August 1977

Fixed Structure Automata in a Multi-Teacher Environment

Daniel E. Koditschek
University of Pennsylvania, kod@seas.upenn.edu

Kumpati S. Narendra
Yale University

NOTE: At the time of publication, author Daniel Koditschek was affiliated with Yale University. Currently, he is a faculty member in the Department of Electrical and Systems Engineering at the University of Pennsylvania.

Fixed Structure Automata in a Multi-Teacher Environment

Abstract

The concept of an automaton operating in a multi-teacher environment is introduced, and several interesting questions that arise in this context are examined. In particular, we concentrate on the consequences of adding a new teacher to an existing n-teacher set as it affects the choice of a switching strategy. The effect of this choice on expediency and speed of convergence is presented for a specific automaton structure.

Comments

NOTE: At the time of publication, author Daniel Koditschek was affiliated with Yale University. Currently, he is a faculty member in the Department of Electrical and Systems Engineering at the University of Pennsylvania.

shown stable behavior, and it proved easy to apply. Although the test problems have been of modest size ($d \leq 6$), the model should be applicable to considerably larger problems ($d \approx 40$).

REFERENCES

[1] M. R. Anderberg, "Cluster analysis for applications." New York and London: Academic, 1973.

[2] M. M. Astrahan, "Speech analysis by clustering, or the hyperphoneme method," Stanford Artificial Intelligence Proj. Mem. AIM-124, AD 709067. Stanford Univ., Stanford, CA, 1970.

[3] E. Diday, "The dynamic clusters method and optimization in nonhierarchial-clustering," in *Lecture Notes in Computer Science*, 3, pp. 211-258, 1974.

[4] L. C. W. Dixon and G. P. Szegö, Eds., *Towards Global Optimization*, Proceedings of a Workshop at the University of Cagliari, Italy, October 1974 (North-Holland/Americal Elsevier, Oxford, and New York), 1975.

[5] L. C. W. Dixon, J. Gomulka, and S. E. Hersom, "Reflections on the global optimization problem," Numerical Optimization Center, Tech. Rep. No. 64, The Hatfield Polytechnic.

[6] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[7] M. G. Kendall, "Cluster analysis," in *Frontiers of Pattern Recognition*, S. Watnabe, Ed. New York: Academic, 1972.

[8] W. S. Meisel, *Computer-Oriented Approaches to Pattern Recognition*. New York: Academic, 1972.

[9] E. A. Patrick, *Fundamentals of Pattern Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1972.

[10] H. R. Ruspini, "Numerical methods for fuzzy clustering," *Information Science*, 2, pp. 319-350, 1970.

[11] A. Törn, "Global optimization as a combination of global and local search," *Simulation Versus Analytical Solutions for Business and Economic Models* (proceedings), Gothenburg, BAS 17, 1973.

[12] ——, "Global optimization as a combination of global and local search," Skriftserie utgiven vid Handelshögskolan vid Åbo Akademi, A: 13, 1974.

[13] ——, "Cluster analysis as a tool in a global optimization model," presented at the 3rd Int. Cong. Cybernetics and Systems. Bukarest, Aug. 1975.

[14] ——, "Probabilistic global optimization, a cluster analysis approach," 2nd European Cong. Operations Research (proceedings), Stockholm, 1976.

[15] ——, "A search-clustering approach to global optimization," in *Towards Global Optimization*, Second Volume, L. C. W. Dixon and G. P. Szegö, Eds. North-Holland, 1977.

Fixed Structure Automata in a Multi-Teacher Environment

DANIEL E. KODITSCHEK AND KUMPATI S. NARENDRA, SENIOR MEMBER, IEEE

Abstract—**The concept of an automaton operating in a multiteacher environment is introduced, and several interesting questions that arise in this context are examined. In particular, we concentrate on the consequences of adding a new teacher to an existing n-teacher set as it affects the choice of a switching strategy. The effect of this choice on expediency and speed of convergence is presented for a specific automaton structure.**

I. INTRODUCTION

THE THEORY of learning automata was introduced in 1961 by Tsetlin [1], and is now studied quite extensively in both the Soviet Union and the United States. Learning, in this field, is investigated through the formal paradigm of an automaton interacting with a generalized environment. The automaton, finding its behavior repertoire confronted by a set of environmental consequences which it preferentially orders, learns to choose a behavior which produces the most favorable response. Narendra and Thathachar survey this theory and review some fundamental questions in [2]. The power and versatility of automata schemes is being slowly recognized, and several applications have recently been suggested which involve learning in complex situations [2], [3].

In spite of these advances, almost all schemes developed have dealt with a single automaton operating in a single environment. Yet much of what we generally call learning takes place outside this context. While some of our actions are better than others, it is often impossible to compare them linearly. Students are exposed to a variety of teachers; businesses are confronted with conflicting expertise; babies (usually) learn from two parents. In all cases learning seems to be effected in spite of the vector character of the teachers' responses. This paper is concerned with further generalizing the concept of environment to include these common situations. More formally, we shall undertake to modify the theory of learning automata as it currently exists by replacing a totally ordered input set to the automaton with a partially ordered set and study the effects upon optimal learning strategies. The modification will be accomplished by exposing automata to several single-teacher environments.

There is an ambiguity in learning automata theory be-

Manuscript received September 27, 1976; revised February 28, 1977. This work was supported by the National Science Foundation under Grant ENG75-08616.

The authors are with the Department of Engineering and Applied Science, Yale University, New Haven, CT 06520.

tween the notion of a teacher and an environment. While this has been unimportant in the past, it becomes crucial in the context of the problem discussed in this paper. We will distinguish between an environment which rewards or penalizes the automaton directly and a teacher which conveys information about an implicit environment by approving or disapproving the automaton's behavior. This distinction arises naturally when we address the question whether "learning" is manifested by the capacity to gain approval from conflicting judges or by the tendency to perform one "correct" action more often than others. Our interest lies primarily in the latter situation: we postulate the existence of an underlying environment within which there exists a correct behavior pattern which is inaccessible to the automaton. A set of teachers which attempts to describe this environment forms the "multi-teacher environment" with which the automaton interacts. The objective of the automaton then is to interrogate the teachers and determine the particular behavior pattern that the underlying environment treats as the best. To make the problem meaningful we postulate, regardless of other differences, that the teachers "agree" with the underlying environment on the ordering of the automaton's actions.

For a more complete understanding of the problems that may arise in multiple-teacher environments, it is necessary to introduce some distinctions between teachers who satisfy the postulate stated above. For discussion purposes we distinguish between benign, harsh, and good teachers. While the first two are defined only qualitatively, the last concept will be defined as a relation over the set of all teachers. Using these, an investigation of the nature of "improvement" in a multi-teacher environment is undertaken in the body of the paper.

We shall first formulate the problem a little more precisely in Section II and then address the following questions in the succeeding sections. It seems intuitively clear that two good teachers must provide more information than one. Is it true that two teachers are better than either one alone, or that $n + 1$ teachers are better than any n of them? If more teachers improve the automaton's accuracy, do they necessarily decrease the rate of the learning process? Is it sometimes preferable to ignore a new teacher's response? Are there different effective learning strategies in the multi-teacher environment, or is the automaton limited to one optimal strategy?

It will be shown in this paper that the switching strategy chosen almost completely affects the answers to the foregoing questions. There exists a strategy by which we can always assure that $n + 1$ teachers are "better" than any n of them. However, such a strategy affects the speed of learning adversely and may consequently not be of great practical interest. If convergence rate is an important consideration, then the automaton may do better to choose strategies whose responses are worsened by the addition of the new teacher. Whether a new teacher is to be ignored or included depends on the performance criterion as well as the old strategy used and the information that is available regarding the new teacher.

