AMRL-TR-67-10

COMPOUND DECISION PROCEDURES FOR PATTERN CLASSIFICATION

KENNETH ABEND

Distribution of this document is unlimited. It may be released to the Clearinghouse, Department of Commerce, for sale to the general public.

FOREWORD

This report was prepared at the Philco-Ford Corporation, System Sciences Laboratory, Blue Bell, Pennsylvania, under Contract AF 33(615)-2966 for the Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command, Wright-Patterson Air Force Base, Ohio, under Project 7233, "Biological Information Handling Systems and Their Functional Analogs," and Task 723305, "Theory of Information Handling." This research was initiated 1 August 1965 and completed 30 April 1967. The technical contract monitor was Hans L. Oestreicher, PhD, Chief, Mathematics and Analysis Branch, Biodynamics and Bionics Division, Biomedical Laboratory, Aerospace Medical Research Laboratories.

The material represents a dissertation in electrical engineering for the degree of Doctor of Philosophy at the University of Pennsylvania. Portions of this work appear in the Proceedings of the 1966 National Electronics Conference and the Proceedings of the 1966 Institute of Electrical and Electronics Engineers Pattern Recognition Workshop.

This technical report has been reviewed and is approved.

WAYNE H. McCANDLESS
Technical Director
Biomedical Laboratory
Aerospace Medical Research Laboratories

ABSTRACT

Compound decision theory is shown to be powerful as a general theoretical framework for pattern recognition, leading to nonparametric methods, methods of threshold adjustment, and methods for taking context into account. The finite-sample-size performance of the Fix-Hodges nearest-neighbor nonparametric classification procedure is derived for independent binary patterns. Classification of binary patterns based on Markov-chain assumptions that account for dependence of a variable on a set of spatial neighbors is shown to require the estimation of a much smaller number of parameters than the general case. The most general nonparametric pattern-recognition problem of "learning with a teacher" followed by "unsupervised" updating is formulated as a distribution-free compound decision problem. Adaptive threshold adjustment procedures for two classes with known distributions but unknown a priori probabilities, are presented. The optimum (Bayes) sequential compound decision procedure, for known distributions and dependent states of nature is derived. When the states of nature form a Markov chain, the procedure is recursive, easily implemented, and immediately applicable to the use of context. A similar procedure, in which a decision depends on previous observations only through the decision about the preceeding state of nature, can (when the populations are not well separated) yield results significantly worse than a procedure that does not depend on previous observations at all. When the populations are well separated, however, an improvement almost equal to that of the optimum sequential rule is achieved. Further improvement is available through the use of nonsequential compound rules. These results are illustrated by error-probability curves for the case of normal densities.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF SYMBOLS USED IN FIGURES 3 to 12

e_1 = error probability for the simple Bayes rule.

e_k = k^{th} component error probability for the sequential compound Bayes rule.

$e_k^{\dagger}$ = k^{th} component error probability for the "decision-directed" rule.

$e^{\dagger}$ = limit of $e_k^{\dagger}$.

e_c = minimum error probability for a sequential compound rule
= error probability for the "decision-directed" rule when the preceeding decision is correct.

e_e = error probability for the "decision-directed" rule when the preceeding decision is in error.

e_c^{*} = minimum error probability for a non-sequential compound rule.

Section I

INTRODUCTION

1.1 PATTERN CLASSIFICATION

Classical decision theory [35, 3] , which has been successfully applied to problems in communication theory and in pattern recognition ever since 1954 [30] and 1957 [4] , is applicable only when a single decision must be made. When the same decision problem occurs many times, advantages might be gained by considering the whole collection of problems as a totality. Compound decision theory, introduced by Hannan and Robbins [21, 10] is powerful as a general theoretical framework into which to imbed pattern recognition. It can lead to methods of threshold adjustment, to methods for taking context into account, and to nonparametric methods.

In most pattern-recognition systems [15, 16, 19, 2, 7] , a set of measurements characterizing a pattern is used to classify the pattern into one of a finite number of categories. We assume that the patterns which are to be classified can be represented by a finite set of properties called observables. The values of these n observables are real numbers, called measurements.

The identification of pattern classification with statistical decision theory is made as follows. We identify a pattern with the outcome of an experiment (a point in a sample space), and the set of observables with a vector-valued function defined on the sample space (a random vector). Thus, the ordered set of measurements $(x_1, x_2, \ldots, x_n)$, called a pattern vector, is identified with the value x of a vector-valued random variable X.

We shall not treat the problem of choosing the observables, but only the problem of classifying the pattern vector.

Let S be the (induced) sample space of values x of the (vector-valued) random variable X. Let the parameter space Ω be an index set for the probability distributions P_ω on S. For a fixed $\omega \in \Omega$, let $p_\omega(x) = p(x/\omega)$ be a (multivariate) probability density* on S. The elements of Ω are called the states of nature [3].

The decision maker has at his disposal a set A of possible actions and suffers a loss $L(\omega, a) \geq 0$ if he takes action $a \in A$ when nature is in state $\omega \in \Omega$. For example, if Ω is the set of letters in the English alphabet, action a_0 may be to decide that ω is a vowel and $L(\omega, a_0) = \begin{cases} 1 \text{ if } \omega \text{ is a consonant} \\ 0 \text{ if } \omega \text{ is a vowel} \end{cases}$. If Ω and A are finite, we write i for ω and j for a, and the loss function L becomes a rectangular matrix with elements L_{ij}. We shall identify the states of nature with the classes i = 1, 2, ..., r from which the pattern vector x may have come.

The identifications we have made are summarized in the following table:

* If $p_\omega(x)$ is a discrete probability distribution instead of a density, all (multiple) integrals with respect to x become sums.

Pattern Classification	Statistical Decision Theory	Symbol
1. Pattern	Outcome	
Pattern space	Sample space	
2. Set of observables	Vector random variable	$X(=X_1,\dots,X_n)$
3. Pattern vector	Value of random variable	$x=(x_1,\dots,x_n)$
Measurement space	Induced sample space	$S = \{x\}$
4. Class	State of nature	ω or i
	Parameter space	$\Omega=\{\omega\}=\{1,2,\dots,r\}$
5. Classification	Action	a or j
	Action space	$A=\{a\}=\{1,2,\dots,s\}$

We shall only draw a distinction between items 1, 2, and 3 when there is danger of confusion. Hence we shall generally write x instead of X and call it a pattern. We shall sometimes refer to S as the pattern space.

In a compound decision problem we have a vector $\underline{\theta}_N = (\theta_1,\dots,\theta_N)$ of states of nature (classes) and a corresponding vector $\underline{x}_N = (x_1,\dots,x_N)$ of (n-dimensional) random variables (patterns). In the k^{th} component problem, we assume that for a given $\theta_k \epsilon \Omega$, x_k is independent of the other x's and θ's and we denote the probability density function of x_k by $p_{\theta_k}(x_k)$. Since a pattern classifier (a machine that classifies pattern vectors) is not expected to be used only once, we identify the problem of pattern classification with the compound decision problem having a finite loss matrix.

1.2 THE SIMPLE DECISION PROBLEM

A finite statistical decision problem involves a set of states of nature $\Omega = \{1, 2, \ldots, r\}$, and a set of actions $A = \{1, 2, \ldots, s\}$. For every $i \epsilon \Omega$, $j \epsilon A$, the element L_{ij} of the $r \times s$ loss matrix denotes the loss incurred by action j when the state of nature is i. The action chosen depends upon the value $x \epsilon S$ of an observable random variable, and we assume that the distribution of x is $P_i(x)$ when the true state of nature is $i \epsilon \Omega$. A randomized decision function $t(j/x)$ is for each x a distribution over A; $t(j/x)$ is the probability with which action j is selected when x is observed. If for every x, $t(j/x) = 1$ for one particular action $j = j(x)$, t is said to be non-randomized. The risk function $R(i, t)$ is the expected loss incurred by the use of t,

$$R(i, t) = \sum_{j=1}^{s} \int_S L_{ij} t(j/x)\, p_i(x)\, dx\ . \tag{1}$$

Let $q(i) = q_i$ be an a priori probability distribution on Ω. The average risk, or Bayes risk of t with respect to q is

$$\overline{R}(q, t) = \sum_{i=1}^{r} R(i, t)\, q_i = \int_S \sum_{j=1}^{s} \sum_{i=1}^{r} L_{ij}\, t(j/x)\, p_i(x)\, q_i\, dx \tag{2}$$

A decision rule t^q that minimizes the Bayes risk $\overline{R}(q, t)$ is said to be Bayes against q. Thus, a Bayes procedure t^q has $t^q(j/x) = 1$ for that j which minimizes the quantity

$$\sum_{i=1}^{r} L_{ij}\, p_i(x)\, q_i\ . \tag{3}$$

The function

$$R(q) = \min_{t} \overline{R}(q, t) = \overline{R}(q, t^{q}) \qquad (4)$$

is called the Bayes envelope. It is obtainable only if q is known.

The a posteriori probability q(i/x), is the conditional probability of i, given x:

$$q(i/x) = \frac{p_i(x)\, q_i}{p(x)} \quad \frac{p_i(x)\, q_i}{\sum\limits_{\omega \in \Omega} p_{\omega}(x)\, q_{\omega}} \,. \qquad (5)$$

Since the denominator is independent of i, the Bayes procedure is equivalent to choosing the action j which minimizes

$$\sum_{i=1}^{r} L_{ij}\, q(i/x) \,. \qquad (3a)$$

For the special case where $A = \Omega$, action j corresponds to deciding that the state of nature is j, and $L_{ij} = \begin{cases} 1 \text{ if } i \neq j \\ 0 \text{ if } i = j \end{cases}$, the Bayes risk is equal to the probability of error and the Bayes rule consists of choosing the i that maximizes either q(i/x) or the product $p(x, i) = p_i(x)\, q_i$.

Example: Consider the classical problem of testing a simple hypothesis against a simple alternative; nature is in one of two possible states, denoted by $\theta = 0$ and $\theta = 1$. Let the probability density of the observed pattern be $p_{\theta}(x)$ under the hypothesis H_{θ}. Let L_{ij} be the loss incurred if H_j is chosen when H_i is true $(i, j = 0, 1)$ let $q_1 = q$ and $q_0 = 1-q$ be the a priori probabilities of H_1 and H_0 respectively, and let $w = (L_{10} - L_{11})/(L_{01} - L_{00})$. Let t be a randomized strategy: t(1/x) denotes the probability of chosing H_1, given x, and $t(0/x) = 1 - t(1/x)$.

By (3), the Bayes risk is minimized by choosing

$$t^q(1/x) = \begin{cases} 1 & \text{if } p_1(x)/p_0(x) > (1-q)/wq \\ 0 & \text{if } \qquad\qquad\qquad < \\ \gamma & \text{if } \qquad\qquad\qquad = \end{cases} \qquad (6)$$

where γ $(0 \leq \gamma \leq 1)$ is arbitrary. In order to specify a unique, non-randomized rule, we shall set $\gamma = 0$. We note that $t^q(1/x)$ can be considered as the characteristic function of the set

$$T = T(q) = \left\{ x\colon \frac{p_1(x)}{p_0(x)} > \frac{1-q}{wq} \right\}. \qquad (7)$$

Thus the Bayes rule is to chose H_1 whenever $x \in T$, i.e., whenever the likelihood ratio $p_1(x)/p_0(x)$ exceeds the threshold $(1-q)/wq$.

Considering the simplified loss matrix $(w > 0)$:

		Action	
		1	0
State of	1	0	w
Nature	0	1	0

the minimum risk (the Bayes envelope) $\overline{R}(q, t^q)$ is the continus concave function

$$R(q) = qw\left[1 - P_1(T)\right] + (1-q)\, P_0(T) \qquad (8)$$

shown in figure 1 ($P_1(T) \equiv \int_T p_1(x)\, dx$). If $T' = T(q')$ is designed against q' when in fact q is the true a priori probability for state 1, the average risk is

$$\bar{R}(q, t^{q'}) = qw \left[1 - P_1(T')\right] + (1 - q)\, P_0(T'), \tag{9}$$

which is linear in q and tangent to R(q) at q = q'. Thus, if q is unknown, the minimum risk cannot generally be achieved. Within the confines of classical decision theory, however, we can minimize the maximum risk by a "minimax" rule t_o, for which $R(1, t_o) = R(0, t_o) = \max\limits_{q} R(q)$. This "safe" procedure $t_o = t^{q_o}$ designs $T_o = T(q_o)$ against the maximum point q_o so that the average risk is constant:

$$\bar{R}(q, t^{q_o}) = w \left[1 - P_1(T_o)\right] = P_o(T_o) \quad . \tag{10}$$

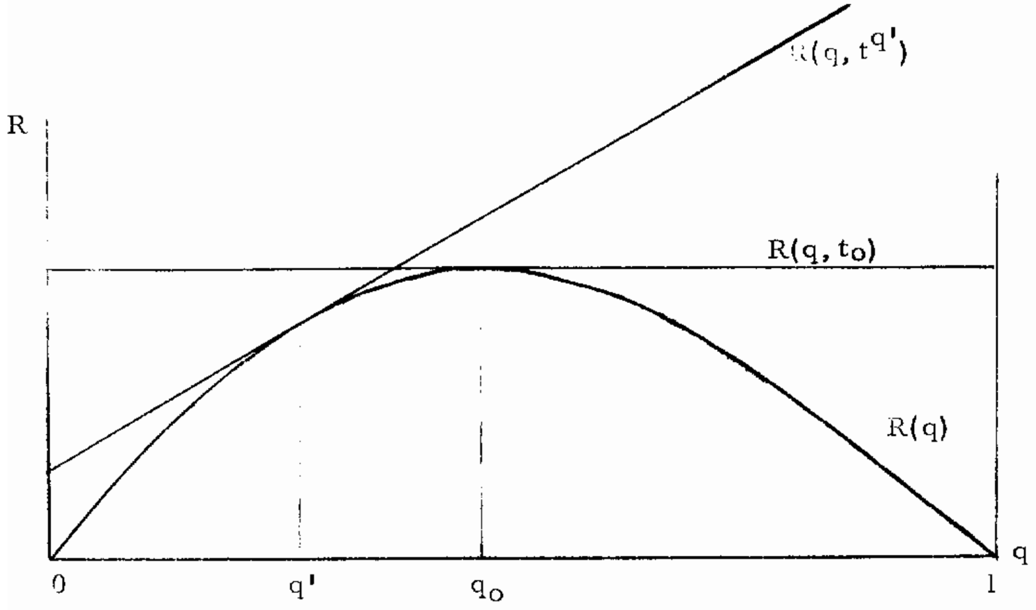

Figure 1 Average Risks and the Bayes Envelope

1.3 THE COMPOUND DECISION PROBLEM

A compound decision problem arises when one is confronted with the same decision problem, called the component problem, not only once, but N times. Thus, there exists a vector $\underline{\theta}_N = (\theta_1, \ldots, \theta_N)$ of states of nature, and a corresponding vector $\underline{x}_N = (x_1, \ldots, x_N)$ of random variables, where θ_k denotes the state in the k^{th} problem, and the distribution of x_k is $P_{\theta_k}(x_k)$. For a given θ_k, x_k is independent of the other x's and θ's:

$$p(x_k/\underline{x}_{k-1}, x_{k+1}, \ldots, x_N, \underline{\theta}_N) = p(x_k/\theta_k) = p_{\theta_k}(x_k) \qquad (11)$$

and hence $p(\underline{x}_k/\underline{\theta}_k) = \prod_{j=1}^{k} p(x_j/\theta_j)$. We do not assume that the θ's are necessarily independent.

The loss in the compound decision problem is taken to be the average of the losses incurred at each of the N decisions, and the compound risk is defined correspondingly. If all observations $\underline{x}_N$ are at hand before the individual decisions must be made, one can use a *compound decision rule* $\underline{t}_N = (t_1, \ldots, t_N)$, where $t_k = (j/\underline{x}_N)$ is for each $\underline{x}_N$ a distribution over A according to which the k^{th} action is chosen. If only the observations $\underline{x}_k$ are at hand when the k^{th} decision must be made, one can use a *sequential* compound decision rule, where $t_k = t_k(j/\underline{x}_k)$. A *simple* rule is one where $t_k = t_k(j/x_k)$, that is, one where the decision about θ_k depends on x_k alone. For a *simple symmetric* rule, $t_k = t(j/x_k)$ for all k. Classical decision theory is restricted to using only simple symmetric rules. The risk for the compound rule $\underline{t}_N$ is

$$R(\underline{\theta}_N, \underline{t}_N) = \frac{1}{N} \sum_{k=1}^{N} R(\underline{\theta}_N, t_k) \tag{12}$$

where

$$R(\underline{\theta}_N, t_k) = \int_{S^N} \sum_{j=1}^{s} L_{\theta_k j}\, t_k(j/\underline{x}_N)\, p(\underline{x}_N/\underline{\theta}_N)\, dx^N \tag{13}$$

is the risk for the k^{th} problem (the k^{th} component risk). The integration is over the N-fold Cartesian product of the measurement space, and

$$p(\underline{x}_N/\underline{\theta}_N) = \prod_{k=1}^{N} p_{\theta_k}(x_k) \quad . \tag{14}$$

We assume that $p_i(x)$ is known for $i = 1, 2, \ldots, r$; but that none of the θ_k's are known.

