The coefficient matrix reduces by (3.5) to

$$(1/k)A'A = (1/k)\begin{bmatrix} k & -1 & \cdot & \cdot & \cdot & -1 \\ -1 & k & \cdot & & & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & & & \cdot & k & -1 \\ -1 & \cdot & \cdot & \cdot & -1 & k \end{bmatrix},$$

which is of the form $yU + zI$, where y and z are real numbers, I is the identity matrix, and U is the matrix whose elements are all 1's. But a matrix of this form with order k, $z \neq 0$, and $yk + z \neq 0$, has for its inverse the matrix $(-y/z(yk + z))U + (1/z)I$. It follows that the inverse of the coefficient matrix is

$$[(1/k)A'A]^{-1} = (k/(k+1))\begin{bmatrix} 2 & 1 & \cdot & \cdot & \cdot & 1 \\ 1 & 2 & \cdot & & & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & & & \cdot & 2 & 1 \\ 1 & \cdot & \cdot & \cdot & 1 & 2 \end{bmatrix}. \quad (3.6)$$

Substituting (3.3) and (3.6) into the matrix equation (3.2), we thus obtain Theorem 2.

IV. An Example

With $r = 3$ we are assuming that

$$c_i = h_1 m_{i-1}^0 m_{i-2}^0 m_{i-3}^1 + \cdots + h_7 m_{i-1}^1 m_{i-2}^1 m_{i-3}^1.$$

Consider the sequence a_i generated by $x^3 + x + 1$ (mod 2):

$$a_i \equiv -a_{i-1} - a_{i-3} \pmod 2, \qquad a_{-1} = a_{-2} = 0,\ a_{-3} = 1$$

and normalized by

$$m_i' = 1 - 2a_i.$$

i	0	1	2	3	4	5	6
a_i	1	1	1	0	1	0	0
m_i'	−1	−1	−1	1	−1	1	1
m_{i+1}'	−1	−1	1	−1	1	1	−1
m_{i+2}'	−1	1	−1	1	1	−1	−1

Note from the table that

$$\{(m_i', m_{i+1}', m_{i+2}') \mid i = 0, \cdots, 6\} = (1,-1)X(1,-1)X(1,-1) - (1,1,1)$$

Therefore, if $u = (i_1 i_2 i_3)_2$, $v = (j_1 j_2 j_3)_2$, then

$$\sum_{t=0}^{6} H(u,t)H(v,t) = \sum_{t=0}^{6} (m_{t'+2}^{i_1+j_1})(m_{t'+1}^{i_2+j_2})(m_{t'}^{i_3+j_3}) = \prod_{k=1}^{3} (1^{i_k+j_k} + (-1)^{i_k+j_k}) - 1. \quad (4.1)$$

If $u = v$, then $i_k + j_k \equiv 0 \pmod 2$ for all k and (4.1) is 7. If $u \neq v$, then $i_k + j_k \equiv 1 \pmod 2$ for some k and (4.1) is -1. The coefficient matrix and its inverse, respectively, are therefore

$$(1/7)\begin{bmatrix} 7 & -1 & \cdot & \cdot & \cdot & -1 \\ -1 & 7 & \cdot & & & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & & & \cdot & 7 & -1 \\ -1 & \cdot & \cdot & \cdot & -1 & 7 \end{bmatrix} \quad (7/8)\begin{bmatrix} 2 & 1 & \cdot & \cdot & \cdot & 1 \\ 1 & 2 & \cdot & & & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & & & \cdot & 2 & 1 \\ 1 & \cdot & \cdot & \cdot & 1 & 2 \end{bmatrix}.$$

References

[1] A. A. Albert, *Fundamental Concepts of Higher Algebra*. Chicago, Ill.: Univ. Chicago Press, 1956, p. 131.

[2] S. W. Golomb, *Shift Register Sequences*. San Francisco: Holden-Day, 1967, pp. 24–37.

[3] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, 4th ed. New York: Oxford Univ. Press, 1960, pp. 49–71.

[4] J. D. Hill and G. J. McMurtry, "An application of digital computers to linear system identification," *IEEE Trans. Automat. Contr.* (Short Papers), vol. AC-9, pp. 536–538, Oct. 1964.