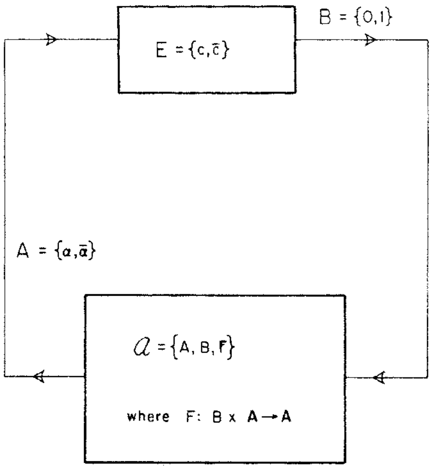

Fig. 1. P-model learning automaton.

II. The Problem

The multi-teacher environment considered in this paper will be confined to the context of a fixed structure automaton. We describe the P-model learning automaton briefly in this section. For a fuller discussion of this topic and for further information regarding other learning models, the reader is referred to [4].

A learning automaton is an automaton-teacher pair $(\mathscr{A},E_i)$ as shown in Fig. 1. The automaton $\mathscr{A}$ is described by the triple $\{A,B,F\}$, where A is an action set $\{\alpha,\bar{\alpha}\}$, B is the input set $\{0,1\}$, and $F\colon A \times B \to A$ is the switching strategy. The teacher is defined by the set of probabilities $\{c_i,\bar{c}_i\}$, where $c_i = \Pr\{\beta = 1 \mid \alpha\}$ and $\bar{c}_i = \Pr\{\beta = 1 \mid \bar{\alpha}\}$ (note: $\beta \in B$). The automaton performs one of the actions α or $\bar{\alpha}$, receives a reward ($\beta = 0$) or penalty ($\beta = 1$) from the teacher, and uses the information in its updating strategy to determine a new action. At any time t, the automaton will perform α with total probability $\pi(t)$ and $\bar{\alpha}$ with total probability $\bar{\pi}(t)$, where the total probability vector

$$\pi(t) = \begin{bmatrix} \pi(t) \\ \bar{\pi}(t) \end{bmatrix}$$

is a Markov sequence. An expedient strategy is defined as [4]

$$\lim_{t\to\infty} M(t) = E[\beta(t) \mid \pi(t)] = c_1\pi + \bar{c}_1\bar{\pi} \le \frac{c_1 + \bar{c}_1}{2}.$$

Expedience results in the expected penalty being less than chance. It also corresponds to the better action being chosen with higher probability than the worse action (i.e., $\bar{\pi} > \pi$).

The single-teacher environment Tsetlin automaton discussed above will be denoted $\mathscr{A}(1)$. To facilitate discussion of higher order environments, we introduce the measure space associated with $\mathscr{A}(1)$, $\{B,\mathscr{P}(B),p_i\}$, where B is the quotient set $\{0,1\}/A$ and p_i is the measure on $\mathscr{P}(B)$ assigning $\{0\}/\alpha \mapsto 1 - c_i$; $\{1\}/\alpha \mapsto c_i$; $\{0\}/\bar{\alpha} \mapsto 1 - \bar{c}_i$; $\{1\}/\bar{\alpha} \mapsto \bar{c}_i$.

Fig. 2 presents the general n-teacher environment $\mathscr{A}(n)$. This is the pair $(\mathscr{A}, \{E_i\}, i = 1, 2, \cdots, n)$ where $\{E_i\}$ represents a set of n teachers. The input set to the automaton is now the set of all n-tuples of zeros and ones and is denoted B^n where

$$B^n = \prod_{i=1}^{n} \{0,1\}_i.$$

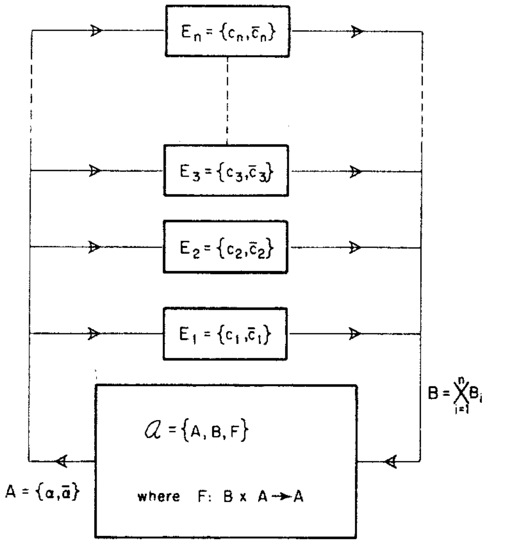

Fig. 2. *n*-teacher environment.

The measure space $(B^n,\mathscr{P}(B^n),P_n)$ corresponds to the environment $\mathscr{A}(n)$, where P_n is the measure on $\mathscr{P}(B^n)$ induced by the Cartesian product of the individual teacher's probabilities making the assignments:

$$(0, 0, \cdots, 0)/\bar{\alpha} \mapsto \prod_{i=1}^{n} (1 - \bar{c}_i)$$

$$\vdots$$

$$(1, 1, \cdots, 1)/\bar{\alpha} \mapsto \prod_{i=1}^{n} \bar{c}_i$$

$$(0, 0, 0, \cdots, 0)/\alpha \mapsto \prod_{i=1}^{n} (i - c_i)$$

$$\vdots$$

$$(1, 1, \cdots, 1)/\alpha \mapsto \prod_{i=1}^{n} c_i.$$

The distinction between a multi-teacher environment and a joint exposure to many environments may now be clarified. In the former case we postulate an underlying environment E, which is inaccessible to $\mathscr{A}$ and has $c = 1$, $\bar{c} = 0$. Thus E defines $\bar{\alpha}$ as the "correct" behavior. It is this information that $\mathscr{A}$ must learn from the multi-teacher environment. As stated in the introduction, we postulate that all teachers "agree" with E, i.e., all of the teachers agree that $\bar{\alpha}$ is the correct behavior so that $c_i > \bar{c}_i$. The set D of acceptable teachers is defined by

$$D \triangleq \{E_i \in [0,1] \times [0,1] \,|\, c_i > \bar{c}_i\} \tag{1}$$

and is shown in Fig. 3. Qualitatively, if c_i is small we will call E_i "benign," and if $\bar{c}_i$ is large we will call the teacher "harsh." A benign teacher will tend to call both actions good, while a harsh teacher will tend to call both actions bad most of the time.

In the case of the n-teacher environment, the former simple situation no longer prevails. The set B^n is unordered, and an output cannot be classified as a reward or a penalty in a straightforward manner. Intuitively, a response $(1, 1, \cdots, 1)$ must be a penalty and $(0, 0, \cdots, 0)$ must be a reward. But there are $2^n - 2$ other actions which must be categorized before a switching strategy can be developed. In this context it no longer makes sense to talk about the minimum expected penalty, since penalty is undefined. Consequently, the obvious interpretation that the expectation of a 1 in the input should be minimized is not meaningful. The automaton's primary objective in this case is to maximize the final probability of choosing the correct action $\bar{\alpha}$, and the outputs of the environments are interpreted as rewards or penalties to aid in achieving this objective. Hence, we will define expedience by the condition