One can also talk about a compound Bayes risk $\overline{R}(G, \underline{t}_N)$ with respect to an a priori distribution $G(\underline{\theta}_N)$ over Ω^N, the N-fold Cartesian product of Ω. The compound Bayes risk is

$$\overline{R}(G, \underline{t}_N) = \sum_{\underline{\theta}_N \in \Omega^N} R(\underline{\theta}_N, \underline{t}_N)\, G(\underline{\theta}_N) = \frac{1}{N} \sum_{k=1}^{N} \overline{R}(G, t_k) \tag{15}$$

where

$$\overline{R}(G, t_k) = \sum_{\underline{\theta}_N \in \Omega^N} R(\underline{\theta}_N, t_k)\, G(\underline{\theta}_N) \quad , \tag{16}$$

is the k^{th} component Bayes risk. A procedure is compound Bayes against G if it minimizes $\overline{R}(G, \underline{t}_N)$. Thus, the compound Bayes procedure $\underline{t}_N{}^G$ is one that, for every k, minimizes

$$\bar{R}(G, t_k) = \int \sum_{j=1}^{s} \sum_{\underline{\theta}_N} L_{\theta_k j}\, t_k(j/\underline{x}_N)\, p(\underline{x}_N/\underline{\theta}_N)\, G(\underline{\theta}_N)\, dx^N. \tag{17}$$

Hence $t_k{}^G(j/\underline{x}_N) = 1$ for that j which minimizes the quantity

$$\sum_{\underline{\theta}_N} L_{\theta_k j}\, p(\underline{x}_N, \underline{\theta}_N) = \sum_{\theta_k} L_{\theta_k j}\ p(\underline{x}_N, \theta_k) \tag{18}$$

where $p(\underline{x}_N, \underline{\theta}_N) = p(\underline{x}_N/\underline{\theta}_N)G(\underline{\theta}_N)$. For the special case where action j corresponds to deciding that $\theta = j$, and $L_{\theta j} = \begin{cases} 1 \text{ if } \theta \neq j \\ 0 \text{ if } \theta = j \end{cases}$, $t_k{}^G$ choses the value of θ_k that maximizes $p(\underline{x}_N, \theta_k)$. This is equivalent to maximizing the a posteriori probability

$$G(\theta_k/\underline{x}_N) = \frac{p(\underline{x}_N, \theta_k)}{p(\underline{x}_N)}$$

Since the denominator $p(\underline{x}_N) = \sum_{\underline{\theta}_N} p(\underline{x}_N/\underline{\theta}_N)\, G(\underline{\theta}_N)$ is independent of θ_k.

When $G(\underline{\theta}_N) = \prod_{k=1}^{N} q_k(\theta_k)$, i.e., when the states of nature are independently distributed as $q_k(i)$, a simple rule [14] will be compound Bayes against G. A simple symmetric rule will be compound Bayes against G when the states of nature are identically and independently distributed, i.e. $q_k(i) = q(i)$ independent of k, and $G(\underline{\theta}_N) = \prod_{k=1}^{N} q(\theta_k)$. However, even in these cases, non simple compound rules have merit because q(i) may not be known. In the case when the states of nature are identically and independently distributed according to an unknown a priori distribution q(i), one may use an Empirical Bayes decision procedure [22] whereby one employs a "simple" procedure which is

Bayes against a consistent estimate of q. Since such an estimate of q is based upon observations associated with the component problems, such a procedure is really compound.

Thus, in the case of known distributions, there are two distinct situations in which compound decision rules are needed. First, when the states of nature are not independent (e.g., when context may be halpful as in recognizing characters in English text). Second, when an a priori distribution is not known (or does not exist). We shall treat the second situation first.

1.4 ASYMPTOTIC SOLUTIONS

For a simple symmetric rule with $t_k = t(j/x_k)$ Equation (12) becomes

$$R(\underline{\theta}_N, \underline{t}_N) = \frac{1}{N} \sum_{k=1}^{N} R(\theta_k, t) = \sum_{i=1}^{r} q^N(i)\, R(i, t) = \overline{R}(q^N, t), \qquad (19)$$

where $R(\cdot, t)$ and $\overline{R}(\cdot, t)$ are defined by Equations (1) and (2) and $q^N(i)$ is the fraction of the N θ's that are equal to i. Thus, the simple symmetric procedure which is Bayes against the empirical a priori distribution $q^N(i)$, would minimize the compound risk $R(\underline{\theta}_N, \underline{t}_N)$. If we knew q^N in advance, we could, by using a simple symmetric rule $t^{q^N}(j/x_k)$ that chooses action j to minimize the quantity

$$\sum_{i=1}^{r} L_{ij}\, p_i(x_k)\, q^N(i) \ , \qquad (20)$$

obtain the Bayes envelope $R(q^N)$. But q^N is not known. To escape this predicament, we use the observations $\underline{x}_N$ to obtain an estimate $\hat{q}_N$ of q^N and use the compound rule $t_k(j/\underline{x}_N) = t^{\hat{q}_N}(j/x_k)$ to choose the k^{th} action.

For known densities, but unknown G, we may take one of two approaches. The empirical Bayes approach of Robbins [22, 26, 23] assumes that the θ's are independent and identically distributed according to an a priori distribution $q(\theta)$ over Ω and examines convergence of the N^{th} component Bayes risk, $\bar{R}(G, t_N)$, to $R(q)$. A procedure Bayes against a consistent estimate of the a priori distribution $q(\theta)$ will be asymptotically optimum in this sense. The compound approach of Robbins and Hannan [21, 10], is to consider the uniform convergence of the compound risk, $R(\underline{\theta}_N, \underline{t}_N)$, to $R(q^N)$ for any sequence $\underline{\theta}_N$. Hannan and Robbins [10, 24] have shown that the rule $\hat{\underline{t}}_N = (\hat{t}_1, \ldots, \hat{t}_N)$, with

$$\hat{t}_k = t^{\hat{q}_N(\underline{x}_N)}(j/x_k) \qquad (21)$$

Bayes against the consistent estimate of q^N given in Equation (25) below, is "optimal in the limit", i.e., given any $\epsilon > 0$, there exists N_ϵ such that for all $N > N_\epsilon$

$$R(\underline{\theta}_N, \hat{\underline{t}}_N) - R(q^N) < \epsilon \qquad (22)$$

uniformly for all $\underline{\theta}_N \in \Omega^N$, where $R(q)$ is the Bayes envelope. Compound rules that satisfy Equation (22) are said to be asymptotically subminimax.

Van Ryzin [31, 11] has shown that the rate of convergence of $R(\underline{\theta}_N, \underline{\hat{t}}_N)$ to $R(q^N)$ is at least of the order $\sqrt{N}$; for there exists a constant c, independent of $\underline{\theta}_N$ and of N such that

$$R(\underline{\theta}_N, \underline{t}_N) - R(q^N) \le cN^{-1/2} . \qquad (23)$$

With restrictions on $p_i(x)$, even faster rates of convergence are obtainable.

If only the first k observations are at hand when the kth decision must be made, a sequential compound decision rule, $\underline{t}_N^*$, introduced by Samuel [27, 24, 28, 29] is used, where $t_k^*(j/\underline{x}_k) = t^{\hat{q}_{k-1}}(j/x_k)$. This rule is also optimal in the limit, and there even exists a constant c independent of $\underline{\theta}_N$ and of N such that Equation (23) is satisfied [32]. Since $t_k^*(j/\underline{x}_k)$ does not depend on N, the sequential compound decision function $\underline{t}_N^* = (t_1^*, \ldots, t_N^*)$ may be used without knowing N in advance.

Let us return to case of the 2 x 2 loss matrix, the problem is to decide for each $k = 1, \ldots, N$ whether $\theta_k = 0$ or 1.

Let h(x) be a bounded unbiased estimator of θ and for any $\underline{x}_k$, $k > 0$, let

$$h_k(\underline{x}_k) = \frac{1}{k} \sum_{i=1}^{k} h(x_i) \quad . \qquad (24)$$

Letting

$$\hat{q}_k(\underline{x}_k) = \begin{cases} 0 & \text{if } h_k < 0 \\ h_k & \text{if } 0 \le h_k \le 1 \\ 1 & \text{if } h_k > 1 \end{cases} \tag{25}$$

be the "truncated" estimate of

$$q^k = \bar{\theta}_k = \frac{1}{k} \sum_{i=1}^{k} \theta_i \ , \tag{26}$$

the rule

$$\hat{t}_k(1/\underline{x}_N) = \begin{cases} 1 \text{ for } \dfrac{p_1\,(x_k)}{p_0\,(x_k)} > \dfrac{1 - \hat{q}_N(\underline{x}_N)}{w\hat{q}_N(\underline{x}_N)} \\ \\ 0 \text{ otherwise} \end{cases} \tag{27}$$

satisfies Equation (22) [10, 24] and Equation (23) with the constant c depending on the second and third absolute central moments of h and on the value w, but independent of $\underline{\theta}_N$ and of N [11] .

In the sequential case we use at the kth decision rule [27, 28]

$$t_k^*(1/\underline{x}_k) = t^{\hat{q}_{k-1}(\underline{x}_{k-1})}\,(1/x_k) = \begin{cases} 1 \quad \text{for } \dfrac{p_1\,(x_k)}{p_0\,(x_k)} > \dfrac{1 - \hat{q}_{k-1}}{w\,\hat{q}_{k-1}} \\ \\ 0 \quad \text{otherwise} \end{cases} \tag{28}$$

with $\hat{q}_0 = 1/2$.

Using a constant threshold (corresponding to q'), the compound risk as a function of $\bar{\theta}_N$ lies along a straight line $R(\bar{\theta}_N, t^{q'})$ as in

Figure 1 (see Equation (19)). With the procedure described above $R(\underline{\theta}_N, \underline{t}_N)$ approaches the Bayes envelope $R(\overline{\theta}_N) \leq R(\overline{\theta}_N, t^{q'})$, for any sequence $\underline{\theta}_N$ (regardless of whether or not the θ_k's are independent).

1.5 SCOPE OF THIS WORK

In this section we identified pattern classification with statistical decision theory (Section 1.1), described the simple and compound decision problems (Sections 1.2 and 1.3), and summarized some asymptotic solutions to the compound decision problem that have appeared in the statistical literature (Section 1.4).

A simple Bayes solution to the compound problem cannot be found if either a) the probability densities $p_i(x)$ are unknown, b) the a priori distribution $G(\underline{\theta}_N)$ is unknown (or does not exist), or c) the states of nature are not independent. These situations are treated in Sections II, III, and IV, respectively.

In Section II the probability distributions are unknown. After discussing "nonparametric" or distribution-free classification procedures (Section 2.1), we investigate the finite sample-size performance of the simplest version of such a procedure applied to an apparently easy problem (Section 2.2). We then show how a Markov-chain assumption, resulting in a classification function with a small number of parameters, can be used to account for spatial dependence in physical patterns (Section 2.3). In Section 2.4 we extend the formalism of Section I by formulating the distribution-free compound decision problem.

In Section IIIthe densities are known but the a priori distribution is unknown. The results summarized in Section 1.4 specify a class of procedures that are asymptotically subminimax. Restricting the discussion to the case of hypothesis testing, specific procedures for adjusting the decision threshold on the likelihood ratio are presented in Sections 3.1 and 3.2.

In Section IV both the densities and the a priori distribution are known, but $G(\underline{\theta}_N) \neq \prod_{k=1}^{N} q_k(\theta_k)$. Compound decision procedures for dependent states of nature "take context into account". Section 4.1 presents a heuristic discussion of the use of context in print reading. In Section 4.2, the optimum (Bayes) sequential compound procedure is derived. For the case of Markov-chain dependence between consecutive states of nature, this procedure can be easily implemented. In Section 4.3, we analyze error probabilities for this rule and for a sub-optimum "decision-directed" rule suggested in Section 4.1. Section 4.4 extends the analysis to non-sequential compound rules.

In Section V we assume a two-class problem with normal densities and actually calculate and graph the error probabilities analyzed in Section IV. Section VI is a list of conclusions.

When specific procedures lend themselves primarily to particular applications, these are pointed out (Sections 2.3 and 4.1). In general, the application is pattern recognition, which includes character recognition, speech recognition, speaker identification, photo interpretation, medical diagnosis, signal detection, weather prediction, etc.

Section II

NONPARAMETRIC CLASSIFICATION

2.1 THE FIX-HODGES PROCEDURE AND WINDOW CARPENTRY

When even the functional forms of the underlying probability density functions are not known, techniques based only on observations of patterns of known classification must be used. In this section, distribution-free techniques for the direct estimation of the values of the density functions used in a likelihood ratio or Bayes procedure are described.

For the case of two classes, the nonparametric (or distribution free) classification problem is often stated as follows:

> Given a sample of size N_1 from a (n-variate) distribution P_1, a sample of size N_2 from a (n-variate) distribution P_2, and a single observation x either from P_1 or P_2; decide from which distribution x came, when neither P_1, nor P_2, nor even their parametric form is known.

> In 1951, Fix and Hodges [8] presented the following procedure: Choose K, a positive integer which is large,but small compared to the sample sizes. Specify a metric in the sample space, for example, the ordinary Euclidean distance. Pool the two samples and count, of the K values in the pooled samples that are nearest to x, those that are from P_1; call this Q_1. Let $Q_2 = K-Q_1$ be the number that are from P_2. Proceed with likelihood ratio dis-

crimination using however Q_1/N_1 in place of $p_1(x)$ and Q_2/N_2 in place of $p_2(x)$. That is, assign x to P_1 if and only if

$$\frac{Q_1/N_1}{Q_2/N_2} > \tilde{\tau} \, . \tag{1}$$

The threshold τ depends on the losses and the a priori probabilities.

If the a priori probabilities q_1 and $q_2 = 1-q_1$ are not known, then the nonparametric classification problem must be reformulated.

By Equation (6) of Chapter 1, the Bayes procedure for known densities and known a priori probabilities is given by

$$t(1/x) = \begin{cases} 1 & \text{if } \dfrac{p_1(x)}{p_2(x)} > \dfrac{q_2}{q_1} W \\ \\ 0 & \text{if } \qquad\quad \leq \end{cases} \tag{2}$$

or

$$t(1/x) = \begin{cases} 1 & \text{if } \dfrac{q(1/x)}{q(2/x)} > W \\ \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where

$$W = \frac{L_{21} - L_{22}}{L_{12} - L_{11}}$$

We consider the problem [14] of classifying x given a random sample (of size N) $\underline{x}_N = (x_1, x_2, \ldots, x_N)$ independent of x. If N_1 of the

x_i's turn out to be from P_1, then N_1/N is a consistent estimate of q_1. Using the Fix-Hodges procedure with $K = K(N)$, a non-decreasing sequence of positive integers such that

$$\lim_{N\to\infty} K = \infty \tag{4}$$

$$\lim_{N\to\infty} K/N = 0, \tag{5}$$

then Q_1/K is a consistent estimate of $q(1/x)$ (and Q_2/K is a consistent estimate of $q(2/x)$). An important theorem in empirical Bayes hypothesis testing [26] states that if $f_N(x, \underline{X}_N)$ is such that for each x

$$f_N(x, \underline{X}_N) \longrightarrow q(1/x) - Wq(2/x) \quad \text{in probability} \tag{6}$$

(where "in probability" refers to the distribution of $\underline{X}_N$), then the decision rule

$$t_N(1/x) = \begin{cases} 1 & \text{if } f_N(x, \underline{X}_N) > 0 \\ 0 & \text{if } f_N(x, \underline{X}_N) \leq 0 \end{cases} \tag{7}$$

has a Bayes risk that approaches the Bayes envelope,

$$\lim_{N\to\infty} \bar{R}(q, t_N) = R(q) \ . \tag{8}$$

Thus the procedure of assigning x to P_1 if and only if

$$Q_1/Q_2 > W \tag{9}$$

is asymptotically optimum.

For the case of two classes with known q, we use (1) with fixed N_1 and N_2 to approximate (2). For unknown q, we use (9) with fixed $N(N = N_1+N_2)$ to approximate (3). For the general case of r classes (and unknown q), we choose a sample of size N, independent of x and a positive integer K depending on N and satisfying (4) and (5). Of the K values of the sample closest to x (in a specified metric) let Q_i be the number from class i,

$$\sum_{i=1}^{r} Q_i = K \ .$$

We assign x to the class j which minimizes the quantity

$$\sum_{i=1}^{r} L_{ij}\, Q_i \ . \tag{10}$$

From (2) or from Equation (3) of Section I, we see that the nonparametric classification problem (with known q) is basically one of estimating probability densities. Van Ryzin* has shown that if $\bar{R}(q, t^q(\hat{p}))$ is the average risk for a procedure designed to be Bayes against the known a priori probabilities q_1 and q_2 on the basis of estimates $\hat{p}_1(x)$ and $\hat{p}_2(x)$ for the densities, then for the case $L_{11} = L_{22} = 0$,

$$0 \le R(q, t^q(\hat{p})) - R(q) \le L_{12}\, q_1 \int \left|\hat{p}_1(x) - p_1(x)\right| dx + L_{21} q_2 \int \left|\hat{p}_2(x) - p_2(x)\right| dx.$$

In general [34],

$$\left|R(q, t^q(\hat{p})) - R(q)\right| \le \sum_{j=1}^{s} \sum_{i=1}^{r} \left|L_{ij}\right| q_i \int \left|\hat{p}_i(x) - p_i(x)\right| dx.$$

* J. R. Van Ryzin: "Bayes Risk Consistency of Classification Procedures Using Density Estimation," Unpublished.

