac{1}{2}\int_0^t (\xi B(u))^2 s(f(u))du\right],$$

for all t and ξ. Choosing $\xi = \lambda B^{-1}(t)$ we finally obtain eqn. (2.3.9). (For a general discussion of this technique, see Courant & Hilbert, 1962, pp. 62–66.)

For any real ξ, let $\xi_n = \xi B(n\theta)$, $H_n = H(n\theta, \xi_n)$, and $h_n = h_n(\xi_n)$. Eqn. (2.3.11) suggests that

$$H_{n+1} - H_n = -\theta s(\nu_n)\xi_n^2 H_n/2 + O(\theta^2), \tag{2.3.12}$$

and this is easily established via Taylor's theorem.

Also, writing

$$h_{n+1} - h_n = (h_{n+1}(\xi_{n+1}) - h_{n+1}(\xi_n)) + (h_{n+1}(\xi_n) - h_n(\xi_n)),$$

and applying eqn. (2.3.10) to the first and eqn. (2.3.7) to the second term on the right, we obtain the analogous equation

$$h_{n+1} - h_n = -\theta s(\nu_n)\xi_n^2 h_n/2 + O(\theta^{3/2}) \tag{2.3.13}$$

for h_n. Subtracting eqn. (2.3.13) from eqn. (2.3.12) we obtain

$$H_{n+1} - h_{n+1} = (1 - \theta s(\nu_n)\xi_n^2/2)(H_n - h_n) + O(\theta^{3/2}),$$

which implies that $H_n - h_n = O(\theta^{1/2})$. But $H_n \to H(t, \xi B(t))$ and $h_n - h_n(\xi B(t)) \to 0$ as $n\theta \to t$, hence $h_n(\xi B(t)) \to H(t, \xi B(t))$ as $n \to \infty$, $\theta \to 0$ and $n\theta \to t$, for all ξ and t. A change of variables $\xi = \lambda B^{-1}(t)$ and an application of the continuity theorem complete the proof. Q.E.D.

Another method of solving the differential equation (2.3.8) sheds light on eqn. (2.3.1). Assuming that eqn. (2.3.9) is valid for some differentiable function g, and substituting into eqn. (2.3.8), we obtain the ordinary differential equation

$$g'(t) = 2w'(f(t))g(t) + s(f(t)), \tag{2.3.14}$$

to which eqn. (2.3.1) gives the solution.

Let $\psi(t, \cdot) = \psi(t, \cdot, x)$ be the normal density with mean 0 and variance $g(t)$. The partial differential equation (2.3.8) can be obtained from the equation

$$\frac{\partial \psi}{\partial t}(t,y) = \tfrac{1}{2}s(f(t))\frac{\partial^2 \psi}{\partial y^2}(t, y) - w'(f(t))\frac{\partial}{\partial y}(y\psi(t, y))$$

by Fourier transformation and integration by parts. This is the Kolmogorov forward or Fokker–Planck equation (Rosenblatt, 1962, p. 137) for a continuous time Markov process $Z(t)$ that may be regarded as an approximation to the process $(X_n^\theta - \mu_n)/\theta^{1/2}$ (t corresponds to $n\theta$) when θ is small. The nature of this and comparable approximations for the processes discussed in Sections 3 and 6, will be considered in a subsequent paper.

3. Generalizations

3.1. *Generalizations of Theorem* 2.3

As is shown in Section 4, Theorem 2.3 covers the linear model for two-choice experiments and the beta model for such experiments when it is formulated in a certain way. Various slight modifications of the hypotheses of Theorem 2.3 yield a new theorem with a much broader range of applicability. The modifications encompass (i) multidimensional X_n^θ, (ii) weak dependence on θ of the distribution of $\Delta X_n^\theta/\theta$ given X_n^θ, and (iii) state spaces I_θ of $\{X_n^\theta\}$ that vary slightly with θ. Both (ii) and (iii) arise, for instance, in M element fixed sample size stimulus-sampling models for two-choice experiments, where $\theta = 1/M$ and X_n^θ is the proportion of stimulus elements conditioned to one of the response alternatives after n trials.

