The Effects of Randomization on Finite-Memory Decision Schemes

MARTIN E. HELLMAN, MEMBER, IEEE

Abstract—**This paper is concerned with the differences between deterministic and randomized finite-memory decision rules. It is shown that for any hypothesis-testing problem there exists a $b < \infty$ such that, for any B, the optimal deterministic rule with $B + b$ bits in memory has a lower error probability than the optimal randomized rule with B bits in memory. Suboptimal deterministic rules with this property are demonstrated. These deterministic rules lose at most b bits. Thus for large memories the fraction of memory, measured in bits, lost by deterministic rules is negligible.**

Introduction

Let $X_1,X_2\cdots$ be a sequence of independent identically distributed random variables each having the probability density function $p(x)$ over an observation space $\mathscr{X}$. We wish to test between the hypotheses $\{H_k, 0 \leq k \leq K-1\}$ where under H_k, $p(x) = p_k(x)$. We assume that $\{p_k(x), 0 \leq k \leq K-1\}$ as well as the *a priori* probabilities $\{\pi_k, 0 \leq k \leq K-1\}$ are known.

A finite-memory algorithm for this problem is a triple (S,f,d), where $S = \{1,2,\cdots,m\}$ is an m-state memory, or state space; f: $S \times \mathscr{X} \rightarrow S$ is the state transition function; and d: $S \rightarrow \{H_k, 0 \leq k \leq K-1\}$ is the decision function. The interpretation is that if at time $n-1$ the memory is in state i and $X_n = x$ then the state of memory at time n is $j = f(i,x)$; decision $d(j)$ is then made and the process repeats. Letting T_n denote the state of memory at time n and d_n denote the decision made at time n, we may summarize the operation of the algorithm as

$$\begin{aligned} T_n &= f(T_{n-1},X_n) \in \{1,2,\cdots,m\} \\ d_n &= d(T_n) \in \{H_0,H_1,\cdots,H_{K-1}\}. \end{aligned} \tag{1}$$

In [1] the size of memory was measured by m, the number of states in memory. Another related measure is

$$B = \log_2 m, \tag{2}$$

the number of bits in memory. An algorithm is said to be deterministic if the mappings f and d are deterministic. Otherwise it is said to be randomized.

The asymptotic probability of error of an algorithm is defined to be

$$P(f,d) = E \lim_{N\rightarrow\infty} \frac{1}{N} \sum_{n=1}^{N} e_n, \tag{3}$$

where $e_n = 0$ or 1 accordingly as $d_n = H_t$ or $d_n \neq H_t$, and H_t denotes the true hypothesis. Define

$$P^*(m) = \inf_{f,d} P(f,d), \tag{4}$$

where the infimum is taken over all m-state algorithms, and define

$$P_d^*(m) = \inf_{f,d} P(f,d), \tag{5}$$

where now the infimum is over all m-state deterministic algorithms. Obviously

$$P^*(m) \leq P_d^*(m), \tag{6}$$

and as shown in [1] the strict inequality holds in many important cases. The relative merits of randomization are discussed in [2]–[4]. Newer work [5] shows that randomized algorithms are arbitrarily better, in the sense that for any $m < \infty$ and any $\delta > 0$ there exist problems such that $P^*(2) \leq \delta$ and yet $P_d^*(m) > \frac{1}{2} - \delta$.

In this paper we show that deterministic rules are asymptotically optimal, which in light of the above statement seems paradoxical. The explanation of the seeming paradox is that here asymptotic optimality refers to the fraction of bits lost, whereas in [5] the measure of optimality concerns the number of states. However, taken together, [5] and this paper go a long way towards answering the questions raised in [2]–[4].