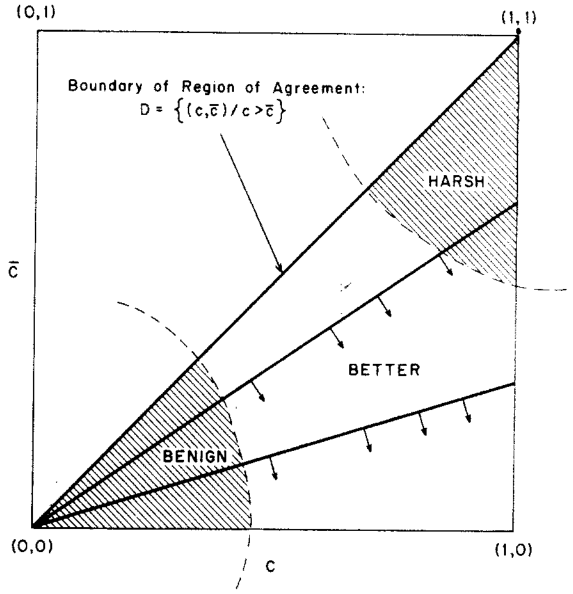

Fig. 3. Characteristics of teachers.

$$\lim_{t\to\infty} \bar{\pi}(t) = \bar{\pi} > \tfrac{1}{2} \tag{2}$$

and if $\mathscr{A}_1$ and $\mathscr{A}_2$ evolve total probabilities $\bar{\pi}_1$ and $\bar{\pi}_2$, we shall say $\mathscr{A}_1$ is more expedient than $\mathscr{A}_2$ if and only if $\bar{\pi}_1 > \bar{\pi}_2$. Thus expedience has been redefined as a total ordering over the set of all automata.

A natural comparison of teachers would determine which produces the most expedient $\mathscr{A}$ if it acts alone. Given two teachers E_i and E_j we find that $\bar{\pi}_j \geq \bar{\pi}_i$ if and only if

$$\frac{\bar{c}_i}{c_i} \geq \frac{\bar{c}_j}{c_j}. \tag{3}$$

Hence, we define a "good teacher" E_j as one whose ratio $\bar{c}_j/c_j$ is low, and we say that it is "better" than E_i if condition (3) holds. By this definition, "goodness" is a total ordering over the set D of all teachers.

III. The Two-Teacher Environment

The two-teacher environment $\mathscr{A}(2)$ is a prototype of all higher order environments. Strategies and solutions in $\mathscr{A}(n)$ are motivated by results and considerations in $\mathscr{A}(2)$. In particular, the interesting questions regarding strategic adjustment from $\mathscr{A}(n)$ to $\mathscr{A}(n+1)$ are addressed in the passage from $\mathscr{A}(1)$ to this case.

A. A Simple Strategy

Given $(A,\{E_1,E_2\})$ as shown in Fig. 2, the probabilities assigned to B^2 by P_2 are the following.

β	$P_2(\beta/\alpha)$	$P_2(\beta/\bar{\alpha})$
(1,1)	c_1c_2	$\bar{c}_1\bar{c}_2$
(0,1)	$(1-c_1)c_2$	$(1-\bar{c}_1)\bar{c}_2$
(1,0)	$c_1(1-c_2)$	$\bar{c}_1(1-\bar{c}_2)$
(0,0)	$(1-c_1)(1-c_2)$	$(1-\bar{c}_1)(1-\bar{c}_2)$

The automaton's response to these inputs will completely determine the behavior of the overall system. Intuitively, $\mathscr{A}$ should switch actions on (1,1) and treat (0,0) as a reward. The appropriate response to (0,1) and (1,0) is less obvious. Recalling from [2] that reward-inaction is an absolutely expedient strategy, one may consider the use of an inaction strategy for $\{(0,1),(1,0)\}$. However, for a fixed structure single-memory-level automaton, inaction is the reward strategy, i.e., in $\mathscr{A}(1)$ the reward transition matrix

$$F^0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Hence such a strategy will lead to

$$F^{(0,0)} = F^{(0,1)} = F^{(1,0)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$F^{(1,1)} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

yielding the stochastic Markov matrix

$$P_{21} = \begin{bmatrix} 1 - c_1c_2 & c_1c_2 \\ \bar{c}_1\bar{c}_2 & 1 - \bar{c}_1\bar{c}_2 \end{bmatrix}.$$

The eigenvector of P_{21} corresponding to the eigenvalue $\lambda = 1$ is

$$\boldsymbol{\pi}_{21} = \begin{bmatrix} \dfrac{\bar{c}_1\bar{c}_1}{c_1c_2 + \bar{c}_1\bar{c}_2} \\ \dfrac{c_1c_2}{c_1c_2 + \bar{c}_1\bar{c}_2} \end{bmatrix}.$$

This is clearly expedient since $c_1c_2 \geq \bar{c}_1\bar{c}_2$. The more interesting subsequent question is whether $\mathscr{A}(2)$ is more expedient under this strategy than was $\mathscr{A}(1)$, i.e., is $\bar{\pi}_1 \leq \bar{\pi}_{21}$?

The inequalities

$$\frac{c_1}{c_1 + \bar{c}_1} \leq \frac{c_1c_2}{c_1c_2 + \bar{c}_1\bar{c}_2} \quad \text{or} \quad \frac{c_2}{c_2 + \bar{c}_2} \leq \frac{c_1c_2}{c_1c_2 + \bar{c}_1\bar{c}_2}$$

are equivalent to

$$c_1\bar{c}_1[c_2 - \bar{c}_2] \geq 0 \quad \text{or} \quad c_2\bar{c}_2[c_1 - \bar{c}_1] \geq 0,$$

hence, they are always true (since $c_i \geq \bar{c}_i$ by (1); agreement).

The Measure Space M_{21}: In the above discussion, if we define $c^* \triangleq c_1c_2$ and $\bar{c}^* \triangleq \bar{c}_1\bar{c}_2$, then c^* and $\bar{c}^*$ become the "penalty probabilities" corresponding to the actions α and $\bar{\alpha}$ of a composite teacher $\{E_1,E_2\}$. This intuitive realization may be derived more formally as follows. The partition on the input set B^2 imposed by the switching strategy results in a "reward" subset which allows its elements to contain one penalty component. These subsets

$$G_{21} \triangleq \{(0,0),(0,1),(1,0)\} \quad \text{and} \quad G_{21}^c \triangleq \{(1,1)\}$$

generate a new measure space M_{21} imposed by the switching strategy. M_{21} is defined by the triple

$$M_{21} \triangleq (B^2,\mathscr{M}_{21},\rho_2)$$

where B^2 is the entire input space,

$$\mathscr{M}_{21} \triangleq \{\phi,G_{21},G_{21}^c,B^2\},$$

and the new measure function ρ_2 is the old P_2 summed over the elements of $\mathscr{M}_{21}$:

$$\rho_2(G) \triangleq \int_G dP_2.$$

By this formal reasoning we see that the switching strategy has "coarsened" the measure set from the power set of B^2 to $\mathscr{M}_{21}$ so that $\rho_2(G_{21}^c) = c_1c_2$ is the penalty probability of a new composite teacher. We will refer to this strategy in $\mathscr{A}(2)$ as defining the automaton (and overall environment) $\mathscr{S}_{21}$. The expedient strategy in $\mathscr{A}(1)$ will be denoted $\mathscr{S}_1$.