A Confidence Model for Finite-Memory Learning Systems

JOHN A. HOROS AND MARTIN E. HELLMAN

***Abstract*—A confidence model for finite-memory learning systems is advanced in this correspondence. The primary difference between this and the previously used probability-of-error model is that a measure of confidence is associated with each decision and any incorrect decisions are weighted according to their confidence measure in figuring total loss. The optimal rule for this model is deterministic, whereas the previous model required randomized rules to achieve minimum error probability.**

I. Introduction

Let $X_1, X_2, \cdots$ be a sequence of independent random variables with common probability density function $p(x)$. Under hypothesis $H_0: p(x) = p_0(x)$, while under $H_1: p(x) = p_1(x)$. The *a priori* probabilities π_0 and $\pi_1 = 1 - \pi_0$, as well as $p_0(x)$ and $p_1(x)$, are assumed known. If no constraint is placed on memory, then a standard likelihood ratio test yields a probability of error that tends to zero exponentially in the number of observations.

In [1] Hellman and Cover investigate the above problem under a finite-memory constraint. They define a rule to have memory of size m if the decision made after the nth observation d_n depends on the data only through an m-valued statistic T whose value T_n at time n is a function of T_{n-1} and X_n. Such a rule may be written in the form

$$T_n = f(T_{n-1}, X_n) \in \{1,2,\cdots,m\}$$
$$d_n = d(T_n) \in \{H_0, H_1\}. \quad (1)$$

Letting e_n equal 1 or 0 according to whether the nth decision is in error or not, define

$$P(f,d) = E \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} e_n \quad (2)$$

to be the (asymptotic) probability of error of the algorithm (f,d). In [1] it is shown that

$$P^*(m) \triangleq \inf_{f,d} P(f,d) = \min\left\{ \frac{2(\pi_0\pi_1\gamma^{m-1})^{1/2} - 1}{\gamma^{m-1} - 1},\ \pi_0, \pi_1 \right\}, \quad (3)$$

where γ is a function only of $p_0(x)$ and $p_1(x)$, and the infimum is over all m-state algorithms, both randomized and deterministic. Later work [2]–[7] discusses the differences between randomized and deterministic rules. Letting $P_d^*(m)$ denote the infimum of $P(f,d)$ over all m-state deterministic rules, it has been shown [5] that problems exist for which $P_d^*(m)/P^*(2)$ is arbitrarily large. Thus for the model advanced in [1] there can be a great difference between the performance of randomized and

Manuscript received March 21, 1972; revised May 22, 1972. This work was supported in part by the NSF under Research Initiation Grant GK5800 and in part by the Joint Services Electronics Program under Contract N00014-67-A-0112-0044.

J. A. Horos was with the Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Mass. 02139. He is now with the Bell Telephone Laboratories, Holmdel, N.J. 07733.

M. E. Hellman was with the Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Mass. 02139. He is now with the Department of Electrical Engineering, Stanford University, Stanford, Calif. 94305.

deterministic rules. There has been some question as to whether or not randomized rules require additional memory.

Here we present a model very similar to that just discussed, yet for which randomization is not needed. By so doing we hope to achieve two goals. First, in those practical situations where the new model is applicable, there is no need for randomization. Secondly, the new model demonstrates the incidental manner in which randomization is required by the optimal rules of [1].

In this new model the rule for updating memory is unchanged. However, the output function d now maps the memory state space $S = \{1,2,\cdots,m\}$ into $\{H_0,H_1\} \times R^+$, where R^+ denotes the nonnegative reals. Thus the new model is characterized by algorithms of the form

$$T_n = f(T_{n-1},X_n) \in \{1,2,\cdots,m\}$$
$$d_n = d(T_n) \in \{H_0,H_1\} \times R^+. \tag{4}$$