Behavior of $P^*(m)$ and $P_d^*(m)$ for Two Hypotheses

Let the likelihood ratio be defined as usual, $l(x) = p_0(x)/p_1(x)$, and let l_u and l_d be the essential supremum and infimum on $l(x)$. (Except for pathological cases l_u and l_d are merely the maximum and minimum values of $l(x)$.) Then as shown in [1]

$$P^*(m) = [2(\pi_0\pi_1\gamma^{m-1})^{1/2} - 1]/(\gamma^{m-1} - 1), \tag{7}$$

where $\gamma = l_u/l_d$. For $\pi_0 = \pi_1 = \frac{1}{2}$, (7) reduces to

$$P^*(m) = 1/(\gamma^{(m-1)/2} + 1). \tag{8}$$

For simplicity we will restrict attention to this case ($\pi_0 = \pi_1$), noting that the extension to $\pi_0 \neq \pi_1$ is straightforward.

For large values of m we have

$$P^*(m) \geq r^m = r^{(2^B)}, \tag{9}$$

Manuscript received July 6, 1971; revised February 28, 1972. This work was supported in part by the National Science Foundation Grant GK-5800 and in part by the Joint Services Electronics Program (U.S. Army, U.S. Navy, and U.S. Air Force) under Contract N00014-67-A-0112-0044. Portions of this paper were presented at the 1970 IEEE Symposium on Adaptive Processes, Austin, Tex., Dec. 7–8, 1970.

The author is with the Department of Electrical Engineering, Stanford University, Stanford, Calif. 94305.

where $r = \gamma^{-1/2} < 1$. Suppose we could find a class of deterministic algorithms with

$$P(f,d) \leq s^m, \tag{10}$$

where $r < s < 1$. Then we could find a $C < \infty$ such that

$$r \geq s^C \tag{11}$$

and hence

$$P_d^*(m) \leq P(f,d) \leq r^{m/C} \leq P^*(m/C). \tag{12}$$

Equation (12) implies that if we have an optimal m-state randomized rule and then consider the optimal deterministic rule with $m' = mC$ states we would be guaranteed of having no higher a probability of error in the second case. Of course, we have lost the possibly large fraction $(C - 1)/C$ of states by using such a deterministic rule. On the other hand, this corresponds to losing $b = \log_2 C$ bits, and since C is independent of m, and hence of B, the fraction of bits lost would tend to zero as $B \to \infty$.

Thus to accomplish our goal, all we must do is find a class of deterministic algorithms for which the probability of error goes to zero exponentially in m.

As a first step, note that we lose no generality by restricting consideration to Bernoulli random variables (i.e., the observation X_n can take on only two different values). This is because we can quantize the observation space as follows. Let $\mathscr{H} = \{x \in \mathscr{X}: l(x) > 1\}$. Then $\Pr(\mathscr{H} \mid H_0) > \Pr(\mathscr{H} \mid H_1)$ unless $p_0(x) = p_1(x)$ almost everywhere.

Now let $\mathscr{T} = \mathscr{X} - \mathscr{H}$. Obviously $\Pr(\mathscr{T} \mid H_0) < \Pr(\mathscr{T} \mid H_1)$. Equating $X \in \mathscr{H}$ with "head," and $X \in \mathscr{T}$ with "tail" we effectively reduce the problem to testing between the hypotheses H_0: Pr (head) $= p_0$, and H_1: Pr (head) $= p_1$ with $p_0 > p_1$ and thus $q_0 = 1 - p_0 < q_1 = 1 - p_1$. If we can show that this latter problem has a probability of error that goes to zero exponentially in m, we will have solved the general two-hypothesis problem.

As an aside, note that a three-level quantization scheme as used in [1] is usually much superior, but for simplicity we use two-level quantization here.

We now prove the following basic theorem.

Theorem: If $p_0 > \frac{1}{2} > p_1$ then the following class of rules (known as saturable counters) has $P(f,d) \leq s^m$ where $s = \max\{(q_0/p_0)^{1/2},(p_1/q_1)^{1/2}\} < 1$:

$$\begin{aligned} f(i, \text{head}) &= i + 1, & 1 \leq i \leq m - 1 \\ f(m, \text{head}) &= m \\ f(i, \text{tail}) &= i - 1, & 2 \leq i \leq m \\ f(1, \text{tail}) &= 1 \\ d(i) &= H_0, & i > m/2 \\ d(i) &= H_1 & i \leq m/2. \end{aligned} \tag{13}$$