Let J be a bounded set of positive real numbers of which 0 is a limit point. Let N be a positive integer. R_N is the set of N-tuples of real numbers, which we regard as column vectors. For every $\theta \in J$, let $\{X_n^\theta\}_{n\geq 0}$ be a Markov process with stationary transition probabilities in a subset I_θ of R_N. Let I be the smallest closed convex set including all I_θ $(\theta \in J)$. Let

$$Y_n^\theta = \Delta X_n^\theta/\theta, \quad w(x, \theta) = E[Y_n^\theta | X_n^\theta = x],$$
$$s(x, \theta) = E[(Y_n^\theta - w(x, \theta))(Y_n^\theta - w(x, \theta))^T | X_n^\theta = x]$$

and

$$r(x, \theta) = E[|Y_n^\theta|^3 | X_n^\theta = x],$$

where T indicates transposition, and $|y|^2 = y^T y$ for $y \in R_N$. Thus $w(x, \theta)$ is the conditional mean vector and $s(x, \theta)$ the conditional covariance matrix of the normalized increment Y_n^θ, given $X_n^\theta = x$. For any $N \times N$ matrix A, let

$$|A|^2 = \sum_{i,j=1}^{N} a_{i,j}^2.$$

We assume that I_θ approximates I as $\theta \to 0$, in the sense that, (a1) for any $x \in I$,

$$\lim_{\theta \to 0} \inf_{y \in I_\theta} |x - y| = 0.$$

Next we suppose that there are functions $w(x)$ and $s(x)$ on I that approximate $w(x, \theta)$ and $s(x, \theta)$ when θ is small, by which we mean that (a2)

$$\sup_{x \in I_\theta} |w(x, \theta) - w(x)| = O(\theta) \quad \text{and} \quad \sup_{x \in I_\theta} |s(x, \theta) - s(x)| \to 0$$

as $\theta \to 0$. The function w is assumed differentiable in the sense that (b1) there is an $N \times N$ matrix valued function $w'(x)$ on I such that

$$\lim_{\substack{y \to x \\ y \in I}} \frac{|w(y) - w(x) - w'(x)(y - x)|}{|y - x|} = 0,$$

for all $x \in I$. We assume that $w'(x)$ is bounded: (b2)

$$\sup_{x \in I} |w'(x)| < \infty,$$

and that $w'(x)$ and $s(x)$ satisfy Lipschitz conditions (b3)

$$\sup_{\substack{x, y \in I \\ x \neq y}} \frac{|w'(x) - w'(y)|}{|x-y|} < \infty \quad \text{and} \quad \sup_{\substack{x, y \in I \\ x \neq y}} \frac{|s(x) - s(y)|}{|x-y|} < \infty.$$

Finally, we suppose that $r(x, \theta)$ is bounded: (c)

$$\sup_{\theta \in J,\, x \in I_\theta} r(x, \theta) < \infty.$$

For $\theta \in J$ and $x \in I_\theta$, let $\mu_n(\theta, x) = E_x[X_n^\theta]$, and $\omega_n(\theta, x) = E_x[|X_n^\theta - \mu_n(\theta, x)|^2]$. Under the above conditions we have the following theorem.

THEOREM 3.1. (i) $\omega_n(\theta, x) = O(\theta)$ uniformly in $x \in I_\theta$ and $n\theta \leqslant L$, for any $L < \infty$.

(ii) For any $x \in I$, the differential equation

$$f'(t) = w(f(t))$$

has a unique solution $f(t) = f(t, x)$ for which $f(0) = x$. $f(t) \in I$ for all $t \geqslant 0$.

$$\mu_n(\theta, x) - f(n\theta, x) = O(\theta)$$

uniformly in $x \in I_\theta$ and $n\theta \leqslant L$.

(iii) For any $x \in I$ the matrix differential equation

$$B'(t) = -w'(f(t, x))^T B(t)$$

has a unique solution $B(t) = B(t, x)$ for which $B(0)$ is the identity matrix. $B(t)$ is non-singular for all $t \geqslant 0$. For every $\theta \in J$ let $\{X_n^\theta\}$ have initial state $x_\theta \in I_\theta$. If $x \in I$, and $x_\theta \to x$ as $\theta \to 0$, then the distribution of $(X_n^\theta - \mu_n(\theta, x_\theta))/\theta^{1/2}$ converges to the normal distribution with mean 0 and covariance matrix

$$g(t, x) = \int_0^t [B(u, x)B^{-1}(t, x)]^T s(f(u, x))[B(u, x)B^{-1}(t, x)]du$$

as $n \to \infty$, $\theta \to 0$ and $n\theta \to t$.

In view of (ii), we can, of course, substitute $f(n\theta, x_\theta)$ for $\mu_n(\theta, x_\theta)$ in (iii). It can be shown that $f(t, x) - f(t, x') = O(|x - x'|)$ uniformly in $t \leqslant L$ for any $L < \infty$, so $f(n\theta, x)$ can also be so substituted if $|x_\theta - x| = o(\theta^{1/2})$.