One interpretation is that if $d_n = (H_0,r_n)$, then r_n units are bet on H_0 being the true hypothesis, whereas if $d_n = (H_1,r_n)$ then r_n units are bet on H_1. Thus r_n is a measure of confidence in the nth decision. We define the cumulative fractional loss up to time N as

$$L_N = \sum_{n=1}^{N} r_n e_n / \sum_{n=1}^{N} r_n, \tag{5}$$

where, as before, $e_n = 1$ or 0 accordingly as the nth choice of hypothesis is in error or correct. Our goal is to find the m-state algorithm that minimizes the expected asymptotic loss

$$L(f,d) = E\{\lim_{N\to\infty} L_N\}. \tag{6}$$

To better understand the reason for considering this new model let us examine the optimal solutions of [1]. Letting

$$l(x) = p_0(x)/p_1(x) \tag{7}$$

denote the likelihood ratio and

$$l_{\max} = \text{ess sup } l(x)$$
$$l_{\min} = \text{ess inf } l(x) \tag{8}$$

then the parameter γ in (3) is

$$\gamma = l_{\max}/l_{\min}. \tag{9}$$

Note that for "nice" problems $l_{\max}$ and $l_{\min}$ are merely the maximum and minimum of $l(x)$. Further let

$$\mathcal{H}_\Delta = \{x: l(x) \geq [(1/l_{\max}) + \Delta]^{-1}\}$$
$$\mathcal{T}_\Delta = \{x: l(x) \leq l_{\min} + \Delta\}. \tag{10}$$

It has been shown that the state transition function with

$$f(i,x) = \begin{cases} i+1, & i \leq m-1 \text{ and } x \in \mathcal{H}_\Delta \\ i-1, & i \geq 2 \text{ and } x \in \mathcal{T}_\Delta \\ i, & \text{otherwise} \end{cases} \tag{11}$$

yields a maximum "spread" [1] between the two hypotheses as $\Delta \to 0$. This is intuitively pleasing since if $x \in \mathcal{H}_\Delta$ then x yields close to maximal evidence in favor of H_0, whereas if $x \in \mathcal{T}_\Delta$ then x yields close to maximal information in favor of H_1. Thus rules that obey (11) throw away all but the best observations, and even on the best observations change memory by only one state. This is a conservative strategy, but for an unlimited sequence of observations it yields an unlimited number of state transitions, which "skim the cream" off the data.

Although this state transition function yields maximal spread it does not yield minimum error probability for two reasons, which are best illustrated by example. Consider a problem where the observations are binary valued, so that

$$p_0(x) = q_0\,\delta(x) + p_0\,\delta(x-1)$$
$$p_1(x) = q_1\,\delta(x) + p_1\,\delta(x-1), \tag{12}$$

where $q_i = 1 - p_i$. Here, as with all discrete distributions, we may take $\Delta = 0$. Thus, for example, if $p_0 > p_1$ then $\mathcal{H}_\Delta = \mathcal{H}_0 = \{X = 1\}$ and $\mathcal{T}_\Delta = \mathcal{T}_0 = \{X = 0\}$. Then the rule specified by (11) moves up one state when $X = 1$ (unless memory is in state m, in which case it stays there), and moves down one state when $X = 0$ (unless memory is in state 1, in which case it stays there).

Now consider a two-state memory to be used in this hypothesis-testing problem. If the problem is symmetrical (i.e., $p_0 = 1 - p_1$ and $\pi_0 = \pi_1$), then using the state transition rule specified by (11) and deciding H_0 in state 2 and H_1 in state 1 results in an optimal rule, $P(f,d) = P^*(2)$. If, on the other hand, $p_0 \neq 1 - p_1$ this rule can be far from optimal. For example, if $p_0 = 10^{-6}$ and $p_1 = 10^{-12}$ then this rule has $P(f,d) \approx \frac{1}{2}$. It moves from state 1 to state 2 on $X = 1$, which occurs with low probability under either hypothesis (10^{-6} and 10^{-12}), while it moves from state 2 to state 1 on $X = 0$, which occurs almost certainly under both hypotheses. Therefore this rule results in a high occupation probability for state 1 and a low occupation probability for state 2 under either hypothesis, and its probability of error is high. By exhaustion all deterministic 2-state rules can be shown to have $P(f,d) \approx \frac{1}{2}$.