Proof: Under either hypothesis the sequence $\{T_n, 0 \leq n < \infty\}$ is a Markov chain. Indeed it is a random walk, and using standard methods [1], [6] we find the long-run probability of occupying state i to be

$$\begin{aligned} \mu_i^0 &= \Pr(i \mid H_0) = a_0 s_0^i \\ \mu_i^1 &= \Pr(i \mid H_1) = a_1 s_1^i, \end{aligned} \tag{14}$$

where $s_0 = (p_0/q_0) > 1$, $s_1 = (p_1/q_1) < 1$ and a_0 and a_1 are chosen so that

$$\sum_{i=1}^{m} \mu_i^0 = \sum_{i=1}^{m} \mu_i^1 = 1.$$

Letting α and β denote the probability of error under H_0 and H_1, and assuming m is even (if m is odd similar results are obtained) we have

$$\alpha = \sum_{i \leq m/2} \mu_i^0 = \frac{s_0^{m/2} - 1}{s_0^m - 1} = \frac{1}{s_0^{m/2} + 1} \tag{15}$$

$$\beta = \sum_{i > m/2} \mu_i^1 = \frac{s_1^{m/2} - s_1^m}{1 - s_1^m} = \frac{s_1^{m/2}}{1 + s_1^{m/2}}. \tag{16}$$

Consequently

$$\alpha \leq s_0^{-m/2} \qquad \beta \leq s_1^{m/2}. \tag{17}$$

For $\pi_0 = \pi_1 = \frac{1}{2}$, we thus have

$$P(f,d) = \tfrac{1}{2}(\alpha + \beta) \leq s^m, \tag{18}$$

where

$$s = \max\{s_0^{-\frac{1}{2}}, s_1^{\frac{1}{2}}\} < 1. \tag{19}$$

Q.E.D.

If $p_0 > p_1 \geq \frac{1}{2}$, or $\frac{1}{2} \geq p_0 > p_1$ the saturable counters used in the above proof do not have a probability of error that goes to zero in m. However, a slight modification yields the desired behavior.

Suppose we change the form of the state transition function f, in that state transitions are made after every N observations, and the new state of memory depends on the old state and the previous N observations, i.e., for $n = 1,2,3,\cdots$

$$\begin{aligned} T_{nN+1} &= f(T_{nN}, X_{nN+1}, X_{nN}, X_{nN-1}, \cdots, X_{nN-N+2}) \\ T_{nN+1} &= T_{nN+2} = \cdots = T_{nN+N}. \end{aligned} \tag{20}$$

A decision is still made after each observation, however. The importance of this formulation is that an m-state algorithm of the form (20) can be implemented by an algorithm of the form (1) using no more than $m(2^N - 1)$ states, provided the X_i can take on only two values. In essence we break the $m(2^N - 1)$ states into m superstates, each consisting of $(2^N - 1)$ substates. When a superstate is entered, the entry is via the first substate. After the next observation a subtransition is made to one of the next two substates depending on whether a head or a tail is observed; after the next observation a subtransition is made to one of the next four states, etc. After the Nth new observation a transition is made from one of the last 2^{N-1} substates to a new superstate.

Now consider the problem where $p_0 > p_1 \geq \frac{1}{2}$ (the problem of $\frac{1}{2} \geq p_0 \geq p_1$ is equivalent to this by interchanging heads and tails). We can always find integers N_1 and N_2 with $N_1 > N_2$, such that

$$\begin{aligned}(p_0^{N_1}/q_0^{N_2}) &> 1\\ (p_1^{N_1}/q_1^{N_2}) &< 1. \qquad (21)\end{aligned}$$