A proof of Theorem 3.1 can be obtained by straightforward but rather tedious modification of that of Theorem 2.3. We shall not give it here. The additional complications in the proof of the more general theorem are due almost entirely to the more complicated dependence of $\{X_n^\theta\}$ on θ, rather than to multidimensionality.

3.2. *Generalization of Corollary* 2.2.2(ii) *and Lemma* 2.2.3(ii)

We begin by noting some implications of negativity of $d=\sup_{x\in I} w'(x)$, within the framework of the assumptions of Section 2.1. The function $w(x)$ is bounded, so $I=[a, b]$ must be bounded, and w has at most one zero in this interval. If $w(a)<0$ then

$$\mu_1(\theta, a)=a+\theta w(a)<a,$$

an impossibility. Thus $w(a)\geqslant 0$. Similarly $w(b)\leqslant 0$. Thus w has a unique zero ρ, and $w'(\rho)<0$. These properties of w and boundedness of I are sufficient to obtain $\omega_n(\theta, x)=O(\theta)$ and $\mu_n(\theta, x)-f(n\theta, x)=O(\theta)$ uniformly in n, even under more general assumptions about the dependence of $\{X_n{}^\theta\}$ on θ.

We now state our assumptions precisely. J, $\{X_n{}^\theta\}$, I_θ, I, $Y_n{}^\theta$ and $w(x, \theta)$ are as in the second paragraph of Section 3.1 with $N=1$. Let

$$z(x, \theta)=E[(Y_n{}^\theta)^2|X_n{}^\theta=x].$$

We assume that (a1) and the first condition of (a2) hold, that $w(x)$ has a bounded derivative that satisfies a Lipschitz condition, and that $z(x, \theta)$ is bounded. Finally, we assume that I is bounded, that $w(x)$ has a unique zero ρ, and that $w'(\rho)<0$. Under these conditions we have this result.

THEOREM 3.2. (i) $\text{var}_x(X_n{}^\theta)=O(\theta)$ uniformly in $x\in I_\theta$ and $n\geqslant 0$. (ii) For any $x\in I$, there is a unique differentiable function $f(t)=f(t, x)$ such that $f'(t)=w(f(t))$ and $f(0)=x$. $E_x[X_n{}^\theta]-f(n\theta, x)=O(\theta)$ uniformly over $x\in I_\theta$ and $n\geqslant 0$.

This theorem can be proved by suitably refining the arguments in the proof of Corollary 2.2.2(ii) and Lemma 2.2.3(ii). We omit the details.

Theorem 3.2(i) yields immediately

$$\limsup_{n\to\infty} \text{var}_x(X_n{}^\theta)=O(\theta).$$

By Lemma 5.1 below, $f(t, x)\to\rho$ as $t\to\infty$, so that (ii) implies

$$\limsup_{n\to\infty} |E_x[X_n{}^\theta]-\rho|=O(\theta).$$

The problem of the asymptotic distribution of $(X_n{}^\theta-\rho)/\theta^{1/2}$ as $\theta\to 0$ *after* $n\to\infty$, under hypotheses like those of Theorem 3.2, has been considered by Norman & Graham (1968) and Norman (1968). As one might suppose on the basis of Theorem 3.1 and Lemma 5.4 below, the limiting distribution is normal with mean 0 and variance $s(\rho)/2|w'(\rho)|$.

4. EXAMPLES

Consider a two-choice (A_1 or A_2) learning experiment in which either response can be reinforced on any trial, regardless of what response the subject makes. Let π_{ij} be the probability that A_j is reinforced if A_i is made, and let p_n denote a subject's A_1 response probability after n trials ($n=0, 1, \ldots$).

In the linear model (see Sternberg, 1963) $\Delta p_n = \theta_{i1}(1-p_n)$ if response A_i on trial $n+1$ is followed by reinforcement of A_1, while $\Delta p_n = -\theta_{i2} p_n$ if A_2 is reinforced after A_i. In the beta model (see Sternberg, 1963) $p_n = v_n/(1+v_n)$, where $v_{n+1} = \beta_{ij} v_n$ if A_i is made and A_j is reinforced ($\beta_{i1} \geqslant 1$, $\beta_{i2} \leqslant 1$). At the level of p_n, the corresponding transformation is

$$p_{n+1} = \beta_{ij} p_n / [(1-p_n) + \beta_{ij} p_n],$$

while at the level of $y_n = \ln v_n$ it is $\Delta y_n = \ln \beta_{ij}$. In the M element stimulus sampling model with fixed sample size ν (see Atkinson & Estes, 1963) p_n is the proportion of stimulus elements conditioned to A_1 after n trials. If the stimulus sample on trial $n+1$ contains k elements conditioned to A_1, then A_1 is made with probability k/ν. If conditioning is effective, all elements in the sample become conditioned to the reinforced response (those already so conditioned do not change), so $\Delta p_n = (\nu - k)/M$ if A_1 is reinforced, while $\Delta p_n = -k/M$ if A_2 is reinforced. If conditioning is ineffective, $\Delta p_n = 0$. Conditioning is effective with probability c_{ij} if response A_j is reinforced after response A_i is made.