Other Strategies: The automaton $\mathscr{S}_{21}$ described above has two teachers and interprets any input with at most one penalty (the second subscript) as a "reward." According to this definition, the automaton $\mathscr{S}_{20}$ with measure space $M_{20} \triangleq (B^2,\mathscr{M}_{20},\rho_2)$ would allow into the "reward" subset only inputs with no penalty component. Under this strategy the final total probability vector is

$$\boldsymbol{\pi}_{20} = \begin{bmatrix} \dfrac{\bar{c}_{20}^*}{c_{20}^* + \bar{c}_{20}^*} \\ \dfrac{c_{20}^*}{c_{20}^* + \bar{c}_{20}^*} \end{bmatrix} = \begin{bmatrix} \dfrac{\bar{c}_1 + \bar{c}_2 - \bar{c}_1\bar{c}_2}{c_1 + c_2 - c_1c_2 + \bar{c}_1 + \bar{c}_2 - \bar{c}_1\bar{c}_2} \\ \dfrac{c_1 + c_2 - c_1c_2}{\bar{c}_1 + \bar{c}_2 - \bar{c}_1\bar{c}_2 + c_1 + c_2 - c_1c_2} \end{bmatrix}.$$

Since $\bar{c}_{20}^* < c_{20}^*$ we know that $\mathscr{S}_{20}$ is expedient. However, the inequality $\bar{\pi}_{21} > \bar{\pi}_{20}$ holds if and only if

$$\frac{c_1c_2}{c_1c_2 + \bar{c}_1\bar{c}_2} \geq \frac{c_1 + c_2 - c_1c_2}{c_1 + c_2 - c_1c_2 + \bar{c}_1 + \bar{c}_2 - \bar{c}_1\bar{c}_2}$$

or

$$c_1\bar{c}_1(c_2 - \bar{c}_2) + c_2\bar{c}_2(c_1 - \bar{c}_1) \geq 0,$$

which is always true by the "agreement" postulate (1). Hence, we have the result that, independent of the characteristics of the two teachers, $\mathscr{S}_{20}$ is always less expedient than $\mathscr{S}_{21}$. Further, the two teachers are not necessarily better than either one taken separately since $\bar{\pi}_{20} > \bar{\pi}_{10}$ if and only if

$$\frac{\bar{c}_2}{c_2} \leq \frac{\bar{c}_1(1 - c_1)}{c_1(1 - \bar{c}_1)} \tag{4}$$

which is not always true.

We may briefly comment on the only remaining viable deterministic switching strategies in $\mathscr{A}(2)$, which ignore one or the other of the two teachers, respectively. For example, if the automaton switches on receiving the inputs (1,1) and

(0,1), then clearly the first teacher E_1 is being disregarded. Similarly if it switches on (1,1), (1,0), the second teacher is being disregarded. In Section III-B we discuss the conditions under which this policy becomes attractive.

The preceding discussion has demonstrated that the strategy used by an automaton determines whether or not two teachers as a composite entity are going to be unconditionally better than any one teacher. We know $\mathscr{S}_{21}$ is more expedient than the strategy in $\mathscr{A}(1)$ using either E_1 or E_2. The composite teacher is better than either one used singly. However, $\mathscr{S}_{20}$ does not enjoy such properties even though it is also expedient; the characteristics of one of the teachers may actually result in a deterioration of the overall performance.

B. *Rate of Convergence and Choice of Strategy*

It is clear from the previous section that $\mathscr{S}_{21}$ is the most expedient strategy in $\mathscr{A}(2)$. Why, then, should we even mention any others? Unfortunately expedience may be bought at cost of speed: rate of convergence to steady-state behavior is an equally important attribute. The automaton $\mathscr{S}_{21}$ does not switch actions unless the outputs of both teachers are penalties. The less frequent this occurrence the fewer the automaton's "experiments" with alternative behaviors. While the convergence problem in $\mathscr{A}(2)$ does not necessarily preclude $\mathscr{S}_{21}$, a large number of teachers will mar the desirability of its higher order analogues precisely because the automaton may continue to perform the same action for long periods of time. In this section we briefly consider the different strategies in a two-teacher environment with regard to convergence rate criteria.

A measure of the convergence rate of an automaton [4] is the distance from unity of the second eigenvalue of the probability transition matrix P. Let $r_{nm} = |1 - c_{nm}^* - \bar{c}_{nm}^*|$ denote this distance for strategy $\mathscr{S}_{nm}$ in the n-environment case. By inspection, in the two-environment case, we have $c_{21}^* + \bar{c}_{21}^* < c_1 + \bar{c}_1 < c_{20}^* + \bar{c}_{20}^*$ all of which values increase monotonically between 0 and 2 as some function of $\|(c_1,c_2,\bar{c}_1,\bar{c}_2)\|$ depicted in Fig. 4. Four regions are constructed in the figure illustrating where the sum of the penalties using the various strategies are greater than unity. In region 1 it is clear that r_{21} is lower than either r_1 or r_{20}. In this region E_1 and E_2 are both harsh enough to bring $c_{21}^* + \bar{c}_{21}^* > 1$. Hence, strategy $\mathscr{S}_{21}$ is both more expedient and converges more rapidly than either $\mathscr{S}_{20}$ or $\mathscr{S}_1$ (with either E_1 or E_2). Outside of this small region, however, we are faced with the choice between speed and accuracy (i.e., between convergence rate and expedience).

Disregarding $\mathscr{S}_{21}$, we are left to choose between $\mathscr{S}_{20}$ and $\mathscr{S}_1(i)$ ($i = 1$, or 2), the latter notation denoting the case where either teacher E_2 or E_1 is ignored, respectively. If the strategy $\mathscr{S}_{20}$ is chosen, the reward set includes only the element (0,0). While this strategy is expedient, it is more or less expedient than $\mathscr{S}_1(1)$ (ignoring the second teacher) depending on whether (4) is true or not, i.e.,

$$\frac{\bar{c}_2}{c_2} \leq \frac{\bar{c}_1(1-c_1)}{c_1(1-\bar{c}_1)}.$$

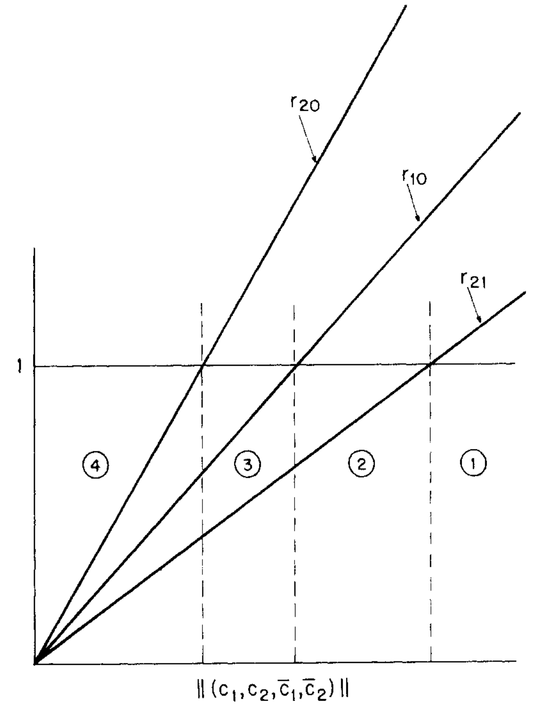

Fig. 4. Convergence relations.

In other words, condition (4) determines when E_2 "improves" E_1 in strategy $\mathscr{S}_{20}$. Hence, we have defined a new strict partial ordering on the set of all teachers yielding properties:

i) E_1 can never improve itself,
ii) E_i and E_j may be found such that neither improves the other.

Fig. 5 illustrates an occurrence of ii).

Fig. 6 magnifies region 2 of Fig. 4, where the inequalities $c_1 + \bar{c}_1 > 1$ and $c_{21}^* + \bar{c}_{21}^* < 1$ are satisfied. We consider $2a$,$2b$,$2c$ defined by g, where $r_{21} = d_1 = r_1$, and h, where $r_{21} = d_2 = r_{20}$, as depicted. In region $2a$, $\mathscr{S}_{21}$ is not only more expedient but also faster than the other strategies (as in region 1). In $2c$ the teachers yield $r_1 < r_{20} < r_{21}$, while in $2b$, $r_1 < r_{21} < r_{20}$. Consequently, we must choose a performance criterion when the teachers have characteristics of these latter two subregions. This choice entails the following considerations. If speed of convergence is more desirable, then we must ignore the second teacher. If a compromise between speed and accuracy is desired, then we must determine whether E_2 improves (4) E_1. If there is no improvement, the choice of $\mathscr{S}_1(1)$ is again immediate. If the second teacher does improve the first, then we will choose $\mathscr{S}_{20}$ in $2c$, and $\mathscr{S}_{21}$ in $2b$.