The Fix-Hodges method is not the only way to obtain consistent estimates of density functions. The following method of "Window Carpentry" is from Murthy's extension [18] of the work in References [20, 25, 17].

Let $X_i = (X_{i1}, \ldots X_{in})$, $i = 1, 2, \ldots, N$, be independent and identically distributed n-dimensional random vectors with cumulative distribution function $F(x) = F(x_1, \ldots, x_n)$ and density function $f(x)$.

The sample distribution function $F_N(x_1, \ldots, x_n) = 1/N$ (No. of observations X_i, $i = 1, 2, \ldots, N$ such that $X_{ij} \leqq x_j$, $j = 1, 2, \ldots, n$) is a binominally distributed random variable whose mean and variance are given by

$$E\left[F_N(x)\right] = F(x)$$

$$\mathrm{Var}\left[F_N(x)\right] = (1/N)\, F(x)\left[1 - F(x)\right].$$

As an estimate of f(x), one might take

$$f_N(x_1, \ldots, x_n) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{1}{h_1 \cdots h_n} K\left(\frac{x_1 - y_1}{h_1}, \ldots, \frac{x_n - y_n}{h_n}\right)$$

$$dF_N(y_1, \ldots, y_n)$$

$$= \frac{1}{Nh_1 \ldots h_n} \sum_{i=1}^{N} K\left(\frac{x_1 - X_{i1}}{h_1}, \ldots, \frac{x_n - X_{in}}{h_n}\right)$$

where $K(x_1, \ldots, x_n)$ is an n-dimensional "window" (or weighting function) and the constants $h_j = h_j(N)$, $j = 1, \ldots, n$, are positive functions of N approaching zero as $N \rightarrow \infty$.

Let the window $K(x)$ satisfy the conditions

(1) $K(x) \geq 0$

(2) $K(x_1, \ldots, x_n) = K(|x_1|, \ldots, |x_n|)$

(3) $|x_{i1}| \geq |x_{i2}|$ for all $i = 1, \ldots, n \Rightarrow K(x_1) \leq K(x_2)$

(4) $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} K(x)\, dx_1 \ldots dx_n = 1.$

Then at all points x at which $f(x)$ is continuous, $f_N(x)$ is asymptotically unbiased, i.e.,

$$\lim_{N\to\infty} E\left[f_N(x)\right] = f(x),$$

and also

$$\lim_{N\to\infty} Nh_1 \ldots h_n \operatorname{Var}\left[f_N(x)\right] = f(x) \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} K^2(y)\, dy_1 \ldots dy_n$$

If in addition to

$$\lim_{N\to\infty} h_j(N) = 0 \;,$$

the positive constants $h_j = h_j(N)$, $j = 1, \ldots, n$ also satisfy

$$\lim_{N\to\infty} Nh_1(N) \ldots h_n(N) = \infty \;,$$

then $f_N(x)$ is a consistent (and asymptotically normal) estimate of $f(x)$ at all points of continuity of $f(x)$.

Examples: Let $n = 2$, let $h_1 = h_2 = h$, and let

$$K(x) = \begin{cases} 1 & \text{if } |x_i| \leq 1/2,\ i = 1, 2 \\ 0 & \text{otherwise} \end{cases}$$

Then, with this "square window of size h",

$$f_N(x_1, x_2) = \frac{F_N(x_1 + \frac{h}{2}, x_2 + \frac{h}{2}) - F_N(x_1 - \frac{h}{2}, x_2 - \frac{h}{2})}{h^2}$$

If, in the Fix-Hodges method, we choose the metric implied by the norm

$$\|x\| = \min |x_i| ,$$

the procedure becomes equivalent to estimation with a square window whose size is a random variable. If the Euclidean metric with norm $|x| = (\Sigma x_i^2)^{1/2}$ is used, it is equivalent to a "circular" window of radius $h(h = h_1 = h_2 \ldots = h_n)$:

$$K(x) = \begin{cases} 1 & \text{if } |x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

with h as a random variable.

We note that for the case $h_1 = h_2 = \ldots = h_n = h$, the estimate of f(x) may be written more compactly as

$$f_N(x) = \frac{1}{Nh^n} \sum_{i=1}^{N} K \left(\frac{x - X_i}{h} \right)$$

with $h \rightarrow 0$ and $Nh^n \rightarrow \infty$ as $N \rightarrow \infty$. Some arguments have been given [35] for taking $h(N) = O(N^{-1/(n+2)})$.

2.2 NEAREST-NEIGHBOR CLASSIFICATION OF BINARY PATTERNS

When K = 1, the Fix-Hodges procedure is called the nearest-neighbor method. For known q we chose N_1 and N_2 such that $N_1/N_2 = q_1/q_2$ while for unknown q we fix only $N = N_1 + N_2$. In either case, we assign the unknown to the class from which its nearest neighbor came.

Cover and Hart [6] have shown that in a large-sample analysis, the probability of error is less than twice the Bayes envelope error probability, R(q). If $R \equiv R(q)$, and R^* is the limit of the nearest neighbor error probability as $N \rightarrow \infty$, then for r classes and $L_{ij} = 1 - \delta_{ij}$, they prove the theorem:

> Let S be a separable metric space. Let $p_1, p_2, \ldots, p_r$ be probability densities with respect to some probability measure μ such that, with probability one, x is either a) a continuity point of $p_1, p_2, \ldots, p_r$, or b) a point of nonzero probability measure. Then
>
> $$R \leq R^* \leq R\,(2 - (r/r-1)\,R)\,.$$
>
> These bounds are as tight as possible, in that they are achieved with particular sets of densities.

For classifying binary vectors, we use Hamming distance as the metric; the distance between two vectors is equal to the number of components that differ. We shall assume a specific form for the distributions, namely independent identically distributed components, and derive the probability of error (for two classes), first using a parametric technique and second using the nearest-neighbor rule. We shall assume equal a priori probabilities, $q_1 = q_2 = 1/2$, throughout.

Let $x = (x_1, \dots, x_n)$ denote an observation from class 1, where the x_i's are independent, binary, and $P\{x_i = 1\} = \alpha$ for all i. Let $y = (y_1, \dots, y_n)$ denote an observation from class 2, where the y_i's are independent, binary, and $P\{y_i = 1\} = \beta$ for all i. Assume further that $\beta = 1 - \alpha$. If α is unknown, it is estimated from N samples of each class, $x^{(1)}, \dots, x^{(N)}$, $y^{(1)}, \dots, y^{(N)}$ as

$$\hat{\alpha} = \frac{1}{2Nn} \sum_{j=1}^{N} \sum_{i=1}^{n} x_i^{(j)} + 1 - y_i^{(j)} .$$

We have

$$P\left\{\hat{\alpha} = \frac{m}{2Nn}\right\} = b(m; 2Nn, \alpha)$$

where $b(k;n, p) = \binom{n}{k} p^k (1-p)^{n-k}$.

A sample $z = (z_1, \dots, z_n)$ of unknown origin is found to contain r ones and n-r zeroes. The likelihood ratio (or Bayes) decision is to decide class 1 if

$$r < n/2 \text{ and } \hat{\alpha} < 1/2$$

or if

$$r > n/2 \text{ and } \hat{\alpha} > 1/2,$$

and class 2 if

$$r < n/2 \text{ and } \hat{\alpha} > 1/2$$

or if

$$r > n/2 \text{ and } \hat{\alpha} < 1/2 .$$

Since $q_1 = q_2 = 1/2$, the probability of misclassification is equal to the probability of misclassifying a sample from class 1:

$$P_{1e} = P_1 \left\{r < n/2\right\} P\left\{\hat{\alpha} > 1/2\right\} + P_1 \left\{r > n/2\right\} P\left\{\hat{\alpha} < 1/2\right\} .$$

For n odd

$$P_1 \left\{r < n/2\right\} = \sum_{r<n/2} P_1(r) = \sum_{r<n/2} \binom{n}{r} \alpha^r (1-\alpha)^{n-r} = B\left(\frac{n-1}{2}; n, \alpha\right)$$

where $B(k;n,p) = \sum_{r=1}^{k} b(r;n,p)$. Thus the probability of error is

$$P_{1e} = B\left(\frac{n-1}{2}; n, \alpha\right) \left[1 - B(Nn;\ 2Nn, \alpha)\right] + \left[1 - B\left(\frac{n-1}{2}; n, \alpha\right)\right] B(Nn-1;\ 2Nn, \alpha)$$

for n odd. Theprobability of rejection (a tie) for n odd, is

$$P_{1r} = b(Nn;\ 2Nn, \alpha) .$$

When we are faced with the preceeding problem without being armed with the knowledge that the components are independent, we might use the nearest-neighbor method. The error probability, P_{1e}, is then equal to the probability that the nearest to z is a y, given that z is an x. Let $P_2(r)$ be the probability that the nearest neighbor to z is a y, given that z has r ones. Then P_{1e} is the expectation of $P_2(r)$ with z distributed according to P_1.

$$P_{1e} = E_1 \left[P_2(r)\right] = \sum_{r=0}^{n} \binom{n}{r} \alpha^r (1-\alpha)^{n-r} P_2(r) .$$

Let $h_r(\delta)$ be the probability that the Hamming distance from z to an x is δ (given that z has r ones). It is the probability that x is the same as z in n-δ places and different in δ places.

$$h_r(\delta) = P\left\{d(x, z) = \delta\right\}.$$

Similarly, we define

$$k_r(\delta) = P\left\{d(y, z) = \delta\right\}$$

$$H_r(\delta) = P\left\{d(x, z) \le \delta\right\} = \sum_{d=0}^{\delta} h_r(d)$$

$$K_r(\delta) = P\left\{d(y, z) \le \delta\right\} = \sum_{d=0}^{\delta} k_r(d) \ .$$

Then

$$P_2(r) = N \sum_{\delta=0}^{n} \left[1-H_r(\delta)\right]^N \left[1-K_r(\delta-1)\right]^{N-1} k_r(\delta),$$

where

$$h_r(\delta) = \sum_{s=0}^{\delta} b(s;r,1-\alpha)\, b(\delta-s;n-r,\alpha) = \sum_{s=0}^{\delta} b(r-s;r,\alpha)\, b(\delta-s;n-r,\alpha)$$

$$H_r(\delta) = \sum_{d=0}^{\delta} \sum_{s=0}^{d} b(r-s;r,\alpha)\, b(d-s;n-r,\alpha) = \sum_{s=0}^{\delta} \sum_{d=s}^{\delta} b(r-s;r,\alpha)\, b(d-s;n-r,\alpha)$$

$$= \sum_{s=0}^{\delta} b(r-s;r,\alpha) \sum_{t=0}^{\delta-s} b(t;n-r,\alpha) = \sum_{s=0}^{\delta} b(r-s;r,\alpha)\, B(\delta-s;n-r,\alpha),$$

with $k_r(\delta)$ and $K_r(\delta)$ defined the same way but with α replaced by β.

Thus, for the nearest neighbor method, the probability of error is

$$P_{1e} = \sum_{r=0}^{n} b(r;n,\alpha)\, P_2(r),$$

where

$$P_2(r) = N \sum_{\delta=0}^{n} \left\{ \left[1 - \sum_{s=0}^{\delta} b(r-s;r,\alpha)\; B(\delta-s;\, n-r,\alpha)\right]^{N} \right.$$

$$\cdot \left[1 - \sum_{s=0}^{\delta-1} b(r-s;\, r,\beta)\; B(\delta-1-s;\, n-r,\beta)\right]^{N-1}$$

$$\left. \cdot \left[\sum_{s=0}^{\delta} b(r-s;r,\beta)\; b(\delta-s;n-r,\beta)\right] \right\} .$$

The probability of rejection (with $\beta = 1-\alpha$) is

$$P_{1r} = 1 - \sum_{r=0}^{n} \left[b(r;n,\alpha) + b(r;n,\beta)\right] P_2(r) .$$

Even for this "trivial" problem, meaningful numerical results are more easily obtained by computer experiments than by attempting to evaluate the equations.

2.3 MARKOV-CHAIN CLASSIFICATION OF BINARY PATTERNS

In a variety of pattern-recognition problems involving the classification of pictorial data, the gray-scale image is first converted into a black-and-white picture by one of several filtering techniques. The purpose of this preprocessing step is to simplify the data for further processing and to remove the nuisance variables of brightness and contrast so that consistent, detailed binary pictures of the original image are obtained. The black-and-white pictures can be considered to be two-dimensional arrays of binary random variables.

Applying statistical classification procedures to joint distributions of binary random variables has for the most part invoked either the assumption of statistical independence of the variables, or the assumption that their joint distributions are multivariate normal. The first assumption, while leading to simple results, obviously is very limiting. The multivariate normal approach is also limited and requires special development when the sample covariance matrices are singular. This present section developes a general procedure for the classification of patterns of binary random variables when neither of the above assumptions is invoked. In this sense, the procedure is "operationally" nonparametric.

Let S denote the set of 2^n states of $x = (x_1, x_2, \ldots, x_n)$, each x_i taking on values 1 and 0. When a pattern x can belong to one of two groups with probability distributions p(x) and q(x) respectively, the logarithm of the likelihood ratio, $\log L(x) = \log p(x) - \log q(x)$, is widely used as an optimal classification function. If the 2^n-1 probabilities associated with each of the n-variate binary distributions were specified and nonzero, then the classification function would be specified.

Generally, obtaining and storing all the 2^n-1 probabilities associated with each of the alternative distributions will be out of the question, even when n is of moderate size. To overcome this, it seems there is no choice other than to sacrifice the generality of our formulation and impose restrictions on the nature of the dependence between the n binary random variables. After independence, the next step is to consider Markovian dependence. The following shows how such dependence can be converted into the type of spatial dependence of interest in the problems being considered.

Assume a first-order Markov chain:

$$p(x_k/x_1x_2\ldots x_{k-1}) = p(x_k/x_{k-1}) \tag{11}$$

for $k = 2, 3, \ldots n$. A well-known property of such a chain is that

$$p(x_k/x_1x_2\ldots x_j) = p(x_k/x_j) \tag{12}$$

for all $j < k$. A not-so-well-known property is that for $k < n$

$$\begin{aligned} p(x_k/x_1x_2\ldots x_{k-1}\, x_{k+1}\ldots x_n) &= \frac{p(x_1\ldots x_n)}{p(x_1\ldots x_{k-1}x_{k+1}\ldots x_n)} \\ &= \frac{p(x_1)p(x_2/x_1)\ldots p(x_k/x_{k-1})p(x_{k+1}/x_k)\ldots p(x_n/x_{n-1})}{p(x_1)p(x_2/x_1)\ldots p(x_{k+1}/x_{k-1})\ldots p(x_n/x_{n-1})} \\ &= \frac{p(x_k/x_{k-1})p(x_{k+1}/x_k)}{p(x_{k+1}/x_{k-1})} = \frac{p(x_{k-1})p(x_k/x_{k-1})p(x_{k+1}/x_k)}{p(x_{k-1})p(x_{k+1}/x_{k-1})} \\ &= \frac{p(x_{k-1}x_kx_{k+1})}{p(x_{k-1}x_{k+1})} = p(x_k/x_{k-1}x_{k+1}) \end{aligned} \tag{13}$$

so that any point is dependent on only its two nearest neighbors, one on either side. Similarly, for the r^th^-order Markov chain,

$$p(x_k/x_1x_2\ldots x_{k-1}) = p(x_k/x_{k-r}\ldots x_{k-1}) \, , \tag{14}$$

we have

$$p(x_k/x_1x_2\ldots x_{k-1}x_{k+1}\ldots x_n) = p(x_k/x_{k-r}\ldots x_{k-1}x_{k+1}\ldots x_{k+r}) \, . \tag{15}$$

The converse is not, in general, true; i.e., the assumption of dependence on the r nearest neighbors on each side does not imply an r^{th}-order Markov chain. The chain is a special case of dependence on the 2r nearest neighbors. Consider the following example. Let $r = 1$ and $n = 5$, that is, we have five variables. Let $x_i = 0$ or 1 and $p(00011) = 1/2$ and $p(11000) = 1/2$ with all other states having probability zero. Then we find that Equation (15) is satisfied but $p(x_4 = 1/x_1 = x_2 = 1, \ x_3 = 0) = 0$ while $p(x_4 = 1/x_3 = 0) = 1/2$ hence

$$p(x_4/x_1x_2x_3) \neq p(x_4/x_3)$$

and (11) is not satisfied.

For the first-order chain (Equation 11) let

$$\alpha_i = p(x_i = 1/x_{i-1} = 0)$$

$$\beta_i = p(x_i = 1/x_{i-1} = 1) \tag{16}$$

It is convenient to define $x_i = 0$ for $i < 1$ and $i > n$ so that

$$\alpha_1 = p(x_1 = 1)$$

and $\alpha_{n+1} = \beta_{n+1} = 0$. There are now 2n-1 parameters rather than 2^n-1. If α_i and β_i were independent of i (which is not the case here), one would have a stationary chain.