The theorems of the previous sections are applicable to all of these models. In the linear model we introduce the parameter θ by letting $\theta_{ij} = \theta d_{ij}$ for constants $d_{ij} \geqslant 0$. Thus the vector

$$(\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$$

approaches 0 along a line in four-dimensional space as θ approaches 0. Taking $I = [0, 1]$ and $J = (0, 1/\max d_{ij})$, it is easy to verify that $X_n^\theta = p_n$ satisfies the hypotheses of Theorem 2.3. Here, as in all of the subsequent examples, the normalized increment $\Delta X_n^\theta/\theta$ is bounded, so the boundedness condition on its conditional third absolute moment is trivially satisfied. Clearly

$$w(p) = d_{11}\pi_{11} p(1-p) + d_{21}\pi_{21}(1-p)^2 - d_{12}\pi_{12} p^2 - d_{22}\pi_{22}(1-p)p.$$

If we take $\beta_{ij} = \exp(b_{ij}\theta)$ ($b_{i1} \geqslant 0$, $b_{i2} \leqslant 0$, $0 < \theta < \delta$), then $\ln \beta_{ij} = \theta b_{ij}$, and the process $X_n^\theta = y_n$ in the beta model satisfies the hypotheses of Theorem 2.3 with $I = (-\infty, \infty)$. Theorem 3.1 is applicable to p_n. For this process $I = I_\theta = [0, 1]$ and

$$w(p) = p(1-p)[a_1 p + a_2(1-p)],$$

where $a_i = \pi_{i1} b_{i1} + \pi_{i2} b_{i2}$. The function

$$E[(\Delta v_n)^2 | v_n = v]$$

of v is typically unbounded, so Theorem 3.1 does not apply to the process v_n, regardless of how the parameter θ is introduced.

The hypotheses of Theorem 3.1 are satisfied for $X_n^\theta = p_n$ in the stimulus sampling model if we take $\theta = 1/M$, $J = \{1/M : M \geqslant 1\}$, $I_{1/M} = \{0, 1/M, \ldots, 1\}$, and $I = [0, 1]$. In this case it can be shown that

$$w(p) = \pi_{11} c_{11} p(1-p)(\nu - 1) + \pi_{21} c_{21}(1-p)[p + \nu(1-p)] \\ - \pi_{12} c_{12} p[(1-p) + \nu p] - \pi_{22} c_{22}(1-p)p(\nu - 1).$$

In the linear model, positivity of π_{ij} and d_{ij} for $i \neq j$ is necessary and sufficient for $w(0) > 0$ and $w(1) < 0$. Since w is quadratic, it then has a unique zero ρ in [0, 1], and $w'(\rho) < 0$, so Theorem 3.2 is applicable. The same observation, with d_{ij} replaced by c_{ij}, is pertinent to the stimulus sampling model. The process p_n for the beta model never satisfies the hypotheses of Theorem 3.2 since $w(p)$ always has at least two zeros in [0, 1].

As an example of a model for which X_n^θ is multidimensional we consider Wyckoff's discrimination learning model (see Bush, 1965, pp. 172–175). This model is partially specified by Bush's Table 1. To complete the specification we replace all plus and minus signs in the table by functions of the form $\theta\zeta(v)$, where v is the variable being transformed, $\zeta(v) \geqslant 0$ in the case of a plus and $\zeta(v) \leqslant 0$ in case of a minus. The functions ζ corresponding to different plus and minus signs may differ. All ζ's must satisfy $0 \leqslant v + \theta\zeta(v) \leqslant 1$ for $0 \leqslant v \leqslant 1$ and $0 < \theta < \delta$ in order to keep v in the unit interval, and all should have two continuous derivatives. Then Theorem 3.1 applies with $N = 4$ if we take $X_n = (x_n, y_n, z_n, u_n)^T$, $J = (0, \delta)$, and $I = I_\theta$ the closed unit cube in R_4.

5. The Function f

In order to facilitate use of the approximation $f(n\theta)$ to $E[X_n^\theta]$ we collect here a number of observations concerning the function f of Theorem 3.1(ii) in the one-dimensional case.