A similar discussion holds in region 3 of Fig. 4, where r_1 is always smaller than r_{21}. Again, the choice of strategy depends on the performance criterion and the specific character of E_2. In region 4, $\mathscr{S}_{20}$ is always faster than both $\mathscr{S}_1$ and $\mathscr{S}_{21}$. We note again that in all four regions expedience is entirely independent of the convergence rate: 1) $\mathscr{S}_{21}$ is the most expedient strategy; 2) E_2 may or may not improve E_1, regardless of the region in question; 3) hence, speed and accuracy must be evaluated separately.

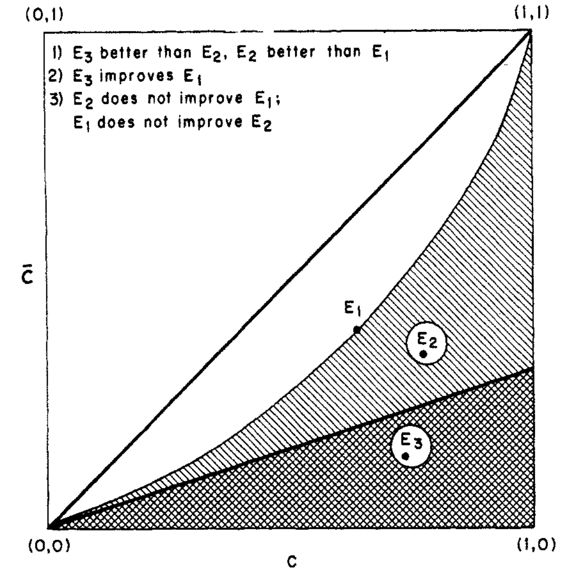

Fig. 5. Improvement relation.

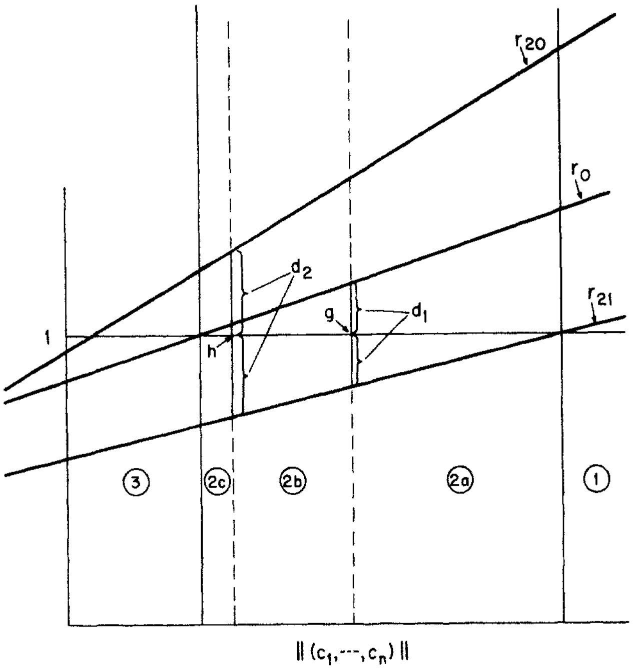

Fig. 6. Magnification of one convergence region.

In summary, it is clear that the choice of strategy in an environment with two teachers depends upon a higher level decision regarding performance criterion. If expedience is the sole criterion, $\mathscr{S}_{21}$ is unquestionably optimal. If convergence rate is the sole criterion, then we must have more information regarding the characteristics of E_1 and E_2. In region 1 the optimal strategy is always $\mathscr{S}_{21}$ and in region 4 it is $\mathscr{S}_{20}$. In regions 2 and 3 it is necessary to know the values of $c_1,\bar{c}_1,c_2,\bar{c}_2$ before a strategy is chosen. In these regions if E_2 does not improve E_1 and convergence rate is the sole criterion, then E_2 should be ignored.

A final observation may motivate our introduction of further requisite knowledge regarding the environment. Clearly, if $c_1,c_2,\bar{c}_1,\bar{c}_2$ are all known, then $\bar{c}_i < c_i$ implies $\bar{\alpha}$ is the better action, and there is nothing to be learned. However, in practice, even if the actual $c_i,\bar{c}_i$ are unknown, there are many situations where estimates of these probabilities are available based on past performance. In such cases the estimates rather than the true values of $c_i,\bar{c}_i$ are used in the decision process described above.

IV. The n-Teacher Environment

As in the previous section, a central concern of this discussion is the formulation of optimal strategies with respect to the different performance criteria which may arise in the context of varying degrees of information about an n-teacher environment. There are a greater number and variety of choices in $\mathscr{A}(n)$ than in $\mathscr{A}(2)$, but the situation is essentially the same. Mathematical proofs of many statements made in the section are lengthy, and have been relegated to the Appendix of the paper.[1] We intend in this section merely to highlight the salient results of that more formal mathematical investigation.

A. $\mathscr{S}$ Strategies and Their Measure Spaces

As described in Section II, the response of n teachers to an input is an n-dimensional vector of ones and zeros. $\mathscr{S}_{nm}$ denotes the strategy whose reward subset G_{nm} contains all vectors with no more than m penalty components (i.e., nonzero entries). In other words, the automaton switches only when at least n-m teachers disapprove of an action. As in Section III, the measure space associated with this strategy is $M_{nm} = (B^n,\mathscr{M}_{nm},\rho_n)$. M_{nm} is proven to be a true probability space in P.1. All such strategies are shown to be expedient in P.2 using the monotonicity of $\rho_n(G_{nm})$ when considered as a function of $\{c_i,\bar{c}_i\}_{i=1}^n$.

B. Central Results of the Analysis

Interesting questions arise when we know nothing about the n teachers and are interested in the relative expedience of two strategies. In P.3 the central result of this paper, that $\mathscr{S}_{nk}$ is more expedient than $\mathscr{S}_{nl}$ if and only if $l + 1 \leq k$, is proved (of course, $k < n$). That this seemingly intuitive result requires a relatively complicated proof may be motivated by the following considerations. Allowing the automaton to consider an input vector with more penalty components as a member of the "reward" subset naturally (tautologously) decreases the incidence of what the automaton regards as a "penalty." But this decrease obtains for α and $\bar{\alpha}$ together. Our definition of expedience (2) reflects the characterization of learning (established in the introduction) as a process of distinguishing between the relative values of two actions. Hence the fact that both yield more "rewards" does not aid the learning process. But the proposition is true. The new composite teacher is not only more benign but is also better: the probability of choosing the correct action increases.

A consequence of this result concerns the effect of adding a new teacher to an n-teacher set. In P.4 it is shown that $\mathscr{S}_{n+1,k}$ is more expedient than $\mathscr{S}_{nl}$ (for any combination of n teachers out of the $n + 1$) if and only if $k \geq l + 1$ (again

[1] The facts, propositions and their corollaries in the Appendix are numbered separately. Hence, e.g., P.4, which denotes the result proven as "Proposition 4" occurs before C.1, "Corollary 1."

$k < n$). In other words, increasing the number of teachers by one results in a more expedient automaton if the strategy is modified to include an extra penalty component in its reward set.

It follows immediately that the most expedient strategy for an n-teacher environment is one where input vector $(1, 1, \cdots, 1)$ is the sole element of the penalty set. A more expedient automaton can only result from the addition of an extra teacher.