Then let $N = N_1$ and have the algorithm move up one superstate when the block of N observations consists of all heads and the old superstate is not the mth, move down one superstate when the first N_2 observations of the block are all tails and the old superstate is not the first; and reenter the same superstate otherwise. It is seen from (21) that under H_0 there is a drift to higher numbered states, while under H_1 there is a drift to lower numbered states. Again, applying standard methods for finding stationary distributions for Markov chains, we can evaluate α and β for the rule that decides H_0 in superstates $i > m/2$ and H_1 otherwise. The results (for even m) are strikingly similar to (15) and (16):

$$\begin{aligned}\alpha &= 1/(s_0^{m/2} + 1)\\ \beta &= s_1^{m/2}/(1 + s_1^{m/2}), \qquad (22)\end{aligned}$$

where now

$$\begin{aligned}s_0 &= (p_0^{N_1}/q_0^{N_2}) > 1\\ s_1 &= (p_1^{N_1}/q_1^{N_2}) < 1. \qquad (23)\end{aligned}$$

Letting

$$s = [\max\{s_0^{-\frac{1}{2}}, s_1^{\frac{1}{2}}\}]^{1/(2^N - 1)} < 1 \qquad (24)$$

we find that for these algorithms

$$P(f,d) \leq s^m \qquad (25)$$

for m (regular) states in memory.

Of course, we have been somewhat conservative. For example, the algorithms just described require at most $3(N - 1)$, not $(2^N - 1)$ substates per superstate, and it is possible to find even more efficient rules.

Generalization to More than Two Hypotheses

If we consider $K > 2$ then neither $P^*(m)$ nor $P_d^*(m)$ is known explicitly. However, it is clear that $P^*(m)$ cannot go to zero faster than exponentially in m. Otherwise, when $K = 2$, we could pretend there were additional hypotheses and achieve a lower probability of error. Thus if we can exhibit a family of deterministic algorithms that have probability of error going to zero exponentially in m we will have generalized the results of the previous section to include any finite number of hypotheses.

Suppose there were two sets of events $\{\mathscr{H}_k, 0 \leq k \leq K - 1\}$ and $\{\mathscr{T}_k, 0 \leq k \leq K - 1\}$ that for all k and $j \neq k$ satisfy

$$\Pr(\mathscr{H}_k \mid H_k)/\Pr(\mathscr{T}_k \mid H_k) > 1$$

and

$$\Pr(\mathscr{H}_k \mid H_j)/\Pr(\mathscr{T}_k \mid H_j) < 1. \qquad (26)$$

(The existence of such sets will be demonstrated in a subsequent construction.) Consider an $m = MK$ state algorithm with states doubly subscripted [i.e., state (k,i)] where $0 \leq k \leq K - 1$ and $1 \leq i \leq M$. Then let

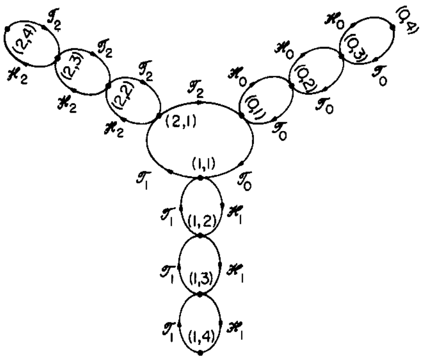

Fig. 1. Deterministic rule for $K = 3$ hypotheses.

$$\begin{aligned}
f((k,i),X) &= (k, i + 1), && i \leq M - 1, X \in \mathscr{H}_k\\
f((k,i),X) &= (k, i - 1), && i \geq 2, X \in \mathscr{T}_k\\
f((k,M),X) &= (k,M), && X \in \mathscr{H}_k\\
f((k,1),X) &= (k + 1, 1), && k \leq K - 2, X \in \mathscr{T}_k\\
f((K - 1, 1),X) &= (0,1), && X \in \mathscr{T}_k\\
f((k,i),X) &= (k,i), && X \notin \mathscr{H}_k U \mathscr{T}_k\\
d((k,i)) &= H_k.
\end{aligned}$$

This algorithm has a star-shaped state transition diagram, with each leg of the star representing one hypothesis. If the state of memory is in the kth leg (i.e., $T_n = (k,\cdot)$, $d_n = H_k$) and H_k is observed, the algorithm moves out one state further in the leg (if possible); if $\mathscr{T}_k$ is observed it moves in one state toward the center, unless it is as close as possible (i.e., in state $(k,1)$), in which case it shifts to the first state in the next leg. Fig. 1 shows the case for $K = 3$ and $M = 4$ with self-loops deleted. It is seen that under hypothesis H_k the stationary distribution has an exponential increase along leg k and an exponential decrease along all other legs. Thus, under $\mathscr{H}_k$, the probability of not being in leg k tends exponentially to zero, as does the probability of error.