Expanding the joint probability:

$$p(x_1 x_2 \ldots x_n) = p(x_1)\, p(x_2/x_1) \ldots p(x_n/x_{n-1})$$

$$= \alpha_1^{x_1}(1-\alpha_1)^{1-x_1} \prod_{i=2}^{n} \Big\{ \beta_i^{x_{i-1}x_i} (1-\beta_i)^{x_{i-1}(1-x_i)}$$

$$\cdot\, \alpha_i^{(1-x_{i-1})x_i} (1-\alpha_i)^{(1-x_{i-1})(1-x_i)} \Big\} . \qquad (17)$$

Taking logarithms and collecting terms, we obtain:

$$\log p(x_1 x_2 \ldots x_n) = A_o + \sum_{i=1}^{n} A_i x_i + \sum_{i=2}^{n} B_i x_{i-1} x_i \,, \qquad (18)$$

where

$$A_o = \sum_{i=1}^{n} \log(1-\alpha_i)$$

$$A_i = \log \frac{\alpha_i}{1-\alpha_i} + \log \frac{1-\beta_{i+1}}{1-\alpha_{i+1}} \qquad (19)$$

$$B_i = \log \frac{\beta_i}{1-\beta_i} - \log \frac{\alpha_i}{1-\alpha_i} \,.$$

For independent variables, $\alpha_i = \beta_i$ and hence $B_i = 0$.

For the second-order Markov chain (dependence on the four nearest neighbors), let

$$\begin{aligned}
\alpha_i &= p(x_i = 1/x_{i-2} = 0,\ x_{i-1} = 0)\\
\beta_i &= p(x_i = 1/x_{i-2} = 0,\ x_{i-1} = 1)\\
\gamma_i &= p(x_i = 1/x_{i-2} = 1,\ x_{i-1} = 0)\\
\delta_i &= p(x_i = 1/x_{i-2} = 1,\ x_{i-1} = 1)
\end{aligned} \qquad (20)$$

Since $x_i = 0$ for $i < 1$ and $i > n$, $\alpha_1 = p(x_1 = 1)$, $\alpha_2 = p(x_2 = 1/x_1 = 0)$, $\beta_2 = p(x_2 = 1/x_1 = 1)$ and there are now $4(n-2) + 3 = 4n - 5$ parameters. The terms in Equation (20) are all zero for $i > n$.

Expanding the joint probability:

$$p(x_1 x_2 \ldots x_n) = p(x_1) p(x_2/x_1) p(x_3/x_1 x_2) \ldots p(x_n/x_{n-2} x_{n-1}). \qquad (21)$$

aking logarithms and collecting terms, we obtain

$$\begin{aligned}
\log p(x_1 x_2 \ldots x_n) = A_o &+ \sum_{i=1}^{n} A_i x_i + \sum_{i=2}^{n} B_i x_{i-1} x_i\\
&+ \sum_{i=3}^{n} C_i x_{i-2} x_i + \sum_{i=3}^{n} D_i x_{i-2} x_{i-1} x_i\ ,
\end{aligned} \qquad (22)$$

where

$$A_o = \sum_{i=1}^{n} \log (1 - \alpha_i)$$

$$A_i = \log \frac{\alpha_i}{1 - \alpha_i} + \log \frac{1 - \beta_{i+1}}{1 - \alpha_{i+1}} + \log \frac{1 - \gamma_{i+2}}{1 - \alpha_{i+2}} \; .$$

$$B_i = \log \frac{\beta_i}{1 - \beta_i} - \log \frac{\alpha_i}{1 - \alpha_i} + \log \frac{1 - \delta_{i+1}}{1 - \gamma_{i+1}} - \log \frac{1 - \beta_{i+1}}{1 - \alpha_{i+1}}$$

$$C_i = \log \frac{\gamma_i}{1 - \gamma_i} - \log \frac{\alpha_i}{1 - \alpha_i}$$

$$D_i = \log \frac{\delta_i}{1 - \delta_i} - \log \frac{\gamma_i}{1 - \gamma_i} - \log \frac{\beta_i}{1 - \beta_i} + \log \frac{\alpha_i}{1 - \alpha_i} \; ,$$

Setting $\gamma_i = \alpha_i$ and $\delta_i = \beta_i$ reduces Equation (22) to Equation (18). The assumption of a third-order chain would lead to summation of x_i, $x_{i-1}x_i$, $x_{i-2}x_i$, $x_{i-3}x_i$, $x_{i-2}x_{i-1}x_i$, $x_{i-3}x_{i-1}x_i$, $x_{i-3}x_{i-2}x_i$, and $x_{i-3}x_{i-2}x_{i-1}x_i$ in terms of 8n - 17 parameters. In general, an r^{th}-order chain (which gives dependence on the 2r nearest neighbors) results in an expansion of the lorarithm of the joint probability up to products of r + 1 adjacent variables, with $2^r(n-r+1)-1$ parameters. Classification is obtained by thresholding the difference between two such expansions.

The Markov assumption of nearest-neighbor dependence, Equation (14), implies a one-dimensional process or sequence. For the classification of two-dimensional patterns, one must scan the pattern and apply the chain assumption to the scanned output. Since we assume that a point depends on only the r points on either side along the scan line, the scan line must be so constructed as to remain as close as possible to a given point for the r succeeding (as well as preceding) points. Hence, for $r > 1$, we should scan the pattern with a continuous curve that, while passing through each point in a rectangular array, is as crimpled as possible. An example of such a curve* is illustrated in Figure 2. The limit of the curves f_m as $m \longrightarrow \infty$ is a continuous curve, called a space-filling curve, [13] that passes through every point of a given area. The curve f_m scans a $2^m \times 2^m$ array of points, while never maintaining the same direction for more than three consecutive points. Whenever it has strayed three points in a straight line, it turns around and comes back.

The curves of Figure 2 do not provide for all spatial dependencies that exist. However, the dependencies that are assumed do get converted into spatial dependencies. Hence, we are assured of doing better than by the simple assumption of independence. A larger class of spatial dependencies is taken care of by extending the Markov-chain methods to two dimensions. The two-dimensional analog so obtained is called a Markov mesh, and is discussed in Reference [2].

* These curves were first presented by David Hilbert [13] in 1891. To the best of our knowledge, they have not found application until now.

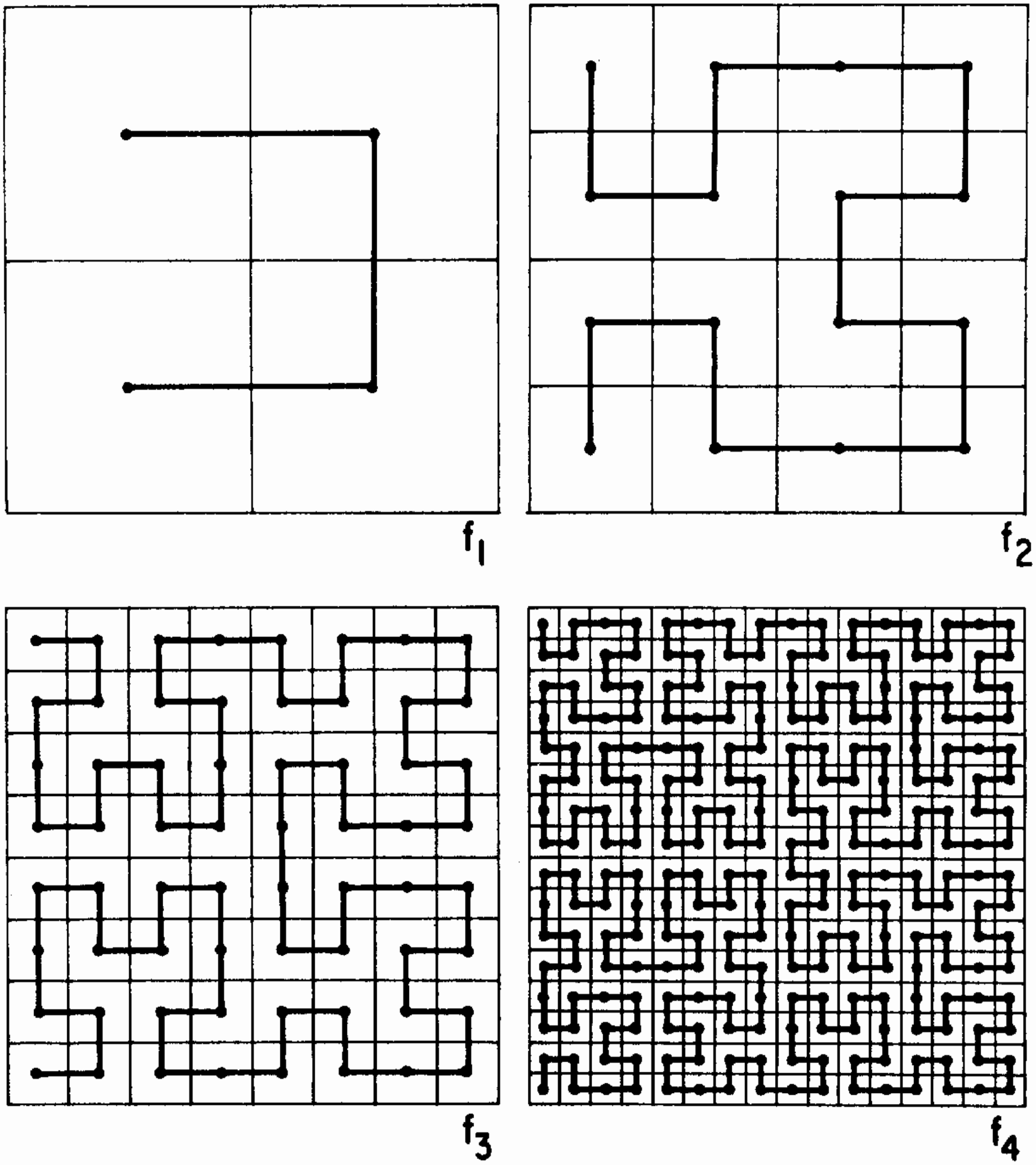

Figure 2 An Open Space-Filling Curve — the First Four Steps

In pattern-recognition problems, the number of variables in the array is usually larger than 25 and often in the thousands. For n binary random variables using the Markov-chain assumption, one only has to estimate on the order of $2^r n$ parameters, with r taken to be 2 or 3. This is much smaller than 2^n, it is also smaller than the parameters that the normal assumption requires, for in that case we have to estimate n means and an n x n covariance matrix. Equipment designed for classification on the basis of an r^{th}-order chain assumption can be used for any chain of order less than r by merely equating certain parameters.

2.4 THE DISTRIBUTION FREE COMPOUND DECISION PROBLEM

For unknown densities $p_i(x)$ (as well as unknown G) one may take one of two approaches. The nonparametric empirical Bayes approach of Johns [14] assumes that the θ's are independent and identically distributed and examines convergence of the component Bayes risk to R(q) for the case $G(\underline{\theta}_N) = \prod_{k=1}^{N} q(\theta_k)$. It has been assumed in the nonparametric problem that when the k^{th} decision must be made there are at hand the true values of all the previous states of nature and a sequential compound decision rule with $t_k = t_k(j/\underline{\theta}_{k-1}, \underline{x}_k)$ is used. One may use $\underline{\theta}_{k-1}$ and $\underline{x}_{k-1}$ to obtain consistent estimates of $P_i(x)$ for all $i \in \Omega$ and act as if these were the true densities to define a procedure "Bayes" against the known empirical a priori distribution q^{k-1}. Such a procedure is asymptotically optimum in the empirical Bayes sense. Johns [14] uses the Fix-Hodges procedure [8, 9] while Van Ryzin [33] uses "Window Caprentry" [20, 18].

Van Ryzin [34] has shown that his above-mentioned procedure $t_k(j/\underline{\theta}_{k-1}, \underline{x}_k)$ is also optimum in the limit in the sense that the compound risk converges to $R(q^N)$ for any sequence $\underline{\theta}_N$. However, the problem he considers is repetitive play in statistical games and it would be artificial to apply his formulation directly to pattern recognition. He assumes that the true value of θ_k is given to us after each deçision is made and examines convergence of the average of the component risks. In a nonparametric pattern-recognition problem one must be given "training samples" from which to estimate the densities, but we should not be concerned with whether or not the risk of a rule applied to these samples converges. What must be examined is convergence of the risk applied to the samples we wish to test and whose true classification is never known. This is done in the empirical Bayes approach, i.e., one investigates convergence of the N^{th} component Bayes risk given N-1 training samples. However, this approach has two major limitations. It assumes that the states of nature are independent and it assumes that only one pattern need really be classified.

In a pattern-recognition problem there is a series of design samples $\underline{x}_m = (x_1, \ldots, x_m)$ with known classification, i.e., $\underline{\theta}_m = (\theta_1, \ldots, \theta_m)$ is known. There is then a sequence of test samples

$\underline{x}_N{}^m = (x_{m+1}, \ldots, x_{m+N})$ with unknown $\underline{\theta}_N{}^m = (\theta_{m+1}, \ldots, \theta_{m+N})$. What really matters is convergence of the compound risk $R(\underline{\theta}_N{}^m, \underline{t}_N)$ for the test samples. The design samples can be used to estimate the densities (learning with a teacher) and the test samples may be used (in a sequential compound rule) to estimate their empirical distribution and possibly to improve the estimates of the densities (unsupervised updating). Unsupervised adaptation has been demonstrated under certain conditions when the distribution functions of the classes differ only in location by Cooper and Cooper [5].

A distribution-free sequential compound decision rule is denoted by $\underline{t}_N = (t_1, \ldots, t_N)$,where $t_k = t_k(j/\underline{\theta}_m, \underline{x}_{m+k})$ (its non-sequential counterpart has $t_k = t_k(j/\underline{\theta}_m, \underline{x}_{m+N})$). If infinitely many design samples from each class are available, then the densities can be estimated exactly and the sequential compound procedure described in Chapter 1 (which estimates the empirical distribution q^{k-1}) is optimal in the limit: $R(\underline{\theta}_N, \underline{t}_N{}^*) \longrightarrow R(q^N)$. If we assume a cost of sampling C_i for each design sample from class i, the compound risk for the rule $\underline{t}_N$ with $t_k = t_k(j/\underline{\theta}_m, \underline{x}_{m+k})$ is

$$R_m(\underline{\theta}_{m+N}, \underline{t}_N) = R(\underline{\theta}_N{}^m, \underline{t}_N) + \frac{1}{N} \sum_{i=1}^{m} C_{\theta_i} .$$

For finite m, $R_m(\underline{\theta}_{m+N}, \underline{t}_N) \longrightarrow R(q^N)$ if and only if $R(\underline{\theta}_N{}^m, \underline{t}_N) \longrightarrow R(q^N)$.

Thus we wish to know under what circumstances it is possible to find a rule such that for some $\underline{\theta}_m$

$$R(\underline{\theta}_N{}^m, \underline{t}_N) \rightarrow R(q^N)$$

uniformly for all $\underline{\theta}_N{}^m \in \Omega^N$. In other words, is there some design procedure and some set of design samples for which the compound risk for the test samples converges to its minimum possible value for any sequence of test samples?

Section III

ADAPTIVE THRESHOLD ADJUSTMENT

3.1 THE TWO-THRESHOLD METHOD

We return to the problem of Section 1.4, page 15, where $A = \Omega = \{0, 1\}$ and only the a priori probability $q = q_1$ is unknown ($q_o = 1-q$). The sequential compound rule of Equation (28), Section 1.4, which chooses class 1 if

$$\frac{p_1(x_k)}{p_o(x_k)} > \frac{L_{01} - L_{00}}{L_{10} - L_{11}} \quad \frac{1 - \hat{q}_{k-1}}{\hat{q}_{k-1}} \tag{1}$$

is asymptotically optimum when $\hat{q}_k$ is a consistent estimator of q. Van Ryzin has shown [32] that this rule even satisfies Equation (23) of Section 1.4, if $\hat{q}_k$ is given by Equations (24) and (25) of Section 1.4, i.e.

$$\hat{q}_{k-1}(\underline{x}_{k-1}) = \left\{ h_{k-1} (\underline{x}_{k-1}) \right\}_{\text{Truncated,}} \tag{2}$$

$$h_k(\underline{x}_k) = \frac{1}{k} \sum_{i=1}^{k} \quad h(x_i), \tag{3}$$

and $h(x_k)$ is a bounded unbiased estimator of θ_k. Since $Eh_k = \overline{\theta}_k$ and Var $h_k \leq (1/k) \max_i$ Var $h(x_i)$, h_k is a consistent estimator of $\overline{\theta}_k$. It remains for us to specify h(x).

Hannan and Robbins [10] have found the h(x) that minimizes the variance of h_k. Their h(x) is given implicitly in a pair of simultaneous integral equations. Samuel [27, 24] presents an h(x) that is much simpler. She sets

$$h(x) = \frac{f_T(x) - P_o(T)}{P_1(T) - P_o(T)} \quad , \tag{4}$$

where T is a set in S for which $P_o(T) \neq P_1(T)$ and $f_T(x)$ is the characteristic function of T. Since

$$Ef_T(x) = P_\theta(T) = \theta P_1(T) + (1-\theta)\ P_o(T), \tag{5}$$

it follows that $Eh(x) = \theta$. (We have omitted the subscript k from the class θ_k and the pattern x_k, θ is either 0 or 1.) It now remains for us to specify the set T.