Lemma 5.1. Let $x \in I$, and $f(t) = f(t, x)$.

(i) If $w(x) = 0$, $f(t) = x$ for all $t \geqslant 0$.

(ii) If $w(x) > 0$, $f'(t) > 0$ for all $t \geqslant 0$. If w has no zeros above x, $\lim_{t\to\infty} f(t) = \infty$. If y is the smallest zero of w above x, $\lim_{t\to\infty} f(t) = y$.

(iii) If $w(x) < 0$, $f'(t) < 0$ for all $t \leqslant 0$. If w has no zeros below x, $\lim_{t\to\infty} f(t) = -\infty$. If y is the largest zero of w below x, $\lim_{t\to\infty} f(t) = y$.

(iv) If $w(x) \neq 0$, $f = F^{-1}$ where

$$F(z) = \int_x^z \frac{1}{w(u)}\, du. \tag{5.1}$$

Proof. For any real z we denote the function on $[0, \infty)$ that is identically equal to z by z^*. If $w(x) = 0$, then x^* solves $x^{*\prime}(t) = w(x^*(t))$ and $x^*(0) = x$. Since $f(t)$ is the only function with these properties, $f(t) = x^*(t) = x$ for all $t \geqslant 0$.

Suppose now that $f'(0) = w(x) > 0$. If $f'(t) = 0$ for some t, let $\tau > 0$ be the smallest such t. Then $w(f(\tau)) = 0$, so $f(\tau)^*$ solves eqn. (2.2.5). Since this solution agrees with $f(t)$ at $t = \tau$, $f(t) = f(\tau)$ for all $t \geqslant 0$ (Birkhoff & Rota, 1962, Theorem 5, p. 20). In particular, $x = f(\tau)$. However, $f'(t) > 0$ for $0 \leqslant t < \tau$, so it is clear that $x < f(\tau)$. This contradiction shows that $f'(t) > 0$ for all $t \geqslant 0$.

It follows that $w(f(t))>0$ for all $t\geqslant 0$. Since this function is continuous, $1/w(f(t))$ is also. Rewriting eqn. (2.2.5) in the form

$$f'(\tau)/w(f(\tau))=1,$$

and integrating both sides between 0 and t we obtain

$$\int_0^t \frac{f'(\tau)}{w(f(\tau))}\,d\tau=t.$$

Under the change of variables $u=f(\tau)$ this becomes

$$\int_x^{f(t)} \frac{1}{w(u)}\,du=t,$$

which is (iv).

If w has no zeros above x, then $w(u)>0$ for all $u\geqslant x$. If $f(t)\leqslant B<\infty$ for all $t\geqslant 0$, then $\infty>F(B)\geqslant F(f(t))=t$, an absurdity. Thus if w has no zeros above x, $f(t)\to\infty$ as $t\to\infty$.

If w has a zero above x, let y be the smallest. If $f(t)=y$ for some t, then $f'(t)=0$. So $f(t)<y$ for all t. Now $w(u)>0$ for all $x\leqslant u<y$. If $f(t)\leqslant y-\delta$ for some $\delta>0$ and all $t\geqslant 0$, then $\infty>F(y-\delta)\geqslant t$ for all $t\geqslant 0$, which is impossible. Thus $f(t)\to y$ as $t\to\infty$.

The case $w(x)<0$ is similar to the case $w(x)>0$ treated above. Q.E.D.

Eqn. (5.1) reduces the determination of f to a quadrature and an inversion. In the three cases in the previous section where we have displayed a formula for w, this function is a polynomial. Whenever w is a polynomial, the quadrature can be reduced to integration of terms of the form $(u-a)^k$ by partial fractions expansion of $1/w(u)$. For instance, if w is quadratic with distinct (real) roots, say $w(u)=K(u-\lambda)(u-\gamma)$, this technique yields

$$f(t)=\frac{\gamma(x-\lambda)e^{K\lambda t}-\lambda(x-\gamma)e^{K\gamma t}}{(x-\lambda)e^{K\lambda t}-(x-\gamma)e^{K\gamma t}}.$$

This result covers the most difficult case that arises in the linear and stimulus sampling models discussed in Section 4. Almost all of the results concerning $E[p_n]$ for these models available in the literature apply only when $w(p,\theta)$ (and thus $w(p)$) is a linear function of p. Except for these cases, $f(n\theta)$ is the only analytical approximation to $E[p_n]$ currently available. A special case of this approximation was proposed by Bush & Mosteller (1955, eqn. 8.64, p. 183) and applied to experimental data with great success in their Sections 13.6 to 13.8. Their derivation suggests that the error will be small if $\operatorname{var}(p_n)$ is small, but they give no conditions for small $\operatorname{var}(p_n)$.