C. Convergence Rate in $\mathscr{A}(n)$

The extended discussion of convergence rate in Section III-B carries over to $\mathscr{A}(n)$ with n functions $\{r_{ni}\}_{i=0}^{n-1}$ defining $n + 1$ regions as they cross unity. Corresponding to region 1 of Fig. 4, if the teachers are very harsh, then the relative frequency of $(1, 1, 1, \cdots, 1)$ will be sufficient to make $\mathscr{S}_{n,n-1}$ the fastest strategy. On the other extreme, corresponding to region 4 of that figure, if the teachers are very benign, then penalty components will be very rare in the input vectors, and switching will be too infrequent unless $\mathscr{S}_{n0}$ is used. In between these extremes are the regions corresponding to 2 and 3 where we require more knowledge of teachers' values and a predetermined performance criterion as described for $\mathscr{A}(2)$. We mention again that in practice this "extra" knowledge will obtain from estimates of $\{c_i,\bar{c}_i\}_{i=1}^{n}$ based on previous and accumulating experience.

D. $\mathscr{C}$ Strategies

An alternative class of strategies in $\mathscr{A}(n)$ is suggested by the argument used to develop $\mathscr{S}_{2i}$ in the transition from $\mathscr{A}(1)$ to $\mathscr{A}(2)$. As has been described in Section III, using an $\mathscr{S}_{n-1,m}$ strategy in $\mathscr{A}(n-1)$ actually produces a composite teacher $\{E_i\}_{i=1}^{n-1}$, with penalty probabilities $\{\rho_{n-1}(G_{n-1,m}^c|\alpha), \rho_{n-1}(G_{n-1,m}^c|\bar{\alpha})\}$. In this light we may view the addition of the extra teacher E_n to the $n-1$ teacher set as forming $\mathscr{A}(2)$ out of $(\mathscr{A}, \{\{E_i\}_{i=1}^{n-1}, E_n\})$ rather than $\mathscr{A}(n)$ out of $(\mathscr{A}, \{E_1\}_{i=1}^{n})$, that is, we will treat $\{E_i\}_{i=1}^{n-1}$ and E_n as two separate teachers rather than simply adjoining E_n into the $n-1$ teacher set. A possible strategy would be to switch actions only when the new teacher responds with a 1 and the $n-1$ teacher input vector falls into the penalty subset $G_{n-1,m}^c$ (i.e., has fewer than $n - 1 - m$ zeros). We will refer to such strategies as $\mathscr{C}_{n,m+1}^1$, denoting an $\mathscr{S}_{21}$ construction of E_n and $\{E_i\}_{i=1}^{n-1}$ (with $\mathscr{S}_{n-1,m}$). Similarly, $\mathscr{C}_{nm}^0$ will denote an $\mathscr{S}_{20}$ construction of E_n and $\{E_i\}_{i=1}^{n-1}$ (with $\mathscr{S}_{n-1,m}$): the automaton will switch strategies unless the new teacher responds with a zero and the $n-1$ teacher input vector falls into the reward subset $G_{n-1,m}$ (i.e., has fewer than m ones). We note that $\mathscr{C}_{nm}^0$ is more expedient than $\mathscr{S}_{n-1,m}$ in $\mathscr{A}(n-1)$ if and only if E_n improves $\{E_i\}_{i=1}^{n-1}$ (under the $\mathscr{S}_{n-1,m}$ strategy). Further, the reader should realize that $\mathscr{S}_{n,n-1} = \mathscr{C}_{n,n-1}^1$ and $\mathscr{S}_{n0} = \mathscr{C}_{n0}^0$.

E. Transition from $\mathscr{A}(n-1)$ to $\mathscr{A}(n)$

It turns out that expedience comparisons between $\mathscr{C}$ and $\mathscr{S}$ strategies provide useful tests of the value of an extra teacher. The interpretation of C.1–C.3 is as follows. We start with an $n-1$ teacher set and employ an $\mathscr{S}_{n-1,m}$ strategy ($m \leq n-3$). We are allowed the use of a new teacher E_n. Should it be ignored; or, if not, what strategy should be used in $\mathscr{A}(n)$? E_n does not improve the $n-1$ teacher set if we know $\mathscr{S}_{n-1,m+1}$ (ignoring E_n) is more expedient than either $\mathscr{S}_{n,m+1}$ or $\mathscr{C}_{n,m+1}^1$, (using E_n). If E_n does improve the $n-1$ teacher set, then $\mathscr{S}_{n,m+1}$ is more expedient than $\mathscr{S}_{n-1,m+1}$ (ignoring E_n); moreover $\mathscr{C}_{n,m+1}^1$ yields a more expedient automaton than $\mathscr{S}_{n,m+1}$, even though both are using E_n. Of course, if $m = n - 2$, a more expedient strategy in $\mathscr{A}(n-1)$ does not exist, and we shall always use the new teacher with a $\mathscr{C}_{n,m+1}^1 = \mathscr{C}_{n,n-1}^1 = \mathscr{S}_{n,n-1}$ strategy.

Since $r_{n-1,m+1}$ is always smaller than $r_{n,m+1}$, and very likely (if E_n is benign enough) smaller than the rate function associated with $\mathscr{C}_{n,m+1}^1$, ignoring E_n and increasing the size of the reward set in $\mathscr{A}(n-1)$ will yield a slower strategy unless $r_{n-1,m+1}$ is closer to unity than the other two functions. Hence, if no performance cost is placed on an additional teacher, we may use it whether or not the $n-1$ teacher set is improved: the difference in accuracy will be made up by speed.

The importance of a performance criterion should be clear from the foregoing discussion. More obviously, all of the comparisons described in this section will require varying degrees of information regarding the teachers.

V. Stochastic Strategies and Memory

Up to this point we have discussed two-state deterministic automata. It is reasonable to suspect that stochastic strategies, allowing a more flexible (than binary) switching choice, will thereby make more effective use of the different "strengths" of approval accorded by an n-teacher set. This, however, is not the case. From [4] we know that memory capacity enhances the automaton's expedience. The multiteacher environment will augment this effect.

A. Stochastic $\mathscr{S}$ Strategies

A stochastic strategy in $\mathscr{A}(1)$ is characterized by state transition matrices

$$F^1 = \begin{bmatrix} \xi_1 & 1 - \xi_1 \\ 1 - \xi_1 & \xi_1 \end{bmatrix}$$

and

$$F^0 = \begin{bmatrix} \xi_0 & 1 - \xi_0 \\ 1 - \xi_0 & \xi_0 \end{bmatrix}$$

where $\xi_i \in [0,1]$. Thus a given input causes a switching response with probability $1 - \xi_i$. While $\xi_0 = 1$ is necessary for expedience, ξ_1 does not effect expedience at all and may increase or decrease convergence rate depending upon whether $c_1 + \bar{c}_1 > 1$ [4]. This uncompelling result may prepare us for equally lackluster consequences in $\mathscr{A}(n)$.