We shall first prove that the set T that minimizes the variance of h_k is given by a threshold on the likelihood ratio:

$$T = \left\{ x:\ p_1(x)/p_o(x) > \tau \right\} \quad . \tag{6}$$

Since $E\ f_T(x) = P_\theta(T)$, we have $Eh(x_i) = \theta_i$, $Eh_k = \overline{\theta}_k$, and $\mathrm{Var} f_T(x_i) = P_{\theta_i}(1-P_{\theta_i}) = \theta_i\ P_1(1-P_1) + (1-\theta_i)\ P_0(1-P_0)$. Thus

$$\mathrm{Var}\ h_k = \frac{\sum_{i=1}^{k} \mathrm{Var}\ f_T(x_i)}{k^2 \left[P_1-P_0\right]^2} = \frac{\overline{\theta}_k P_1(1-P_1)+(1-\overline{\theta}_k)P_0(1-P_0)}{k(P_1-P_0)^2} \quad . \tag{7}$$

Hence

$$k\ \mathrm{Var}\ h_k = \frac{\overline{\theta}_k(P_1-P_1^{\ 2}-P_0+P_0^{\ 2}) + P_0(1-P_0)}{(P_1-P_0)^2}$$

$$k \text{ Var } h_k = \overline{\theta}_k \frac{1-(P_1+P_0)}{(P_1-P_0)} \qquad \frac{P_0(1-P_0)}{(P_1-P_0)^2} \quad .$$

Let T^+ minimize Var h_k. Let $P_0(T^+) = P_0$, $P_1(T^+) = P_1{}^+$. Without loss of generality we may assume $P_0 < P_1{}^+$. Consider the set

$$T^* = \left\{ x: \; p_1(x)/p_0(x) > \tau \; , \; P_0(T^*) = P_0 \right\} .$$

By the Neyman-Pearson lemma, $P_1{}^+ \leq P_1{}^*$. Therefore, $(P_1{}^* - P_0) \geq (P_1{}^+ - P_0)$ and since both sides are positive, $(P_1{}^* - P_0)^2 \geq (P_1{}^+ - P_0)^2$. Furthermore, $1 - (P_1{}^* + P_0) < 1 - (P_1{}^+ + P_0)$. Thus k Var $h_k{}^* \leq$ k Var $h_k{}^+$. But, since Var $h_k{}^+$ was assumed to be the minimum, Var $h_k{}^*$ = Var $h_k{}^+$ and the set T^* minimizes the variance.*

From (7) we see that to minimize the maximum (over $\overline{\theta}_k$) variance of h_k, we must choose τ such that $P_0(T) = 1 - P_1(T)$. From Equation (10) of Section 1.2 we see that this is simply the threshold for minimax probability of error in the simple decision problem ($L_{ij} = 1 - \delta_{ij}$ implies $w = 1$).

In the two-class pattern-recognition problem we somehow implement the likelihood ratio as a classification function [19]. We set two thresholds on it, a fixed threshold (6) and an adjustable decision threshold (1). The fixed threshold, τ , is set to give a minimax probability of error. Letting $\overline{f}_T(\underline{x}_{k-1})$ be the fraction of times the fixed threshold was exceeded, we use

*. This proof is due to T. J. Harley, Jr., Philco Corporation.

$$\hat{q}_{k-1}(\underline{x}_{k-1}) = \left\{ \frac{\bar{f}_T(\underline{x}_{k-1}) - P_0(T)}{P_1(T) - P_0(T)} \right\}_{\text{Truncated}} \tag{8}$$

to adjust the decision threshold in Equation (1), where $\hat{q}_0 = 1/2$. This procedure has the advantage of being easily implemented, and our choice of T yields a minimax variance for the unbiased quantity in brackets in Equation (8).

The following section presents an alternate choice for the estimator $\hat{q}_{k-1}$.

3.2 THE BAYESIAN APPROACH

Let us assume that the a priori probability, q, is distributed according to an a priori density, f(q), on the unit interval. The conditional average risk for the k+1st decision, given the first k patterns is

$$E_q\left[\bar{R}(q, t_{k+1})/\underline{x}_k\right] = \int_0^1 R(q, t_{k+1})\, f(q/\underline{x}_k)\, dq \quad , \tag{9}$$

where $\bar{R}(q, t_{k+1})$ is defined as in Section 1.2. It is minimized by choosing $t_{k+1}(1/x_{k+1}) = 1$ if

$$\frac{p_1(x_{k+1})}{p_0(x_{k+1})} > \frac{L_{01} - L_{00}}{L_{10} - L_{11}} \quad \frac{1 - \hat{q}_k(\underline{x}_k)}{\hat{q}_k(\underline{x}_k)} \quad , \tag{10}$$

where

$$\hat{q}_k = E_q(q/\underline{x}_k) = \int_0^1 q\, f(q/\underline{x}_k)\, dq. \tag{11}$$

The density used in (11) to compute q_k can be defined itteratively in terms of $f(q/\underline{x}_{k-1})$ and $\hat{q}_{k-1}$:

$$f(q/\underline{x}_k) = \frac{p(x_k, q/\underline{x}_{k-1})}{p(x_k/\underline{x}_{k-1})} = \frac{p(x_k/q)\, f(q/\underline{x}_{k-1})}{\int p(x_k/q)\, f(q/\underline{x}_{k-1})\, dq}$$

$$= \frac{\left[q\, p_1(x_k) + (1-q)\, p_0(x_k)\right] f(q/\underline{x}_{k-1})}{\hat{q}_{k-1}\, p_1(x_k) + (1-\hat{q}_{k-1})\, p_0(x_k)} \quad , \tag{12}$$

where $f(q/\underline{x}_0) = f(q)$. The denominator in (12) is a normalization factor to insure $\int f(q/\underline{x}_k)\, dq = 1$. In closed form,

$$\hat{q}_k(\underline{x}_k) = \frac{\int_0^1 q f(q) \prod_{i=1}^{k} \left[q\, p_1(x_i) + (1-q)\, p_0(x_i)\right] dq}{\int_0^1 f(q) \prod_{i=1}^{k} \left[q\, p_1(x_i) + (1-q)\, p_0(x_i)\right] dq} \quad . \tag{13}$$

The first procedure (Section 3.1) for obtaining $\hat{q}_k$, though consistent, is not very efficient for small k; $\hat{q}_1(x_1)$, for example, can only take the values 0 or 1. Thus, $t_2^*(1/\underline{x}_2)$ does not even depend on x_2. The estimate $\hat{q}_k$ depends only on whether or not the likelihood ratio, for each x_i, exceeds a fixed threshold, and not on by how much the threshold is exceeded.

The second procedure, though not as simple, can be implemented iteratively by Equations (11) and (12). Letting

$$a_i = p_1(x_i)/p_0(x_i) - 1$$

and assuming f(q) = 1, Equation (13) becomes

$$\hat{q}_k = \frac{\int_0^1 q \prod_{i=1}^{k} \left[1 + a_i q\right] dq}{\int_0^1 \prod_{i=1}^{k} \left[1 + a_i q\right] dq} \quad . \tag{14}$$

Thus $\hat{q}_0 = 1/2$, $1/3 \leq \hat{q}_1(x_1) \leq 2/3$, and in general $q_k(\underline{x}_k)$ depends on the values of $x_1, \ldots, x_k$, with

$$\frac{1}{k+2} \leq \hat{q}_k(\underline{x}_k) \leq \frac{k+1}{k+2} \quad . \tag{15}$$

The choice of a uniform a priori density for q, leading to $\hat{q}_0 = 1/2$, is reasonablewhen the Bayes envelope has its maximum at q = 1/2, since the first decision is then minimax.

In this chapter, the densities, $p_i(x)$ were assumed to be known. In the following chapters, we treat the case where the a priori distribution, $G(\underline{\theta}_N)$, is also known.

Section IV

COMPOUND PROCEDURES FOR DEPENDENT STATES OF NATURE

4.1 USE OF CONTEXT IN PRINT READING

The objective here is to obtain as accurate and complete character recognition as is possible, by using context to assist in deciding character identity. The redundancy of the English language makes such an approach not only feasible, but extremely promising. As a "quick fix" for an existing character reader, one is tempted to adjust thresholds to allow greater rejection (lowest error rate) and to use context in identifying only the rejected characters. However, if optimal recognition is to be obtained, no character should ever be identified without regard to context. Similarly, no rejected character should be identified solely by context and without regard to "what it looks like".

Let x be an observation of what may be a character of the English language, i.e., a vector in the pattern space S. Let θ be a character of the language (including a "space"), i.e., a pattern class or a point in the parameter space, Ω. Suppose we have a method for computing the approximate conditional probability of an observation given the class, i.e., we can compute $p(x/\theta)$ (the probability function or density of x, given that x is an observation on character θ). For a given x, $p(x/\theta)$ is called the likelihood of θ.

Assuming a "constant" loss matrix $L_{ij} = 1 - \delta_{ij}$, the Bayes decision (minimum risk) is to choose that character θ which maximizes either

$$p(\theta/x) = \frac{p(x/\theta)\ p(\theta)}{p(x)} \tag{1}$$

or the product $p(x/\theta)\ p(\theta)$, where $p(\theta)$ is the a priori probability of θ.

When the a priori probability of θ depends on the context, c, (assuming that x is independent of c when θ is given, $p(x/\theta,\ c) = p(x/\theta)$) Equation (1) becomes

$$p(\theta/x, c) = \frac{p(x/\theta)\ p(\theta/c)}{p(x/c)}\ , \tag{2}$$

so that the decision is based on the product

$$p(x/\theta)\ p(\theta/c)\ . \tag{3}$$

The calculation of $p(x/\theta)$ in (3) has been examined, for example, in references [2, 16, 19]. The Markov chain (discussed in Section 2.3 in relation to the calculation of $p(x/\theta)$) can be used very effectively in the calculation of $p(\theta/c)$. For example, a tabulation of trigram frequencies can be converted into a description of the language in terms of a second-order Markov chain. For such a chain the probability of each character is conditioned on its four nearest neighbors, two on each side. If θ_k denotes the k^{th} character, and if a second-order chain is assumed, then

$$p(\theta_k/\text{all the other characters}) = p(\theta_k/\theta_{k-2}, \theta_{k-1}, \theta_{k+1}, \theta_{k+2}) \ . \qquad (4)$$

For sequential (rather than compound) processing

$$p(\theta_k/\text{the \underline{previous} characters}) = p(\theta_k/\theta_{k-2}, \theta_{k-1}) \ . \qquad (5)$$

This Markovian development is less complicated than in Section 2.3, since here the chain is stationary. For an alphabet of 27 characters, a table of trigram frequencies contains about 6,000 nonzero entries out of a possible total of $27^3 \approx 20,000$ possibilities.

The trouble with the preceeding development is that, in practice, the neighboring characters are not known. The context c is only available through observations on preceding characters. Even if the character ensembles form a Markov chain, it is by no means true that $p(\theta_k/\text{the previous \underline{observations}})$ is given by (5) with θ_{k-1} and θ_{k-2} replaced by the decisions made on the last two observations. If context were used in this way, errors would tend to "propagate". What is needed is the optimum (minimum risk) sequential compound decision procedure for dependent states of nature.

4.2 THE SEQUENTIAL COMPOUND BAYES PROCEDURE FOR DEPENDENT STATES OF NATURE

The optimum (Bayes) sequential compound decision procedure for known distributions and dependent states of nature is derived below. The decision on the k^{th} state θ_k, given the first k observations

$x_1, \dots, x_k$, does not depend on the unknown values of $\theta_1, \dots, \theta_{k-1}$ nor on the decisions about them.

We shall show that when the θ_k's form a first-order Markov chain, the kth decision depends on $x_1, \dots, x_{k-1}$ only through quantities which had already been calculated in order to make the previous decision. The calculations needed to decide on the kth state will be defined recursively in terms of x_k and quantities previously calculated for the decision on θ_{k-1}. Since the first decision is simple, the procedure is well defined and easily implemented.

Let $\Omega = \{1, 2, \dots, r\}$ be a set of states of nature and $A = \{1, 2, \dots, s\}$ be a set of actions. For $i \in \Omega$, $j \in A$, L_{ij} denotes the loss incurred by action j when the state of nature is i. In a compound decision problem, there exists a vector $\underline{\theta}_N = (\theta_1, \dots, \theta_N)$ of states of nature and a corresponding vector $\underline{x}_N = (x_1, \dots, x_N)$ of random variables, where θ_k denotes the state of nature in the kth component problem, and the probability density of x_k is $p_{\theta_k}(x_k)$. For a given θ_k, x_k is independent of the other x's and θ's:

$$p(x_k/x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_N, \underline{\theta}_N) = p(x_k/\theta_k) = p_{\theta_k}(x_k) , \qquad (6)$$

and hence $p(\underline{x}_k/\underline{\theta}_k) = \prod_{i=1}^{k} p(x_i/\theta_i)$. We do not assume that the θ's are independent.

If only the first k observations, $\underline{x}_k = (x_1, \dots, x_k)$ are at hand when the kth decision must be made, one can use a sequential compound decision rule $\underline{t}_N = (t_1, \ldots, t_N)$, where $t_k = t_k(j/\underline{x}_k)$ is for each $\underline{x}_k$ a distribution over A according to which the k^{th} action is chosen. The risk for such a rule is

$$R(\underline{\theta}_N, \underline{t}_N) = \frac{1}{N} \sum_{k=1}^{N} R(\underline{\theta}_N, t_k) \ ,$$

whereby Equation (13) of Chapter 1, the k^{th} component risk is

$$R(\underline{\theta}_N, t_k) = \int \sum_{j=1}^{s} L_{\theta_k j} t_k(j/\underline{x}_k)\, p(\underline{x}_N/\underline{\theta}_N)\, dx^N$$

$$= \int \sum_{j=1}^{s} L_{\theta_k j} t_k(j/\underline{x}_k)\, p(\underline{x}_k/\underline{\theta}_k)\, dx^k = R(\underline{\theta}_k, t_k) \ . \qquad (7)$$

We assume that $p_i(x)$ is known for all $i \in \Omega$, but that none of the θ'_ks is known.

The compound Bayes risk with respect to an a priori distribution $G(\underline{\theta}_N)$ over Ω^N is

$$\overline{R}(G, \underline{t}_N) = \sum_{\underline{\theta}_N \in \Omega^N} R(\underline{\theta}_N, \underline{t}_N)\, G(\underline{\theta}_N) = \frac{1}{N} \sum_{k=1}^{N} \overline{R}(G, t_k) \ ,$$

where

$$\overline{R}(G, t_k) = \sum_{\underline{\theta}_N \in \Omega^N} R(\underline{\theta}_N, t_k)\, G(\underline{\theta}_N) \ .$$

A procedure is compound Bayes against G when it minimizes $\overline{R}(G, \underline{t}_N)$. Thus, the sequential compound Bayes procedure $\underline{t}_N{}^G$ is the one that minimizes the k^{th} component Bayes risk, for every k,

$$\bar{R}(G, t_k) = \sum_{\underline{\theta}_N} R(\underline{\theta}_k, t_k)\, G(\underline{\theta}_N) = \sum_{\underline{\theta}_k} R(\underline{\theta}_k, t_k)\, G(\underline{\theta}_k)$$

$$\tag{8}$$

$$= \int \sum_{j=1}^{s} \sum_{\underline{\theta}_k} L_{\theta_k j}\, p(\underline{x}_k/\underline{\theta}_k)\, G(\underline{\theta}_k)\, t_k(j/\underline{x}_k)\, dx^k ,$$

where $G(\underline{\theta}_k)$ is the (marginal) a priori distribution over Ω^k. Hence $t_k{}^G(j/\underline{x}_k) = 1$ for that j which minimizes the quantity

$$Q = \sum_{\underline{\theta}_k} L_{\theta_k j}\, p(\underline{x}_k/\underline{\theta}_k)\, G(\underline{\theta}_k) . \tag{9}$$

Theoretically, the problem is solved. Practically, it is not. We have a sum of r^k terms,where k may be in the thousands.. Letting $p(\underline{x}_k, \underline{\theta}_k) = p(\underline{x}_k/\underline{\theta}_k)\, G(\underline{\theta}_k)$,

$$Q = \sum_{\underline{\theta}_k} L_{\theta_k j}\, p(\underline{x}_k, \underline{\theta}_k) = \sum_{\theta_k} L_{\theta_k j}\, p(\underline{x}_k, \theta_k) . \tag{10}$$

We note that for the special case where action j corresponds to deciding that $\theta = j$ and $L_{\theta j} = \begin{cases} 1 \text{ if } \theta \neq j \\ 0 \text{ if } \theta = j \end{cases}$, $t_k{}^G$ chooses the value of θ_k that maximizes $p(\underline{x}_k, \theta_k)$. This is equivalent to maximizing the a posteriori probability

$$G(\theta_k/\underline{x}_k) = \frac{p(\underline{x}_k,\ \theta_k)}{p(\underline{x}_k)} ,$$

since the denominator, $p(\underline{x}_k) = \sum_{\underline{\theta}_k} p(\underline{x}_k/\underline{\theta}_k)\, G(\underline{\theta}_k)$ is independent of θ_k.