It is not always possible to express f in terms of elementary functions, and even when it is the expression need not be illuminating. Some information about f can be obtained directly from eqn. (2.2.5). For example, differentiating both sides of this equation we obtain

$$f''(t)=w'(f(t))f'(t).$$

If $w(x)\neq 0$, then $f'(t)\neq 0$, so $f''(t)=0$ if and only if $w'(f(t))=0$.

Lemma 5.2 gives simple asymptotic estimates for the deviation of $f(t)$ from its asymptote. Part (i) can be obtained directly from eqn. (2.2.5), while (ii) is a corollary of Lemma 5.1(iv). The proofs are omitted.

LEMMA 5.2. Suppose that $w(x)\neq 0$, and $f(t)\to y\in I$ as $t\to\infty$. Then $w(y)=0$ and $w'(y)\leqslant 0$.

(i) If $w'(y)<0$, then

$$G(x)=\lim_{z\to y}\int_x^z\left[\frac{w(u)-w(y)}{u-y}-w'(y)\right]\frac{du}{w(u)}$$

exists if z approaches y from the direction of x, and

$$f(t)-y\sim(x-y)\exp[G(x)+w'(y)t]$$

as $t\to\infty$.

(ii) If $|w(u)|\sim|u-y|^{1+\delta}\Lambda$ for $\delta>0$ as u approaches y from the direction of x, then

$$|f(t)-y|\sim(\delta\Lambda t)^{-1/\delta}$$

as $t\to\infty$.

The final results in this section concern $g(t)$.

LEMMA 5.3. (i) If $w(x)=0$, then

$$g(t)=s(x)(e^{2w'(x)t}-1)/2w'(x).$$

(ii) If $w(x)\neq 0$, $g(t)=\Gamma(f(t))$, where

$$\Gamma(z)=w^2(z)\int_x^z\frac{s(u)}{w^3(u)}\,du.$$

Part (i) (read $g(t)=s(x)t$ if $w'(x)=0$) is obtained from eqn. (2.3.1) and Lemma 5.1(i) by straightforward computation. Part (ii) can be established by introducing f as a new variable of integration in both of the integrals in eqn. (2.3.1).

LEMMA 5.4. If $f(t)\to y\in I$ as $t\to\infty$, and if $w'(y)<0$, then

$$\lim_{t\to\infty}g(t)=s(y)/2|w'(y)|.$$

We will not prove this here. However, we note that the form of the limit is easily obtained by assuming that the limit exists and that $g'(t)\to 0$, and letting $t\to\infty$ in eqn. (2.3.14).

6. SMALL PROBABILITY

Let $\{X_n\}_{n\geqslant 0}$ be a Markov process with stationary transition probabilities

$$K(x, A)=P(X_{n+1}\in A|X_n=x)$$

in an abstract space I. The higher transistion probabilities

$$K^{(j)}(x, A)=P(X_{n+j}\in A|X_n=x) \tag{6.1}$$

are given recursively by

$$K^{(j+1)}(x, A)=\int_I K(x, dy)K^{(j)}(y, A). \tag{6.2}$$

From eqn. (6.1) we see that $K^{(0)}(x, A)=\delta_x(A)$, where

$$\delta_x(A)=\begin{cases}1 & \text{if } x\in A\\ 0 & \text{if } x\in' A,\end{cases}$$

and $K^{(1)}(x, A)=K(x, A)$. For $0<c\leqslant 1$, let $\{X_n^c\}$ be any Markov process with transition probabilities

$$K_c(x, A)=cK(x, A)+(1-c)\delta_x(A). \tag{6.3}$$

In the next section we will show that the parameter c discussed in Section 1 is of this type.

In order to see what we should expect of

$$P(X_n^c\in A|X_0^c=x)=K_c^{(n)}(x, A)$$

when n is large and c is small, we consider an example of a Markov process with transition probabilities (6.3). If $\{W_j\}_{j\geqslant 1}$ is a sequence of random variables with $P(W_j=1)=c$ and $P(W_j=0)=1-c$ that are independent of each other and of $\{X_n\}$, and if $S(n)=\sum_{j=1}^{n} W_j$, then $X_{S(n)}$ is such a process. Let $\{N(t)\}_{t\geqslant 0}$ be a Poisson process with mean inter-event time 1 that is independent of $\{X_n\}$. As $n\to\infty$, $c\to 0$ and $nc\to t$, the distribution of $S(n)$ converges to that of $N(t)$. Hence we expect the distribution of $X_{S(n)}$ to converge to that of $X_{N(t)}$, i.e.

$$K_c^{(n)}(x, A)\to L_t(x, A), \tag{6.4}$$

where

$$L_t(x, A)=\sum_{j=0}^{\infty} K^{(j)}(x, A)e^{-t}t^j/j!.$$

The Markov process $\{X_{N(t)}\}_{t\geqslant 0}$ is of pseudo-Poisson type (Feller, 1966, pp. 312–314).