The obvious extension of stochastic capability in $\mathscr{A}(n)$ is characterized by transition matrices

$$F^\beta = \begin{bmatrix} \xi_i & 1 - \xi_i \\ 1 - \xi_i & \xi_i \end{bmatrix}$$

where the n components of β yield

$$\sum_{j=1}^{n} \beta_j = i.$$

The $\{\xi_i\}_{i=1}^n$ express the automaton's interpretation of the "degree" of disapproval implied by an input which contains i penalties. By a stochastic $\mathscr{S}_{nm}$ strategy we denote the situation where $\xi_i \in [0,1]$ $(i \leq m)$ and $\xi_i = 0$ $(m < i \leq n)$: the automaton will switch actions when the input is an element of G^c_{nm} and will consider elements of G_{nm} as rewards only with probability $1 - \xi_i$ (i is the number of component penalties). The result of P.5 demonstrates that every deterministic $\mathscr{S}_{nm}$ strategy is just as expedient or more expedient than its stochastic cousins. The most expedient stochastic strategy, $\mathscr{S}_{n,n-1}$ with $\xi_i = 1$ $(i < n)$ and $\xi_n \in [0,1]$ (i.e., never switches when input is in $G_{n,n-1}$, switches with probability $1 - \xi_n$ when input is all ones), is slower than $\mathscr{S}_{n,n-1}$ unless $r_{n,n-1} > 1$, which as mentioned in Sections III-B and IV-C, happens extremely rarely.

B. Stochastic $\mathscr{C}$ Strategies

Alternatively, we might implement a stochastic strategy that switches actions with probability $1 - \xi_1$ if the nth component of the input vector is a penalty, and if the first $n - 1$ components fall into $G^c_{n-1,m}$, while considering everything else as a reward. This is a stochastic $\mathscr{C}^1_{n,m+1}$ strategy. It is no more expedient than the deterministic $\mathscr{C}_{n,m+1}$, and will be slower unless the associated r (rate) function is greater than unity.

C. Memory

Allowing the automaton extra states provides the possibility of schemes with memory. In $\mathscr{A}(1)$ it is well known [4] that a Tsetlin m-memory level automaton is ε-optimal if $\bar{c}_1 < \frac{1}{2}$, and a Krinskiy model becomes ε-optimal as memory depth approaches infinity. Placing the automaton in a higher order environment employing the reward–penalty partitions discussed in this report amounts to substituting a particular composite teacher for the single teacher in $\mathscr{A}(1)$. Hence, if the composite teacher is better, the automaton must be more expedient. Since we know when a higher order environment becomes a better teacher, we know how to improve the performance of these established memory schemes. But, substituting a better teacher does not affect the essential characteristics of memory management. No switching strategies discussed in this report can make a Krinskiy or Tsetlin automaton ε-optimal (unless, of course, in the latter case, $\bar{c}^*_{nm} < \frac{1}{2}$, where $\bar{c}_1 > \frac{1}{2}$ previously).

However, the input sets from a multi-teacher environment suggest many new memory schemes which may very possibly induce stronger expedience characteristics than yet seen in deterministic fixed structure automata. Unfortunately, these schemes entail nonsparse $2m \times 2m$ (where m is the memory depth) algebraic matrices whose solutions are rather complicated. Those solutions are now being developed.

VI. Conclusion

We have replaced a single environment with a set of environments, and, positing some *a priori* correct behavior, have investigated fixed structure automata learning schemes that increase the probability of its being performed. By emphasizing the importance of learning a "correct" behavior, we have turned away from the problem of minimizing an an expected incidence of penalties, establishing a distinction between teachers and environments, and thereby motivating a strong initial postulate that all teachers agree (as to the correct behavior).

Initial analysis indicates that the relative success of a particular strategy depends upon how it partitions an unordered input set into reward and penalty subsets. Somewhat surprisingly, the maximally expedient partitions evince a single element penalty subset containing the vector with no reward components. Since the incidence of one out of 2^n inputs decreases drastically as n increases, these maximally expedient strategies will usually be very slow. Exact decisions regarding speed and accuracy may be analytically determined, but require further knowledge of the environment.

Future study of the n-teacher environment might fruitfully consider variable structure or sample mean automata. An investigation of the expected penalty minimization problem would also be interesting.

Appendix

We only present the proofs for the two central propositions of the paper, and simply list all other results (which have been proven in [5]).

We establish the following notational conventions:

$$B^n \triangleq \prod_{i=1}^{n} \{0,1\}_i$$

is the set of n-dimensional vectors with zero or one entries.

$$G_{np} \triangleq \left\{\beta \in B^n \,\middle|\, \sum_{i=1}^{n} \beta_i \leq p\right\}$$

is the set of such vectors containing p or fewer ones. $\mathscr{M}_n$ is the σ-algebra on B^n generated by the nested sequence $\{G_{ni}\}_{i=1}^n$.

$$\rho_n\colon \mathscr{M}_n \to \boldsymbol{R} \text{ by } \rho_n(G) = \int_G dP_n$$

(where P_n is the probability induced by the Cartesian product of teachers $\prod E_i$) is a measure on $\mathscr{M}_n$. $f_{nm} \triangleq \rho_n(G_{nm} \mid \alpha)$ is the ρ-measure of the set of vectors with no more than m penalty components, and hence, the reward probability for $\mathscr{S}_{nm}$. $\eta_{npq} \triangleq \rho_n(G_{np} - G_{nq} \mid \alpha) = f_{np} - f_{nq}$ is the ρ-measure of those vectors with exactly p penalty components.

Finally, let $\boldsymbol{x} \in [0,1]^n$. Then $f_{nm}(\boldsymbol{x})$, $\eta_{npq}(\boldsymbol{x})$ will denote the appropriate mappings from $[0,1]^n \to [0,1]$ formed by evaluating $\rho_n(G_{nm} \mid \alpha)$ when $c_i = x_i$ $(i = 1,n)$, and $\bar{f}_{nm}(\boldsymbol{y})$, $\bar{\eta}_{npq}(\boldsymbol{y})$ will denote those mappings formed by evaluating $\rho_n(G_{nm} \mid \bar{\alpha})$ when $\bar{c}_i = y_i$ $(i = 1,n)$. For any vector $\boldsymbol{u} \in [0,1]^n$, the $n - 1$ dimensional vector $\boldsymbol{u}_j$ will denote $(u_1, \cdots, u_{j-1}, u_{j+1}, \cdots, u_n)^T$.

Proposition 1: $(B^n, \mathscr{M}_n, \rho_n)$ is a probability space.

Proposition 2: For all n,m, and $\boldsymbol{x} \in [0,1]^n$, the reward probability function $f_{nm}(\boldsymbol{x})$ is monotonically decreasing in each x_i.

Proposition 3: Given the automaton-environment pair $\mathscr{A}(n)$, strategy $\mathscr{S}_{np}$ is more expedient than $\mathscr{S}_{nq}$ independent of teacher characteristics if and only if $n > p > q \geq 0$.

Proof: The hypothesis is equivalent to the condition $\bar{\pi}_{np} > \bar{\pi}_{nq}$, which, in turn, is equivalent to the inequality:

$$(*) \qquad \frac{\bar{\eta}_{nnp}}{\eta_{nnp}} < \frac{\bar{\eta}_{nnq}}{\eta_{nnq}}.$$

It will suffice to prove this for $p = q + 1$, which is done by induction on n.

i) *True for* $n = 2$*:* We need only consider the case $p = 1 = q + 1$: $(*)$ may be rewritten

$$\frac{1 - \bar{f}_{21}}{1 - f_{21}} < \frac{1 - \bar{f}_{20}}{1 - f_{20}}$$

which is equivalent to $c_1\bar{c}_1(c_2 - \bar{c}_2) + c_2\bar{c}_2(c_1 - \bar{c}_1) > 0$. This is true by the agreement postulate.

ii) *True for* n *implies true for* $n + 1$*:* We must show, for $y,x \in [0,1]^{n+1}$

$$\frac{\bar{\eta}_{n+1,n+1,q+1}(y)}{\eta_{n+1,n+1,q+1}(x)} < \frac{\bar{\eta}_{n+1,n+1,q}(y)}{\eta_{n+1,n+1,q}(x)}.$$