In Equation (10) (making repeated use of (6)) ,

$$p(\underline{x}_k, \theta_k) = p(x_k/\theta_k)\, p(\underline{x}_{k-1}, \theta_k) = p(x_k/\theta_k) \sum_{\underline{\theta}_{k-1}} p(\underline{x}_{k-1}, \underline{\theta}_k)$$

$$= p(x_k/\theta_k) \sum_{\underline{\theta}_{k-1}} p(\underline{x}_{k-1}/\underline{\theta}_{k-1})\, G(\underline{\theta}_k) \tag{11}$$

$$= p(x_k/\theta_k) \sum_{\underline{\theta}_{k-1}} G(\theta_k/\underline{\theta}_{k-1})\, p(\underline{x}_{k-1}, \underline{\theta}_{k-1}) .$$

If the states of nature form a Markov chain,

$$G(\theta_k/\underline{\theta}_{k-1}) = G(\theta_k/\theta_{k-1}) \quad , \tag{12}$$

with known transition probabilities $G(\omega/\nu)$, for ω and ν in Ω, then

$$p(\underline{x}_k, \theta_k) = p_{\theta_k}(x_k) \sum_{\theta_{k-1}=1}^{r} G(\theta_k/\theta_{k-1})\, p(\underline{x}_{k-1}, \theta_{k-1}) . \tag{13}$$

Note that $p_i(x)$ and $G(i/\nu)$ are known and $p(\underline{x}_{k-1}, \theta_{k-1})$ is defined recursively in terms of $p(x_1, \theta_1) = p_{\theta_1}(x_1)G(\theta_1)$. The kth decision $t_k{}^G(j/\underline{x}_k)$, depends on x_k and a function of θ_{k-1} which had already been calculated in order to make the previous decision. It does not depend on the (unknown) value of θ_{k-1} nor on the decision about it.

We incidentally note that if the states of nature are independent, $G(\theta_k/\underline{\theta}_{k-1}) = q(\theta_k)$, then

$$p(\underline{x}_k, \theta_k) = p_{\theta_k}(x_k)\ q(\theta_k)\ p(\underline{x}_{k-1})\ .$$

Since $p(\underline{x}_{k-1})$ does not involve θ_k, we have a simple decision problem; the quantity to be minimized is $\sum_{i=1}^{r} L_{ij}\ p_i(x)\ q(i)$.

Example: For a 2 x 2 loss matrix $A = \Omega = \{0, 1\}$ (a binary alphabet), with $w = (L_{10} - L_{11})/L_{01}/L_{00})$, we decide $\theta_k = 1$ if $wp(\underline{x}_k, 1) > p(\underline{x}_k, 0)$, where (when the θ_k's form a Markov chain)

$$p(\underline{x}_k, i) = p_i(x_k)\ \left[G(i/0)\ p(\underline{x}_{k-1}, 0) + G(i/1)\ p(\underline{x}_{k-1}, 1)\right],\ i=0, 1;\ k=2, 3 \ldots$$

and $p(\underline{x}_1, i) = p_i(x_1)\ G(i)$.

4.3 SEQUENTIAL RULES FOR MARKOV-CHAIN DEPENDENCE

For a non-randomized decision, let $d_k(\underline{x}_k)$ denote the value of $j \in A$ for which $t_k(j/\underline{x}_k) = 1$; the decision function d_k maps S^k into A, i.e. $d_k(\underline{x}_k) = j \in A$. Such a function d_k determines (and can be considered as) a partition of S^k into mutually exclusive sets $\Gamma_k(j) = \{\underline{x}_k : d_k(\underline{x}_k) = j\} = \{\underline{x}_k : t(j/\underline{x}_k) = 1\}$ whose union is S^k. If $\underline{x}_k$ is an element of $\Gamma_k(j)$, action j is taken in the kth problem.

For the square loss matrix $L_{ij} = 1 - \delta_{ij}$ (action j corresponds to deciding that $\theta_k = j$) the kth component Bayes risk

$$\bar{R}(G, t_k) = \int_{S^k} \sum_{j=1}^{s} \sum_{\theta_k} L_{\theta_k j}\ p(\underline{x}_k, \theta_k)\ t(j/\underline{x}_k)\ dx^k$$

becomes the probability of error, $e_k = \Pr\{d_k \neq \theta_k\}$:

$$e_k = \bar{R}(G, t_k) = 1 - \sum_{j=1}^{r} \int_{\Gamma_k(j)} p(\underline{x}_k, j)\, dx^k ,$$

where $p(\underline{x}_k, \cdot)$ for a first-order Markov chain is given by

$$p(\underline{x}_k, \theta_k) = p(x_k/\theta_k) \sum_{\theta_{k-1}} G(\theta_k/\theta_{k-1})\, p(\underline{x}_{k-1}, \theta_{k-1}) . \qquad (13)$$

The error probability is minimized by choosing $d_k(\underline{x}_k)$ equal to the value of θ_k that maximizes the expression in Equation(13), that is,

$$\Gamma_k(j) = \left\{ \underline{x}_k : p(\underline{x}_k, j) \geq p(\underline{x}_k, \theta_k) \text{ for all } \theta_k \epsilon \Omega \right\}.$$

Since $p(x_k, \theta_k/\underline{x}_{k-1}) = p(\underline{x}_k, \theta_k)/p(\underline{x}_{k-1})$, this rule is equivalent to choosing $d_k(\underline{x}_k)$ equal to the value of θ_k that maximizes

$$p(x_k, \theta_k/\underline{x}_{k-1}) \doteq p(x_k/\theta_k)\, G(\theta_k/\underline{x}_{k-1}) . \qquad (14)$$

An alternate (sub-optimum) rule (suggested in Section 2.1) would be to choose $d_k^{\dagger}(\underline{x}_k)$ as equal to the value of θ_k that maximizes

$$p(x_k, \theta_k/d^{\dagger}_{k-1}\,(\underline{x}_{k-1})\,) = p(x_k/\theta_k)\, G(\theta_k/d^{\dagger}_{k-1}) . \qquad (15)$$

Letting e_k and $e_k^{\dagger}$ represent the error probabilities for d_k and $d_k^{\dagger}$ respectively, we have $e_k \leq e_k^{\dagger}$. Let e_1 represent the error probability for the simple rule $t_1(j/x_k)$,which does not take context into account but merely chooses $d_1(x_k)$ equal to the value of θ_k that maximizes

$$p(x_k, \theta_k) = p(x_k/\theta_k)\; G(\theta_k)\; . \qquad (16)$$

We note that for k = 1, $d_1^{\dagger}(\underline{x}_1) = d_1(\underline{x}_1) = d_1(x_1)$ and hence $e_1^{\dagger} = e_1$.

By expanding the class of possible decision functions, error probabilities less than e_k can be attained. For example, if d_k is allowed to depend on $\underline{\theta}_k$ as well as $\underline{x}_k$, the minimum error probability of a rule $d_k(\underline{\theta}_k, \underline{x}_k)$ is zero,as is easily seen by letting $d_k(\underline{\theta}_k, \underline{x}_k) = \theta_k$.

We define the general k$^{\text{th}}$-component risk as

$$\overline{R}(G, t_k) = \int_{S^N} \sum_{j=1}^{s} \sum_{\underline{\theta}_N} L_{\theta_k j}\; t_k(j/\underline{\theta}_N, \underline{x}_N)\; p(\underline{x}_N, \underline{\theta}_N)\; dx^N \qquad (17)$$

where $t_k(j/\underline{\theta}_N, \underline{x}_N)$ is a general decision rule. For the case where $A = \Omega$ and $L_{ij} = 1 - \delta_{ij}$, $\overline{R}(G, t_k)$ is the probability of error for the k$^{\text{th}}$ problem:

$$\overline{R}(G, t_k) = \int_{S^N} \sum_{\underline{\theta}_N} \sum_{j=1}^{r} t_k(j/\underline{\theta}_N, \underline{x}_N)\; p(\underline{x}_N, \underline{\theta}_N)\; dx^N$$

$$- \int_{S^N} \sum_{\underline{\theta}_N} \sum_{j=1}^{r} \delta_{\theta_k j}\; t_k(j/\underline{\theta}_N, \underline{x}_N)\; p(\underline{x}_N, \underline{\theta}_N)\; dx^N \qquad (18)$$

$$= 1 - \int_{S^N} \sum_{\underline{\theta}_N} t_k(\theta_k/\underline{\theta}_N, \underline{x}_N)\; p(\underline{x}_N, \underline{\theta}_N)\; dx^N .$$

For a non-randomized decision rule, $t_k(\theta_k/\underline{\theta}_N, \underline{x}_N)$ is the characteristic function of the set of $(\underline{\theta}_N, \underline{x}_N) \in \Omega^N \times S^N$ for which $d_k = \theta_k$. Since the average value of the characteristic function of a set is the probability of the set:

$$\bar{R}(G, t_k) = 1 - \Pr\left\{d_k = \theta_k\right\}. \tag{19}$$

By identifying a randomized decision rule t_k with an equivalent probability distribution over the class of decision functions $d_k(\underline{\theta}_N, \underline{x}_N)$,

$$t_k(j/\underline{\theta}_N, \underline{x}_N) = \Pr\left\{d_k = j/\underline{\theta}_N, \underline{x}_N\right\},$$

the same result

$$\bar{R}(G, t_k) = \Pr\left\{d_k \neq \theta_k\right\} \tag{20}$$

is obtained.

We define a general sequential decision rule by $t_k(j/\underline{\theta}_N, \underline{x}_N) = t_k(j/\theta_{k-1}, \underline{x}_k)$. For such sequential rules, the risk becomes

$$\bar{R}(G, t_k) = \int_{S^k} \sum_{j=1}^{s} \sum_{\underline{\theta}_k} L_{\theta_k j}\, p(\underline{x}_k, \underline{\theta}_k)\, t_k(j/\underline{\theta}_{k-1}, \underline{x}_k)\, dx^k . \tag{21}$$

Obviously,

$$\min_{t_k(j/\underline{\theta}_{k-1}, \underline{x}_k)} \bar{R}(G, t_k) \leq \min_{t_k(j/\underline{x}_k)} \bar{R}(G, t_k) \quad , \tag{22}$$

where the left-and right-hand sides are the k^{th}-component sequential Bayes envelopes for known and unknown $\underline{\theta}_{k-1}$, respectively.

Since

$$\bar{R}(G, t_k) = \int_{S^k} \sum_{\underline{\theta}_{k-1}} G(\underline{\theta}_{k-1}) \sum_{j=1}^{s} \sum_{\theta_k} L_{\theta_k j}\, p(\underline{x}_k, \theta_k/\underline{\theta}_{k-1})\, t_k(j/\underline{\theta}_{k-1}, \underline{x}_k)\, dx^k$$

$$= \sum_{\underline{\theta}_{k-1}} G(\underline{\theta}_{k-1}) \int_{S^k} p(\underline{x}_{k-1}/\underline{\theta}_{k-1}) \sum_{j=1}^{s} \sum_{\theta_k=1}^{r} L_{\theta_k j}\, p(x_k/\underline{\theta}_{k-1})$$

$$\cdot\, t_k(j/\underline{\theta}_{k-1}, \underline{x}_k)\, dx^k \,, \tag{23}$$

the sequential Bayes rule for known $\underline{\theta}_{k-1}$ has $t_k(j/\underline{\theta}_{k-1}, \underline{x}_k) =$ $t_k(j/\underline{\theta}_{k-1}, x_k) = 1$ for the value of j that minimizes

$$\sum_{\theta_k = 1}^{r} L_{\theta_k j}\, p(x_k, \theta_k/\underline{\theta}_{k-1}) = \sum_{\theta_k=1}^{r} L_{\theta_k j}\, p(x_k/\theta_k)\, G(\theta_k/\underline{\theta}_{k-1}) \,. \tag{24}$$

For the case of a Markov chain, $t_k(j/\underline{\theta}_{k-1}, \underline{x}_k) = t_k(j/\theta_{k-1}, x_k) = 1$ for the value of j that minimizes

$$\sum_{\theta_k} L_{\theta_k j}\, p(x_k/\theta_k)\, G(\theta_k/\theta_{k-1}) \,. \tag{25}$$

For this t_k (call it $t_k{}^*$)

$$\bar{R}(G, t_k^*) = \sum_{\theta_{k-1}} G(\theta_{k-1}) \int_S \sum_{j=1}^{s} \sum_{\theta_k=1}^{r} L_{\theta_k j}\, p(x_k/\theta_k)\, G(\theta_k/\theta_{k-1})$$

$$\cdot\, t_k^*(j/\theta_{k-1}, x_k)\, dx_k$$

$$= \sum_{\nu=1}^{r} G(\nu) \int_S \sum_{j=1}^{s} \sum_{i=1}^{r} L_{ij}\, p_i(x)\, G(i/\nu)\;\; t(j/\nu, x)\, dx \,, \tag{26}$$

where $t(j/\nu, x) = 1$ for the value of j that minimizes

$$\sum_i L_{ij}\, p_i(x)\, G(i/\nu)\ .$$

Hence, when the states of nature form a stationary Markov chain, the k^{th} component sequential Bayes envelope for known $\underline{\theta}_{k-1}$ is independent of k.

For the case where $A = \Omega$ and $L_{ij} = 1 - \delta_{ij}$, Equation (22) becomes

$$e_c \leq e_k \quad , \tag{27}$$

where e_c is the minimum probability of error for the k^{th} decision when $\underline{\theta}_{k-1}$ is known. The sequential Bayes rule for known $\underline{\theta}_{k-1}$ is nonrandomized, with $d_k(\underline{\theta}_{k-1}, \underline{x}_k) = d_k(\theta_{k-1}, x_k)$ independent of $\underline{x}_{k-1}$ and $\underline{\theta}_{k-2}$ and equal to the value of θ_k that maximizes

$$p(x_k, \theta_k/\theta_{k-1}) = p(x_k/\theta_k)\ G(\theta_k/\theta_{k-1})\ . \tag{28}$$

Now e_k may be considered as the minimum error probability of the rule $t_{k+1}(j/x_2, \ldots, x_{k+1})$ (we have assumed a stationary Markov chain) and hence is not less than e_{k+1}, the minimum error probability of the rule $t_{k+1}(j/x_1, \ldots, x_{k+1})$. Hence $e_1 \geq e_2 \geq \ldots \geq e_k \geq e_{k+1} \geq e_c$, $\{e_k\}$ is a monotonically non-increasing sequence bounded below by e_c and hence has a limit e with $e_c \leq e \leq e_1$.

Using the rule of Equation (15), when the k-1st decision is correct (i.e., when $d_{k-1}^{\dagger} = \theta_{k-1}$) the probability of error is e_c. The probability that $d_{k-1}^{\dagger} = \theta_{k-1}$ is $1 - e_{k-1}^{\dagger}$. Let e_e be the probability of error when the k-1st decision is wrong. Then

$$e_k^{\dagger} = e_{k-1}^{\dagger}\, e_e + (1 - e_{k-1}^{\dagger})\, e_c \ , \tag{29}$$

and, since e_c is the minimum error probability for known θ_{k-1}, $e_c \leq e_e$.

Since $0 \leq e_c \leq e_e \leq 1$, $0 \leq e_e - e_c \leq 1 - e_c$. Thus $e_e - e_c = 1$ if and only if $e_c = 0$ and $e_e = 1$. For this case $e_k^{\dagger} = e_{k-1}^{\dagger} = \ldots = e_1$. Otherwise, $0 \leq e_e - e_c < 1$ and hence, letting $a = e_e - e_c$,

$$\begin{aligned} e_{k+1}^{\dagger} &= a\, e_k^{\dagger} + e_c \\ &= a^k\, e_1 + (1 + a + a^2 + \ldots + a^{k-1})\, e_c \\ &= a^k\, e_1 + \frac{1 - a^k}{1 - a}\, e_c \ . \end{aligned}$$

Thus $e_k^{\dagger}$ converges to a limit $e^{\dagger}$ given by

$$c^{\dagger} = \frac{e_c}{1 - (e_e - e_c)} \ . \tag{30}$$

Since $e_k \leq e_k$ for all k, $e \leq e$. Since

$$e_{k+1}^{\dagger} = (e_e - e_c)\, e_k^{\dagger} + e_c \ , \tag{31}$$

we obtain

$$e_1 - e^{\dagger} = \frac{\left[1 - (e_e - e_c)\right] e_1 - e_c}{1 - (e_e - e_c)} = \frac{e_1 - e_2^{\dagger}}{1 - (e_e - e_c)} \ . \tag{32}$$

Hence, $e^{\dagger} < e_1$ if and only if $e_2^{\dagger} < e_1$.