In general, it is not possible to find $\{\mathscr{H}_k\}$ and $\{\mathscr{T}_k\}$ with the desired properties (26). However, by considering transitions that depend on blocks of N observations we can guarantee the existence of such sets. For N large but finite, the relative frequencies of the various possible outcomes must, by the law of large numbers, be close to their true probabilities. Thus by making N large enough and letting $\mathscr{H}_k = \{\text{typical } N\text{-sequences under } H_k\}$, $\mathscr{T}_k = \mathscr{X}^N - \mathscr{H}_k$ we can satisfy (26). If $\mathscr{X}$ is a finite observation space each superstate will, as in the last section, consist of a finite number (independent of m) of substates. If the cardinality of $\mathscr{X}$ is not finite, a quantization scheme can be employed. In either case the exponential decay of error probability is unaffected.

By way of example suppose we have three hypotheses concerning the bias of a coin. Under H_0, $P(\text{head}) = 0.75$; under H_1, $P(\text{head}) = 0.4$; and under H_2, $P(\text{head}) = 0.2$.

By letting $N = 100$, $\mathscr{H}_1 = \{\text{number of heads} \geq 60\}$, $\mathscr{H}_2 = \{\text{number of heads is between 30 and 60}\}$, and $\mathscr{H}_3 = \{\text{number of heads} \leq 30\}$ we obtain the desired relations. The extension to non-Bernoulli problems is obvious.

Structure of the Optimal Algorithm

We have not attempted to derive an optimal algorithm. However it is possible to show that for a finite observation space $\mathscr{X}$ and large memory size the optimal deterministic algorithm must have at least a partially star-shaped state transition diagram. In particular, for all k, no transition can occur from the state i_k that has maximum probability of occupation under H_k to a state i for which $d(i) \neq H_k$. Otherwise the probability of error under H_k would be at least p/m, where p is the minimum probability of an observation under H_k and m is the number of states in memory. This follows since the probability under H_k of occupying state i_k obeys

$$\mu_{i_k}{}^k \geq 1/m.$$

However, we know that the probability of error goes to zero exponentially in m, implying that for large m and $p > 0$ such a transition cannot be included in the optimal machine. If $p = 0$ we have a degenerate situation and such transitions are allowable, although they never occur.

If the Markov chain is irreducible, then the state j_k having the second highest probability of occupation under H_k obeys

$$\mu_{j_k}{}^k \geq p/m.$$

Thus if there were a transition from j_k to a state i for which $d(i) \neq H_k$ the probability of error under H_k would be at least p^2/m, which again precludes an exponential decrease in m.

Proceeding in a like manner, we see that for an irreducible chain, more and more transitions are ruled out as m becomes large. This implies a star-like structure closely resembling Fig. 1. Star-like structures have also been investigated by Tsetlin [7].

Discussion

One of the main objections to randomized rules has been that general-purpose digital computers have no natural source of randomization available and, if they are to implement randomized rules, additional memory would be required. However, in such a machine the cost of memory is best measured in bits, so that the optimality measure employed in this paper is applicable, and the theory of optimal randomized algorithms gives asymptotically tight bounds on the relative size of memory required to achieve a desired level of performance. Further, it is interesting that the machines proposed do not use a portion of memory to generate pseudorandom numbers. Indeed, it seems that such a division of memory is far from optimal.

References

[1] M. E. Hellman and T. M. Cover, "Learning with finite memory," *Ann. Math. Statist.*, vol. 41, pp. 765–782, June 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] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 1. New York: Wiley, 1968.

[7] M. L. Tsetlin, "On the behavior of finite automata in random media," *Automat. Telemekh.*, vol. 22, pp. 1345–1354, 1961.