Theorem 6.1 is somewhat stronger than expression (6.4). For any real y, $[y]$ is the least integer greater than or equal to y.

THEOREM 6.1. For all n, $c>0$,

$$|K_c^{(n)}(x, A)-L_{nc}(x, A)|\leqslant \exp(c[nc]\zeta(nc))-1,$$

where $\zeta(y)=1-([y]-1)/2y$.

It is easy to show that $\frac{1}{2}<\zeta(y)\leqslant 1$ for all $y>0$, $\zeta(y)=1$ for $0<y\leqslant 1$, and $\zeta(y)\to\frac{1}{2}$ as $y\to\infty$.

Proof. The formula

$$K_c^{(n)}(x, A) = \sum_{j=0}^{n} K^{(j)}(x, A) b_j, \tag{6.5}$$

where

$$b_j = \binom{n}{j} c^j (1-c)^{n-j}$$

follows easily from eqn. (6.3) by induction. Let $b_j = 0$ if $j > n$, and let

$$p_j = e^{-nc}(nc)^j / j!.$$

Clearly

$$K_c^{(n)}(x, A) - L_{nc}(x, A) = \sum_{j=0}^{\infty} K^{(j)}(x, A)(b_j - p_j)$$
$$\leqslant \sum_{j: b_j > p_j} (b_j - p_j) = D$$

and

$$L_{nc}(x, A) - K_c^{(n)}(x, A) \leqslant \sum_{j: b_j \leqslant p_j} (p_j - b_j) = E.$$

But

$$0 = \sum_{j=0}^{\infty} (b_j - p_j) = D - E,$$

so $D = E$, and

$$|K_c^{(n)}(x, A) - L_{nc}(x, A)| \leqslant D. \tag{6.6}$$

Let $a_j = b_j / p_j$. Then

$$D = \sum_{j: a_j > 1} p_j(a_j - 1)$$
$$\leqslant \max_{j: a_j > 1} (a_j - 1) \sum_{j: a_j > 1} p_j$$
$$\leqslant \max_{0 \leqslant j \leqslant n} a_j - 1. \tag{6.7}$$

A routine computation yields

$$a_j = \frac{(n)_j (1-c)^{n-j}}{n^j e^{-nc}},$$

where $(n)_j = n \cdot (n-1) \ldots (n-j+1)$, so

$$\frac{a_{j+1}}{a_j} = \frac{n-j}{n-nc}$$

for $j \leqslant n$. Thus $a_{j+1} > a_j$ if $j < nc$, $a_{j+1} = a_j$ if $j = nc$ and $a_{j+1} < a_j$ if $j > nc$. It follows that

$$\max_{0 \leqslant j \leqslant n} a_j = a_{[nc]}. \tag{6.8}$$

Writing

$$a_j = (1 - 1/n) \ldots (1 - (j-1)/n)(1-c)^{n-j} e^{nc},$$

and applying the inequality $e^{-x} \geqslant (1-x)$ to all terms but the last, we obtain

$$a_j \leqslant \exp[jc(1-(j-1)/2nc)]. \tag{6.9}$$

The proof is completed by taking $j=[nc]$ in eqn. (6.9) and combining the resulting inequality with eqns. (6.6) to (6.8). Q.E.D.

7. Examples

The parameter c can be introduced into a broad class of learning models as follows. Suppose that the variable X_n that describes the state of learning after n trials takes on values in an abstract space I. Any one of k events can occur on trial $n+1$. The jth of these effects the transformation

$$X_{n+1} = F_j(X_n) \tag{7.1}$$

in the state of learning, and occurs with probability $\phi_j(X_n)$ $(j=1,\ldots,k)$. Naturally $\sum_{j=1}^{k} \phi_j(x) \equiv 1$ and $F_j(x) \in I$ if $\phi_j(x) > 0$. We now modify this model to take account of the possibility that the effect ascribed to the jth event in eqn. (7.1) occurs only with probability $c_j \in [0, 1]$, while with probability $1-c_j$, $X_{n+1} = X_n$. Thus eqn. (7.1) is applicable with probability $c_j\phi_j(X_n) = \phi_j^*(X_n)$, and

$$X_{n+1} = F_0(X_n), \tag{7.2}$$

where F_0 is the identity operator, with probability

$$\phi_0^*(X_n) = \sum_{j=1}^{k} (1-c_j)\phi_j(X_n).$$

The possibility that some of the F_j $(j \geqslant 1)$ are identity operators is not excluded.