For the case of two classes, $\theta_k = 0$ or 1,

$$d_1(x_k) = 1 \quad \text{if} \quad \frac{p_1\,(x_k)}{p_0\,(x_k)} > \frac{1 - G(1)}{G(1)} \tag{16a}$$

$$d^{\dagger}(\underline{x}_k) = 1 \quad \text{if} \quad \frac{p_1\,(x_k)}{p_0\,(x_k)} > \frac{1 - G(1/d_{k-1}^{\dagger})}{G(1/d_{k-1})}\ , \tag{15a}$$

and by (14)

$$d_k(\underline{x}_k) = 1 \quad \text{if} \quad \frac{p_1\,(x_k)}{p_0\,(x_k)} > \frac{1 - G_k(1/\underline{x}_{k-1})}{G_k(1/\underline{x}_{k-1})}\ , \tag{14a}$$

where $G_k(1/\underline{x}_m) \equiv G(\theta_k = 1/\underline{x}_m)$. For all $k > 0$ we have

$$e_c \leq e_k \leq e_1$$

and

$$e_c \leq e_k \leq e_k^{\dagger}.$$

In the next chapter we will treat the example of normal populations and show, in Section 5.1, that when the populations are not well separated $e_k^{\dagger} > e_1$ for all $k > 1$, while when they are well separated $e_k^{\dagger} < e_1$.

4.4 NON-SEQUENTIAL COMPOUND RULES

A procedure is compound Bayes if it minimizes the k^{th}-component Bayes risk,

$$\overline{R}(G, t_k) = \sum_{\underline{\theta}_N} R(\underline{\theta}_N, t_k)\, G(\underline{\theta}_N)$$

$$= \int \sum_{j=1}^{s} \sum_{\underline{\theta}_N} L_{\theta_k j}\, t_k(j/\underline{x}_N)\, p(\underline{x}_N/\underline{\theta}_N)\, G(\underline{\theta}_N)\, dx^N, \qquad (33)$$

for each k. In Section 4.2, we derived the Sequential Compound Bayes procedure. The compound Bayes procedure choses $t_k(j/\underline{x}_N)$ equal to one for that j which minimizes

$$\sum_{\underline{\theta}_N} L_{\theta_k j}\, p(\underline{x}_N/\underline{\theta}_N)\, G(\underline{\theta}_N) = \sum_{\underline{\theta}_N} L_{\theta_k j} p(\underline{x}_N, \underline{\theta}_N) = \sum_{\theta_k} L_{\theta_k j}\, p(\underline{x}_N, \theta_k)\ . \qquad (34)$$

We note that the minimum N^{th}-component Bayes risk is the same in the sequential and non-sequential case. The minimum k^{th}-component risk for $k < N$ for the non-sequential case is less than or equal to the corresponding minimum for the sequential case.

Let us consider the case $N = 2$, $A = \Omega$, and $L_{ij} = 1 - \delta_{ij}$. The k^{th}-component Bayes risk is equal to the probability of error

$$\overline{R}(G, t_k) = e_k \qquad (35)$$

and the compound Bayes risk is equal to the average probability of error

$$\overline{R}(G, \underline{t}_2) = \frac{e_1 + e_2}{2} = \langle e \rangle\ . \qquad (36)$$

The compound Bayes procedure chooses d_1 equal to the value of θ_1 that maximizes

$$p(\underline{x}_2, \theta_1) = p(x_1/\theta_1) \sum_{\theta_2} p(x_2/\theta_2)\; G(\theta_1, \theta_2) \tag{37}$$

and d_2 equal to the value of θ_2 that maximizes

$$p(\underline{x}_2, \theta_2) = p(x_2/\theta_2) \sum_{\theta_1} p(x_1/\theta_1)\; G(\theta_1, \theta_2)\;. \tag{38}$$

We note that d_2 is the same as in the sequential case. From (33), the component risks are

$$e_k = 1 - \sum_{\underline{\theta}_2} \int_{S^2} t_k(\theta_k/\underline{x}_2)\; p(\underline{x}_2, \underline{\theta}_2)\; dx^2\;. \tag{39}$$

In particular

$$e_2 = 1 - \sum_{\theta_2} \int_{S^2} t_2(\theta_2/\underline{x}_2)\; p(\underline{x}_2, \theta_2)\; dx^2\;. \tag{40}$$

If $G(\theta_1, \theta_2) = G(\theta_2, \theta_1)$, then $e_1 = e_2 = \langle e \rangle$ for the compound Bayes procedure.

Let e_k be the minimum error probability for the k^{th} decision using a sequential compound procedure. The average error probability (the compound Bayes risk) is then $\langle e_N \rangle = 1/N \sum_{k=1}^{N} e_k$. If $G(\underline{\theta}_N)$ is symmetric, the minimum error probability for the k^{th} decision using a compound procedure is $e_k^{*} = e_N$ for all k and hence $\langle e_N^{*} \rangle = e_N$. In such a case, $G(\underline{\theta}_N)$ is stationary, and as in Section 4.3, e_k is monotonically non-increasing with the limit e. Therefore $<e_N{}^*>$ and $<e_N>$ are monotonically non-increasing with the same limit, e. In general, however, $e_k{}^* \leq e_k$ and hence, $<e_N{}^*> \leq <e_N>$.

In the next chapter we analyze, in more detail, the error probabilities discussed in this chapter.

Section V

THE TWO CLASS PROBLEM WITH NORMAL DISTRIBUTIONS

5.1 SEQUENTIAL RULES FOR MARKOV-CHAIN DEPENDENCE

Let $p_1(x)$ and $p_0(x)$ be univariate normal densities with the same variance, σ^2, and with a difference in means, $\mu = \mu_1 - \mu_0 > 0$. Thus

$$p_i(x) = \phi\left[(x-\mu_i)/\sigma\right] = \frac{1}{\sqrt{2\pi}\,\sigma} \; e^{-\frac{(x-\mu_i)^2}{2\sigma^2}} \, ,$$

and there is no loss of generality in assuming $\mu_0 = 0$ and $\mu_1 = \mu$. The set T in Equation (7) of Chapter 1 is given by

$$T(q) = \left\{ x: \; \frac{p_1(x)}{p_0(x)} > \frac{1-q}{q} \right\} = \left\{ x: x > c \right\} \, ,$$

where

$$c = (\sigma^2/\mu) \log\left[(1-q)/q\right] + \mu/2 \, .$$

Then

$$P_i(T) = \int_c^{\infty} \phi\left[(x-\mu_i)/\sigma\right] \, dx = 1 - \Phi\left[(c-\mu_i)/\sigma\right]$$

$$P_0(T) = 1-\Phi(c/\sigma) = 1-\Phi\left[(\sigma/\mu) \log\left((1-q)/q \right) + \mu/2\sigma\right]$$

$$P_1(T) = 1-\Phi\left[(c-\mu)/\sigma\right] = 1-\Phi\left[(\sigma/\mu) \log\left((1-q)/q \right) - \mu/2\sigma\right] .$$

Letting $m = \mu/\sigma$, the probability of error given that $\theta = 1$ is

$$\gamma(q) = 1-P_1(T) = \Phi\left[(1/m) \log\left((1-q)/q \right) - m/2\right] ,$$

and the probability of error given that $\theta = 0$ is

$$P_0(T) = \Phi\left[(1/m)\log(q/(1-q)) - m/2\right] = \zeta(1-q) .$$

We consider the Markov chain in Section 4.3. Let $q = G(1)$, $\alpha = G(1/0)$, and $\beta = G(1/1)$. Since

$$q = G(1) = G(1/0)\,G(0) + G(1/1)\,G(1) = \alpha(1-q) + \beta q ,$$

we have

$$q = \frac{\alpha}{1-\beta+\alpha}$$

and

$$1-q = \frac{1-\beta}{1-\beta+\alpha} .$$

From Equation (9) of Chapter 1 and Equation (23) of Chapter 4

$$e_1 = \bar{R}(q, t^q) = R(q)$$

$$e_c = (1-q)\,\bar{R}(\alpha, t^\alpha) + q\,\bar{R}(\beta, t^\beta)$$

$$e_e = (1-q)\,\bar{R}(\alpha, t^\beta) + q\,\bar{R}(\beta, t^\alpha) ,$$

where

$$\bar{R}(q, t^{q'}) = q\,\zeta(q') + (1-q)\,\zeta(1-q') .$$

We shall assume $\alpha = 1-\beta$ so that $q = 1/2$. Then

$$e_1 = \zeta(1/2) \qquad (1)$$

$$e_c = \alpha \zeta(\alpha) + (1-\alpha) \zeta(1-\alpha) \tag{2}$$

and

$$e_e = (1-\alpha) \zeta(\alpha) + \alpha \zeta(1-\alpha) . \tag{3}$$

From (2) and (3) we see that $e_c(\alpha) = e_c(1-\alpha)$, $e_e(\alpha) = e_e(1-\alpha)$, and

$$e_e - e_c = (1-2\alpha) \left[\zeta(\alpha) - \zeta(1-\alpha) \right] , \tag{4}$$

where

$$\zeta(\alpha) = \Phi \left[(1/m) \log ((1-\alpha)/\alpha) - m/2 \right] . \tag{5}$$

and

$$\zeta(1-\alpha) = 1-\Phi \left[(1/m) \log ((1-\alpha)/\alpha) + m/2 \right] . \tag{6}$$

The case where $p_1(x)$ and $p_0(x)$ are multivariate normal densities with the same covariance matrix but different mean vectors is the same, but with $m = \sigma_\ell$ defined by Blackwell and Girshick [2], page 158.

From (1) we calculate

$$e_1 = 1-\Phi(m/2), \tag{7}$$

and using (5) and (6) in (2) and (3) we compute $e_c(\alpha)$ and $e_e(\alpha)$. We compute $e^{\dagger}(\alpha)$ and $e_k^{\dagger}(\alpha)$ from Equations (30) and (31) of Chapter 4 with $e_1^{\dagger} = e_1$. For $\alpha = 1/2$ we have $e_c = e_1 = e_k^{\dagger} = e^{\dagger} = e_e$, while for $\alpha = 0$ (or $\alpha = 1$) we have $e_c = 0$, $e_1 = e_k^{\dagger} = e^{\dagger}$, and $e_e = 1$. For all other α, the results are highly dependent on the separation, m.

For extremely small α (as well as for small m), we have

$$m/2 << (1/m)\left[\log\ (1-\alpha)/\alpha\right] ,$$

so that

$$\zeta(\alpha) \approx 1 - \zeta(1-\alpha).$$

Thus

$$e_c \approx \alpha + (1-2\alpha)\ \zeta(1-\alpha)$$

and

$$e_e - e_c \approx 1 - 2e_c.$$

Hence for $\alpha << 10^{-m^2}$ (very roughly),

$$e^{\dagger} \approx \frac{e_c}{1-(1-2e_c)} = \frac{1}{2}$$

and

$$e_2^{\dagger} \approx e_c + (1-2e_c)\ e_1 = e_1 + e_c(1-2e_1) = e_1 + \left[2\Phi(m/2)-1\right] e_c > e_1.$$

The limit $e^{\dagger}$ has a discontinuity at $\alpha = 0$ (and $\alpha = 1$) with $e^{\dagger}(0) = e_1$ and $e^{\dagger}(0+) = 0.5$. A proof of $\lim_{\alpha \to 0+} e^{\dagger}(\alpha) = 1/2$, using L'Hospitals rule, is given below:

Since $e_c(\alpha) = R(\alpha)$ and

$$\bar{R}(q, tq') = R(q') + \frac{dR(q')}{d\ q'}\ (q-q') ,$$

we have

$$e_e(\alpha) = \overline{R}(1-\alpha, t^{\alpha}) = e_c(\alpha) + (1-2\alpha)\, e_c'(\alpha) \;,$$

where $e_c'(\alpha) = de_c(\alpha)/d\alpha$. Hence, by (4)

$$e_c'(\alpha) = \zeta(\alpha) - \zeta(1-\alpha).$$

From (2)

$$e_c'(\alpha) = \zeta(\alpha) + \alpha\, \zeta'(\alpha) - \zeta(1-\alpha) - (1-\alpha)\, \zeta'(1-\alpha)$$

and hence

$$\alpha\, \zeta'(\alpha) = (1-\alpha)\, \zeta'(1-\alpha) \;.$$

From (4)

$$e_c'(\alpha) - e_e'(\alpha) = 2\Big[\zeta(\alpha) - \zeta(1-\alpha)\Big] - (1-2\alpha)\Big[\zeta'(\alpha) + \zeta'(1-\alpha)\Big]$$

$$= 2e_c'(\alpha) - \frac{1-2\alpha}{1-\alpha}\, \zeta'(\alpha).$$

From Equation (30) of Chapter 4

$$\lim_{\alpha \to 0+} e^{+}(\alpha) = \frac{e_c'(0)}{e_c'(0) - e_e'(0)} = \frac{1}{2} \;,$$

since $e_c'(0) = 1$ and $\zeta'(0) = 0$.

For large m (as well as for α very close to 1/2) such that

$$m/2 >> (1/m) \log\Big[(1-\alpha)/\alpha\Big]$$

we have from Equation (4)

$$e_e - e_c = (1-2\alpha)\left\{\Phi\left[\frac{m}{2} + \frac{1}{m}\log\frac{1-\alpha}{\alpha}\right] - \Phi\left[\frac{m}{2} - \frac{1}{m}\log\frac{1-\alpha}{\alpha}\right]\right\} << 1,$$

so that

$$e^{\dagger} = \frac{e_c}{1-(e_e - e_c)} \approx e_c$$

and even

$$e_2^{\dagger} = e_c + (e_e - e_c)e_1 \approx e_c.$$

In Figures 3 through 6 we see (roughly)that while $e_c \leq e_1 \leq e_k^{\dagger} \leq e^{\dagger} \leq e_e$ when $m = 1$, $e_c \leq e^{\dagger} \leq e_k^{\dagger} < e_1 \leq e_e$ when $m = 2$ or 4. We further note that for $m = 4$ convergence of $e_k^{\dagger}$ to $e^{\dagger}$ is extremely rapid, with $e_3^{\dagger} \approx e^{\dagger}$ for $\alpha > 0.02$. The reversal of the inequality $e_k^{\dagger} < e_1$ to $e_1 < e_k^{\dagger}$, which occurs for $\alpha < 0.2$ in Figure 4,occurs for all m, but cannot be seen in Figures 3, 5 and 6 because of the scale (for $m = 2$ the reversal occurs for $\alpha < 10^{-3}$). Figure 7 illustrates the relationship between e_1 and $e_k^{\dagger}$ as a function of m for a fixed α.

With $0 < \alpha < 1/2$ and $m > 0$, suppose

$$e_2^{\dagger} \leq e_1.$$

Then

$$e_c + (e_e - e_c)\, e_1 \leq e_1$$

$$\alpha\zeta(\alpha) + (1-\alpha)\,\zeta(1-\alpha) + \zeta(1/2)\,(1-2\alpha)\left[\zeta(\alpha) - \zeta(1-\alpha)\right] \leq \zeta(1/2)$$

$$\left[\alpha + (1-2\alpha)\,\zeta(1/2)\right]\left[\zeta(\alpha) - \zeta(1-\alpha)\right] \leq \zeta(1/2) - \zeta(1-\alpha)$$

$$\alpha + (1-2\alpha)\,\zeta(1/2) \leq$$

$$\frac{\Phi(m/2+(1/m)\log((1-\alpha)/\alpha) - \Phi(m/2)}{\Phi(m/2+(1/m)\log((1-\alpha)/\alpha) - \Phi(m/2-(1/m)\log((1-\alpha)/\alpha)}\ .$$

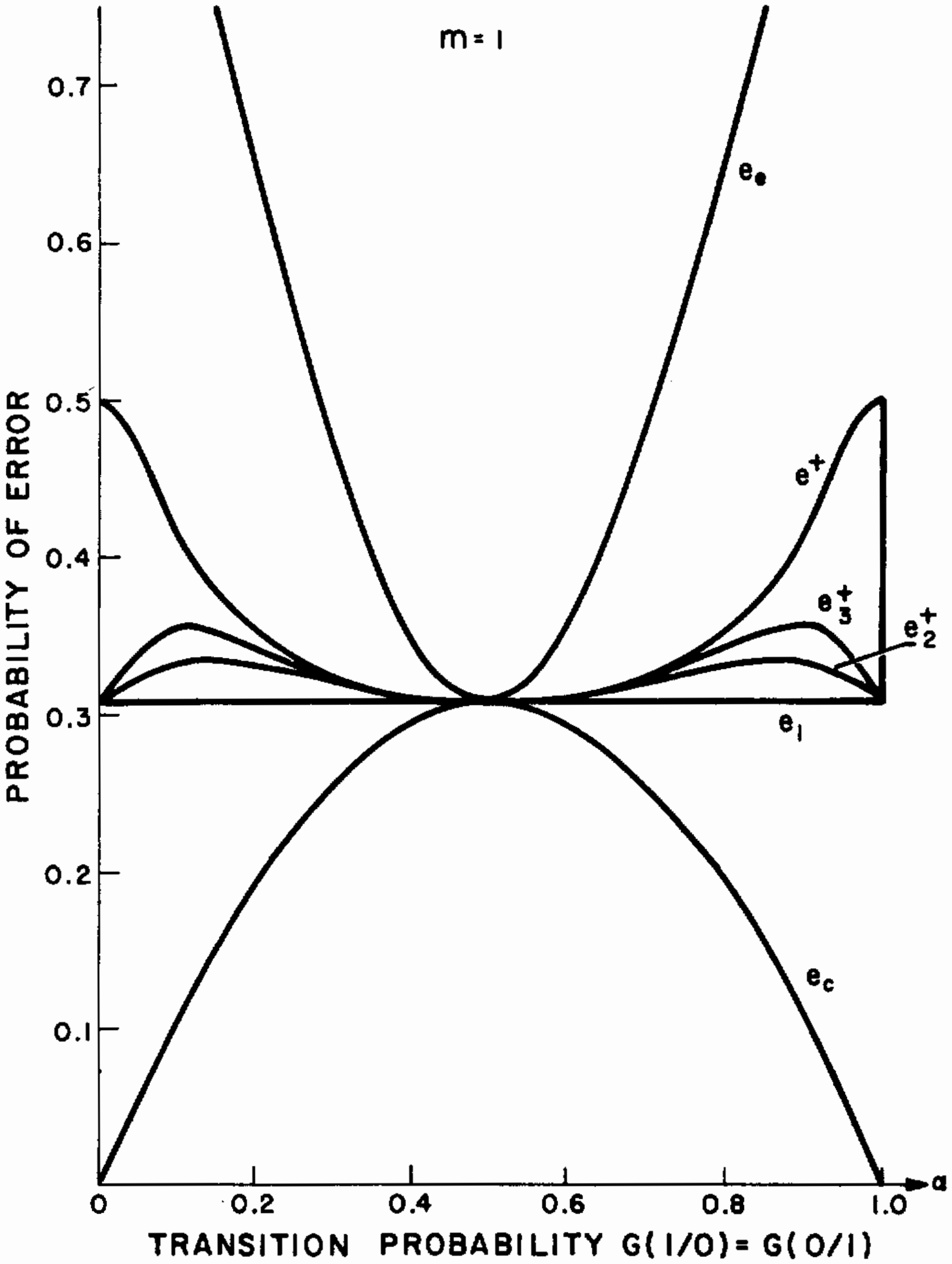

Figure 3 Error Probabilities for Normal Alternatives Displaced 1 Standard Deviation (S/N = 0 db)*

*See list of symbols on page vii.