Now consider the state transition function

$$f(1,x) = 2, \quad x \in \mathcal{H}_0$$
$$f(2,x) = \begin{cases} 1, & \text{with probability } 10^{-9},\ x \in \mathcal{T}_0 \\ 2, & \text{with probability } 1 - 10^{-9},\ x \in \mathcal{T}_0 \end{cases}$$
$$f(i,x) = i, \quad \text{otherwise.} \tag{13}$$

Rule (11) had a strong drift into state 1 under both hypotheses, but Rule (13) cancels that drift under H_0 and has $P(f,d) \approx P^*(2)$. It is seen that if $\pi_0 \neq \pi_1$ a similar asymmetry results, which also requires randomization. Thus the first reason for using randomization is to offset the effects of asymmetries in the problem.

Now consider a symmetric problem, say $p_0 = \frac{3}{4}$, $p_1 = \frac{1}{4}$, and $\pi_0 = \pi_1 = \frac{1}{2}$. For this problem the optimal 2-state algorithm has the deterministic state transition rule (11). However, using (11) for $m \geq 3$ is not optimal since decisions made in states 1 and m have the least probability of being in error, while decisions made in states near $m/2$ have the highest [1]. Thus, the following modifications to the deterministic state transition rule (11) lower $P(f,d)$:

$$f(1,x) = \begin{cases} 2, & \text{with probability } \delta,\ x \in \mathcal{H}_\Delta \\ 1, & \text{with probability } 1 - \delta,\ x \in \mathcal{H}_\Delta \end{cases}$$
$$f(m,x) = \begin{cases} m-1, & \text{with probability } \delta,\ x \in \mathcal{T}_\Delta \\ m, & \text{with probability } 1 - \delta,\ x \in \mathcal{T}_\Delta. \end{cases} \tag{14}$$

If δ is close to zero then this algorithm stays in states 1 and m a much larger proportion of time than the unmodified algorithm. However, as long as $\delta > 0$ the relative proportion of time spent in state 1 as opposed to state m is unchanged. Thus [1] as $\delta \to 0$ this algorithm has $P(f,d) \to P^*(m)$. Note, however, that when $\delta = 0$, $P(f,d) > P^*(m)$. $P^*(m)$ is an unachievable, but approachable lower bound.

Returning to the general hypothesis-testing problem, if $m > 2$ and asymmetries exist, then both types of randomization are necessary and $P^*(m)$ is approached by state transition rules of the form

$$\begin{aligned} f(i,x) &= i + 1, && 2 \le i \le m - 1 \text{ and } x \in \mathscr{H}_\Delta \\ f(i,x) &= i - 1, && 2 \le i \le m - 1 \text{ and } x \in \mathscr{T}_\Delta \\ f(1,x) &= 2, && \text{with probability } \delta,\ x \in \mathscr{H}_\Delta \\ f(m,x) &= m - 1, && \text{with probability } k\delta,\ x \in \mathscr{T}_\Delta \\ f(i,x) &= i, && \text{otherwise.} \end{aligned} \tag{15}$$

The optimal value of k is given by k^* in [1, eq. (57)].

At this point let us see how the flexibility allowed by the new model may be used to eliminate the need for randomization. First note that if bets of size $1/\delta$ and $1/k\delta$ are made in states 1 and m, respectively, with bets of size one in other states, then the algorithm with the deterministic state transition function (11) has the same $L(f,d)$ as the algorithm that bets unity in all states (i.e., has probability-of-error loss criterion), and has a randomized state transition function of the form (15). But as $\Delta, \delta \to 0$ with $k = k^*$, the second algorithm, and therefore the first, has $L(f,d) \to P^*(m)$. Next, note that the same effect is achieved by betting 1 unit in state 1, δ units in states 2 through $m - 1$, and $1/k$ units in state m.

Now let

$$L^*(m) = \inf_{f,d} L(f,d), \tag{16}$$

where the infimum is over all randomized and deterministic m-state algorithms of the form (4).

From the preceding construction we see that for any $\varepsilon > 0$ there is a deterministic rule for which $L(f,d) \le P^*(m) + \varepsilon$. Thus,

$$L^*(m) \le P^*(m). \tag{17}$$

Therefore, the following theorem establishes the optimality of deterministic rules for the new model.

Theorem: For any m-state algorithm of the form (4)

$$L(f,d) \ge P^*(m), \tag{18}$$

where $P^*(m)$ is given by (3).