Suppose that $c = \max_j c_j \neq 0$, and let $d_j = c_j/c$. Then

$$c_j = cd_j \quad \text{and} \quad \max_j d_j = 1.$$

To study the behaviour of the model when the c_j are small, we fix the vector $(d_1, \ldots, d_k)$ and let $c \to 0$. Let $\{X_n^c\}$ be the sequence of states corresponding to the superscripted value of the parameter, let

$$K_c(x, A) = P(X_{n+1}^c \in A \mid X_n^c = x)$$

be the transition probabilities for this Markov process, let $X_n = X_n^1$, and let $K(x, A) = K_1(x, A)$. To justify this notation and show that the theory developed in Section 6 is applicable, we must show that K_c and K satisfy eqn. (6.3). Now

$$K_c(x, A) = \sum_{0 \leqslant j \leqslant k:\, F_j(x) \in A} \phi_j^*(x)$$

and

$$\phi_0^*(x) = c \sum_{j=1}^{k} (1-d_j)\phi_j(x) + (1-c),$$

so that

$$K_c(x, A) = c \sum_{1 \leq j \leq k:\, F_j(x) \in A} d_j \phi_j(x) + \left[c \sum_{j=1}^{k} (1 - d_j)\phi_j(x) + (1 - c) \right] \delta_x(A)$$
$$= cK(x, A) + (1 - c)\delta_x(A).$$

Thus, indeed, eqn. (6.3) holds. It follows that Theorem 6.1 and its corollary (6.4) apply to $\{X_n^c\}_{n=0}^{\infty}$.

The construction described above may be used to introduce the parameter c into the linear, beta and Wyckoff models of Section 4. Linear models involving such a parameter have been considered by Norman (1964) and Yellott (1965). The parameters c_j have already been introduced into the stimulus sampling model discussed in Section 4, and to complete the construction it remains only to write $c_j = cd_j$. The parameter c is a standard feature of stimulus sampling theory.

Acknowledgement

This research was supported by the National Science Foundation under grant GP–7335.

References

Atkinson, R. C. & Estes, W. K. (1963). Stimulus sampling theory. In R. D. Luce, R. R. Bush & E. Galanter (eds.), *Handbook of Mathematical Psychology*, vol. II, pp. 121–268. New York: Wiley.

Birkhoff, G. & Rota, G. (1962). *Ordinary Differential Equations*. New York: Ginn.

Bush, R. R. (1965). Identification learning. In R. D. Luce, R. R. Bush & E. Galanter (eds.), *Handbook of Mathematical Psychology*, vol. III, pp. 161–203. New York: Wiley.

Bush, R. R. & Mosteller, F. (1955). *Stochastic Models for Learning*. New York: Wiley.

Courant, R. & Hilbert, D. (1962). *Methods of Mathematical Physics*, vol. II. New York: Interscience.

Feller, W. (1966). *An Introduction to Probability Theory and Its Applications*, vol. II. New York: Wiley.

Hull, C. L. (1943). *Principles of Behavior*. New York: Appleton-Century-Crofts.

Norman, M. F. (1964). Incremental learning on random trials. *J. math. Psychol.* **1**, 336–350.

Norman, M. F. (1968). Mathematical learning theory. In G. B. Dantzig & A. F. Veinott (eds.), *Mathematics of the Decision Sciences*. Providence, R.I.: American Mathematical Society. (In press.)

Norman, M. F. & Graham, N. V. (1968). A central limit theorem for families of stochastic processes indexed by a small average step size parameter, and some applications to learning models. *Psychometrika* **33.** (In press.)

Rosenblatt, M. (1962). *Random Processes*. New York: Oxford University Press.

Sternberg, S. (1963). Stochastic learning theory. In R. D. Luce, R. R. Bush & E. Galanter (eds.), *Handbook of Mathematical Psychology*, vol. II, pp. 1–120. New York: Wiley.

Yellott, J. I., Jr. (1965). Some effects of noncontingent success in human probability learning. *Technical Report No.* 89. Stanford, Calif.: Institute for Mathematical Studies in the Social Sciences.