This may be recursively expanded and simplified, for each $j \leq n + 1$, as

$$\begin{aligned}(1 - x_j)(1 - y_j)&[\eta_{nnq+1}(x_j)\bar{\eta}_{nnq}(y_j) - \bar{\eta}_{nnq+1}(y_j)\eta_{nnq}(x_j)]\\ &+ x_j y_j[\eta_{nnq}(x_j)\bar{\eta}_{nnq-1}(y_j) - \eta_{nnq-1}(x_j)\bar{\eta}_{nnq}(y_j)]\\ &+ x_j(1 - y_j)[\eta_{nnq}(x_j)\bar{\eta}_{nnq}(y_j) - \eta_{nnq-1}(x_j)\bar{\eta}_{nnq+1}(y_j)]\\ &- y_j(1 - x_j)[\eta_{nnq}(x_j)\bar{\eta}_{nnq}(y_j) - \bar{\eta}_{nnq-1}(y_j)\eta_{nnq+1}(x_j)] > 0.\end{aligned}$$

By inductive hypothesis, there is some j for which the first and second terms are positive. Note that if $q = 0$, the second is, nevertheless, positive since $\eta_{nnk} > \bar{\eta}_{nnk}$ by P.2. It requires a separate proof by induction to show that the third term is positive. Finally, it can be shown that the third term is larger than the fourth since $x_j(1 - y_j) > y_j(1 - x_j)$ and $\bar{\eta}_{nnq-1}\eta_{nnq+1} - \eta_{nnq-1}\bar{\eta}_{nnq+1} > 0$ by inductive hypothesis. Hence the inequality is true.

Proposition 4: Given the automaton-environment pair $\mathscr{A}(n)$ and an extra teacher E_{n+1}, strategy $\mathscr{S}_{n+1,p}$ is more expedient than $\mathscr{S}_{nq}$ independent of teacher characteristics if and only if $n \geq p > q$.

Proof: The hypothesis is equivalent to $\bar{\pi}_{n+1p} > \bar{\pi}_{nq}$ which is true if and only if

$$\frac{\bar{\eta}_{n+1,n+1,p}}{\eta_{n+1,n+1,p}} < \frac{\bar{\eta}_{nnq}}{\eta_{nnq}}.$$

By recursively expanding the left-hand side, and simplifying, this yields the inequality

$$(*) \qquad y < \frac{\bar{\eta}_{nnq}\eta_{np,p-1}}{\eta_{nnq}\bar{\eta}_{npp-1}}x + \frac{1}{\eta_{nnq}\bar{\eta}_{npp-1}}[\bar{\eta}_{nnq}\eta_{nnp} - \eta_{nnq}\bar{\eta}_{nnp}].$$

Define $D' = \{(x,y) \in [0,1]^2 \mid (*) \text{ holds}\}$. By the agreement postulate on teachers we must have $(c_i,\bar{c}_i) \in D = \{(x,y) \in [0,1]^2 \mid x > y\}$.

i) *Sufficiency:* Assume $p \geq q + 1$ – show $D \subseteq D'$. By P.3 we know $\eta_{nnq}\bar{\eta}_{nnp-1} \leq \bar{\eta}_{nnq}\eta_{nnp-1}$ which may be rewritten as $\eta_{nnq}(\bar{\eta}_{npp-1} + \bar{\eta}_{nnp}) \leq \bar{\eta}_{nnq}(\eta_{npp-1} + \eta_{nnp})$, and this is equivalent to $\eta_{nnq}\bar{\eta}_{npp-1} - \bar{\eta}_{nnq}\eta_{npp-1} \leq \bar{\eta}_{nnq}\eta_{nnp} - \eta_{nnq}\bar{\eta}_{nnp}$. Since the inequality is made strict by multiplying the left-hand side by $x < 1$, we have, rearranging,

$$x < \frac{\bar{\eta}_{nnq}\eta_{npp-1}}{\eta_{nnq}\bar{\eta}_{npp-1}}x + \frac{1}{\eta_{nnq}\bar{\eta}_{npp-1}}[\bar{\eta}_{nnq}\eta_{nnp} - \eta_{nnq}\bar{\eta}_{nnp}].$$

But if $(x,y) \in D$, then $x > y$, hence $(x,y) \in D'$.

ii) *Necessity:* Assume $p \leq q$: show $D \nsubseteq D'$. If $p < q$ choose $\varepsilon > 0$ such that $\bar{\eta}_{nnq}\eta_{nnp} = \eta_{nnq}\bar{\eta}_{nnp} - 2\varepsilon$. Choose

$$\left(\frac{\varepsilon}{\eta_{nnq}\bar{\eta}_{npp-1}}, \frac{\varepsilon}{2\eta_{nnq}\bar{\eta}_{npp-1}}\right) \in D.$$

This is not an element of D'. If $p = q$ choose

$$\left(x, \frac{\eta_{nnq}\bar{\eta}_{npp-1}}{\bar{\eta}_{nnq}\eta_{npp-1}}\right) \in D.$$

This is not an element of D'.

Corollary 1: A necessary and sufficient condition for $\mathscr{S}_{n+1,k}$ in $(\mathscr{A},\{E_i\}_{i=1}^{n+1})$ to be more expedient than $\mathscr{S}_{nk}$ in $(\mathscr{A},\{E_j\}_{j=1}^{n})$ where $\{E_j\}_{j=1}^{n} \subseteq \{E_i\}_{i=1}^{n+1}$ is

$$\frac{\bar{c}_{n+1}}{c_{n+1}} < \frac{\bar{\eta}_{nnk}(\bar{c}_1, \cdots, \bar{c}_n)\bar{\eta}_{nkk-1}(\bar{c}_1, \cdots, \bar{c}_1)}{\eta_{nnk}(c_1, \cdots, c_n)\eta_{nkk-1}(c_1, \cdots, c_n)}.$$

The next two corollaries are consequences of the definitions of the $\mathscr{C}$ strategies and the inequalities in P.4.

Corollary 2: If $\mathscr{S}_{n+1,k}$ is more expedient than $\mathscr{C}^1_{n+1,k}$ in $(\mathscr{A},\{E_i\}_{i=1}^{n+1})$ then $\mathscr{S}_{nk}$ in $(\mathscr{A},\{E_i\}_{i=1}^{n})$ is more expedient than $\mathscr{S}_{n+1,k}$.

Corollary 3: If $\mathscr{C}^0_{n+1,k}$ in $(\mathscr{A},\{E_i\}_{i=1}^{n+1})$ is more expedient than $\mathscr{S}_{nk}$ in $(\mathscr{A},\{E_i\}_{i=1}^{n})$, then $\mathscr{C}^1_{n+1,k+1}$ is more expedient than $\mathscr{S}_{n+1,k+1}$ in $(\mathscr{A},\{E_i\}_{i=1}^{n+1})$.

The reader is reminded that a more complete exposition of stochastic strategies as well as the proof of the final proposition are to be found in [5].

Proposition 5: For all n, $\mathscr{S}_{nm}$ is as expedient as any stochastic strategy of its type.

References

[1] M. L. Tsetlin, *Automaton Theory and Modeling Biological Systems.* New York: Academic, 1973.

[2] K. S. Narendra and M. A. L. Thathachar, "Learning automata—A survey," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-4, pp. 323–334, July 1974.

[3] K. S. Narendra and E. A. Wright, "Application of learning automata to telephone traffic routing problems," Becton Center Tech. Rep. CT-69, Yale University, May 1976.

[4] K. S. Narendra and M. A. L. Thathachar, *Learning Automata*, to be published.

[5] D. E. Koditschek and K. S. Narendra, "Fixed structure automata in a multi-teacher environment," Becton Center Tech. Rep. CT-72, Yale University, Sept. 1976.