Proof: Let $\boldsymbol{\mu}^0$ and $\boldsymbol{\mu}^1$ denote the stationary distributions under H_0 and H_1 on S, the state space for memory. Let S_0 and S_1 denote the sets of states in which H_0 and H_1 are chosen, and let ρ_i be the amount bet on the decision made in state i. Then

$$\alpha = \sum_{i \in S_1} \mu_i^0 \rho_i / \sum_{i \in S} \mu_i^0 \rho_i \tag{19}$$

is the asymptotic fractional loss under H_0 and

$$\beta = \sum_{i \in S_0} \mu_i^1 \rho_i / \sum_{i \in S} \mu_i^1 \rho_i \tag{20}$$

is the asymptotic fractional loss under H_1. Therefore

$$L(f,d) = \pi_0 \alpha + \pi_1 \beta. \tag{21}$$

From [1, theorem 2] we know that for any rule, randomized or deterministic, there exists $0 < c < 1$ such that

$$c \le \mu_i^0 / \mu_i^1 \le c\gamma^{m-1}, \qquad 1 \le i \le m. \tag{22}$$

Letting

$$s^j = \sum_{i \in S} \mu_i^j \rho_i, \qquad j = 0,1 \tag{23}$$

we have

$$\begin{aligned} \alpha &= \sum_{S_1} \rho_i \mu_i^0 / s^0 \ge c(s^1/s^0) \sum_{S_1} \mu_i^1 \rho_i / s^1 \\ &= (cs^1/s^0)(1 - \beta) \end{aligned} \tag{24}$$

and

$$\begin{aligned} \beta &= \sum_{S_0} \rho_i \mu_i^1 / s^1 \ge (s^0/s^1 c\gamma^{m-1}) \sum_{S_0} \mu_i^0 \rho_i / s^0 \\ &= (s^0/s^1 c\gamma^{m-1})(1 - \alpha). \end{aligned} \tag{25}$$

Multiplying (24) and (25) we obtain

$$\alpha\beta / [(1 - \alpha)(1 - \beta)] \ge (1/\gamma^{m-1}). \tag{26}$$

Minimizing $L(f,d)$ as given by (21) subject to the constraint (26) yields a lower bound on $L(f,d)$. This problem is equivalent to that treated in [1, th. 3] and results in the bound

$$L(f,d) \ge P^*(m). \tag{18}$$

Q.E.D.

Combining (17) and (18) yields the desired result

$$L^*(m) = P^*(m). \tag{27}$$

II. Discussion

It is interesting to note that when the observation space is finite, $L^*(m)$ is achieved by setting $\Delta = 0$ and $\delta = 0$. This is in contrast to the original model, where $P^*(m)$ is generally not achievable by any rule unless $m = 2$. Further, note that for symmetric problems $k^* = 1$ and the resultant algorithm is particularly simple, with bets of zero and unity only.

Acknowledgment

The authors wish to thank T. Cover for helpful comments on the organization of this correspondence.

References

[1] M. E. Hellman and T. M. Cover, "Learning with finite memory," *Ann. Math. Statist.*, vol. 41, pp. 765–782, 1970.
[2] B. Chandrasekaran, "Finite-memory hypothesis testing—A critique," *IEEE Trans. Inform. Theory* (Corresp.), vol. IT-16, pp. 494–496, July 1970.
[3] T. M. Cover and M. E. Hellman, "Finite-memory hypothesis testing—Comments on a critique," *IEEE Trans. Inform. Theory* (Corresp.), vol. IT-16, pp. 496–497, July 1970.
[4] B. Chandrasekaran, "Reply to 'Finite memory hypothesis testing—Comments on a critique'," *IEEE Trans. Inform. Theory* (Corresp.), vol. IT-17, pp. 104–105, Jan. 1971.
[5] M. E. Hellman and T. M. Cover, "On memory saved by randomization," *Ann. Math. Statist.*, vol. 42, pp. 1075–1078, 1971.
[6] J. A. Horos, "Deterministic finite memory learning algorithms," S.M. thesis, Dep. Elec. Eng., Massachusetts Inst. Technol., Cambridge, 1971.
[7] M. E. Hellman, "The effects of randomization on finite memory decision schemes," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 499–502, July 1972.