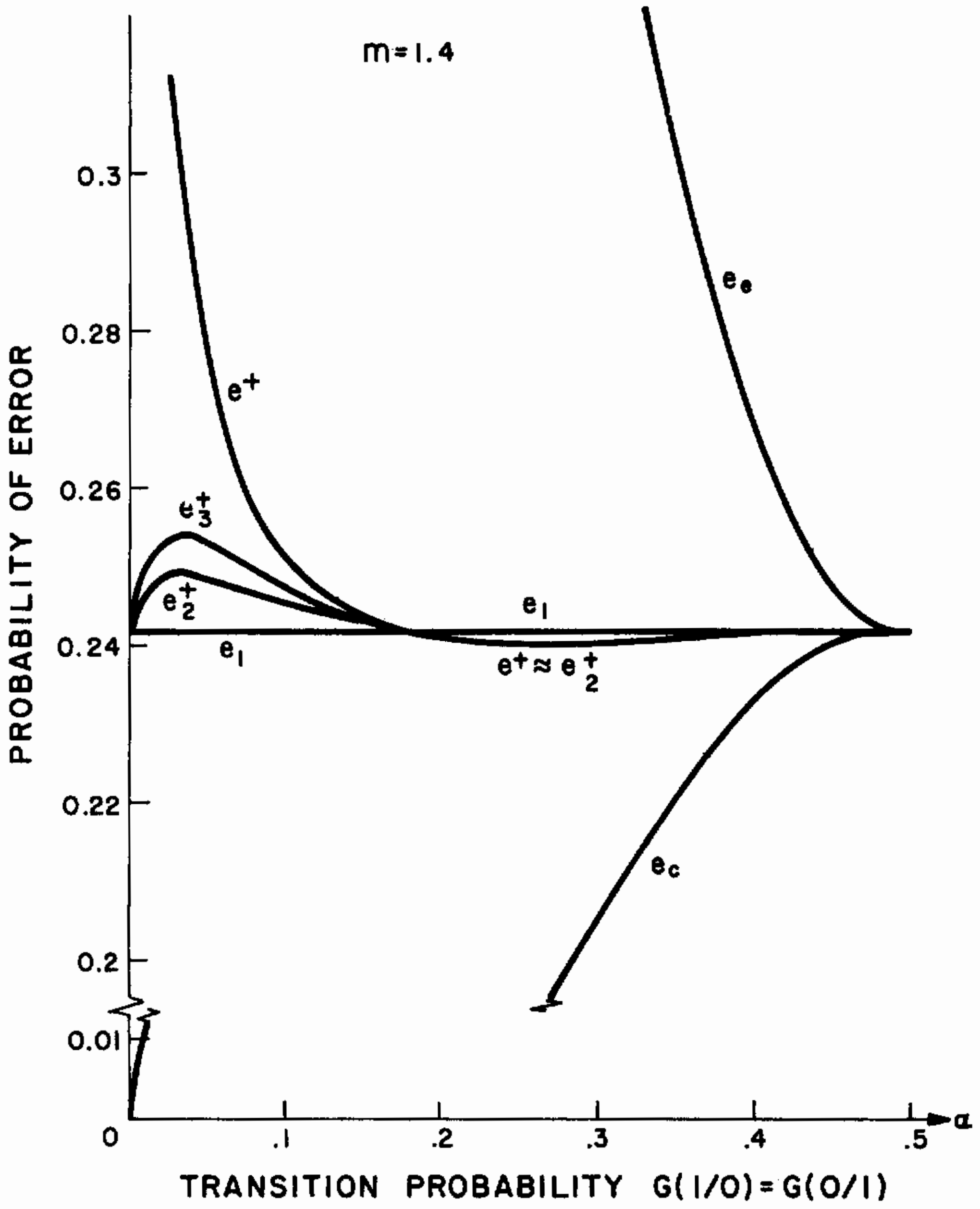

Figure 4 Error Probabilities for Normal Alternatives Displaced 1.4 Standard Deviation (S/N = 3 db)

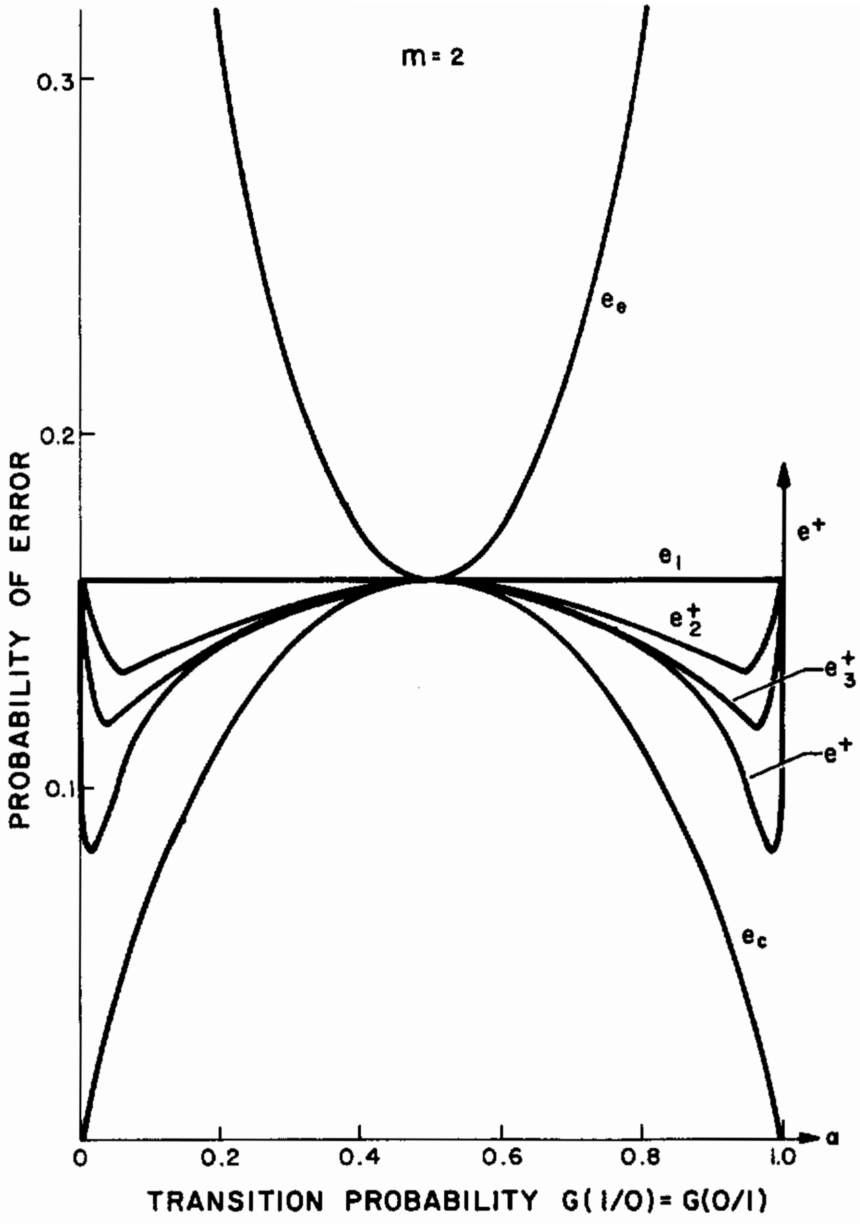

Figure 5 Error Probabilities for Normal Alternatives Displaced 2 Standard Deviations (S/N = 6 db)

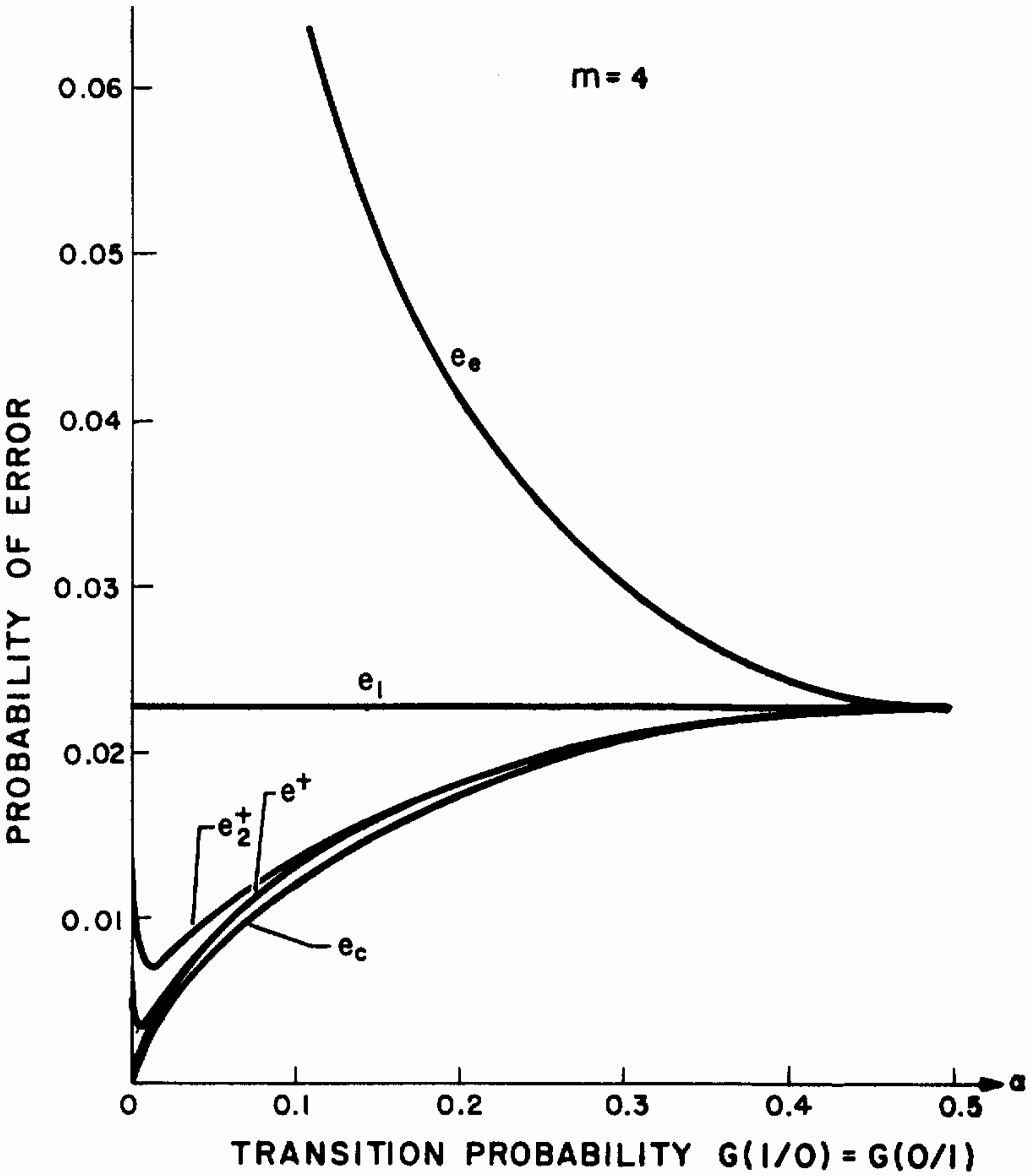

Figure 6 Error Probabilities for Normal Alternatives Displaced 4 Standard Deviations (S/N = 12 db)

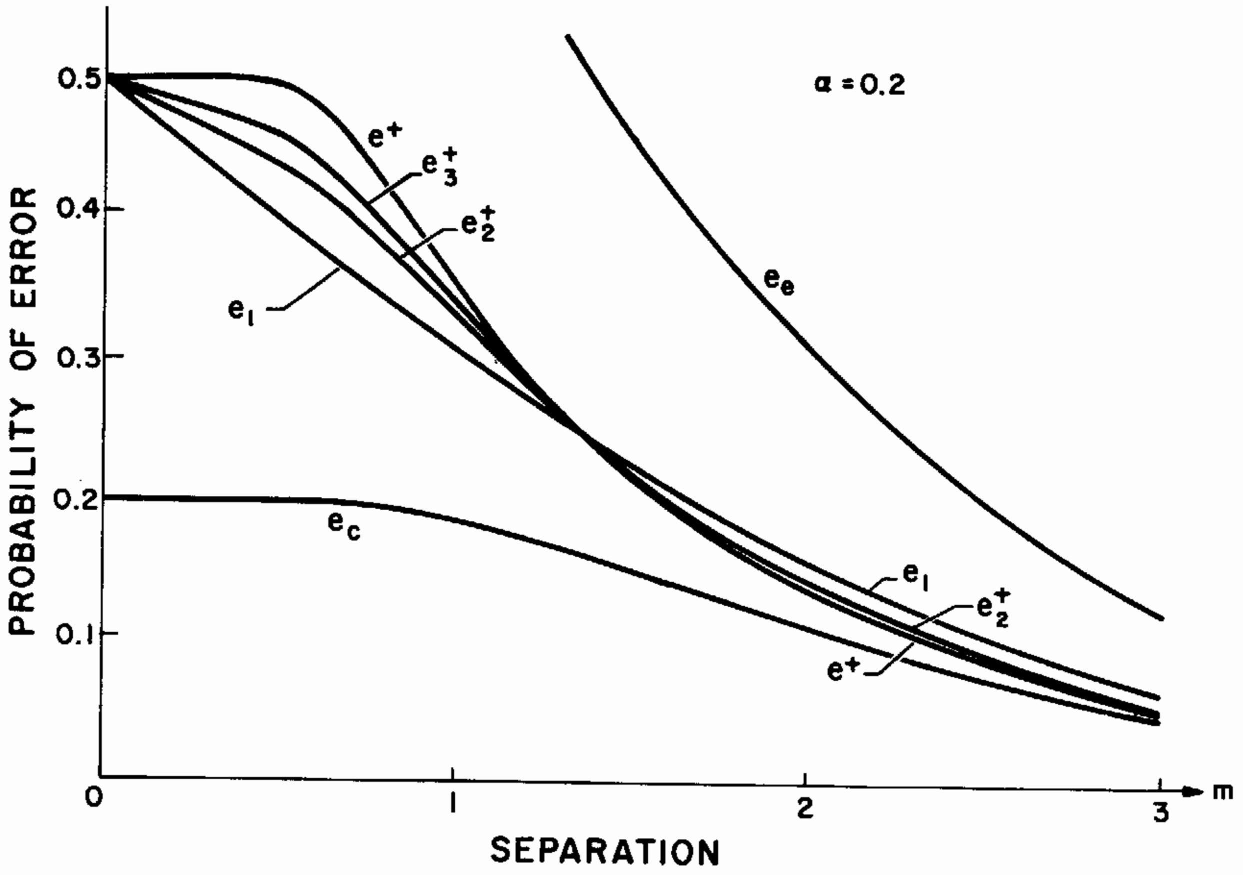

Figure 7 Error Probabilities Vs. Displacement for Normal Alternatives With a Transition Probability of 0.2

Since $\zeta(1/2) = 1-\Phi(m/2) < 1/2$,

$$\zeta(1/2) < \zeta(1/2) + \left[1-2\,\zeta(1/2)\right]\alpha = \alpha + (1-2\alpha)\,\zeta(1/2).$$

Hence

$$1-\Phi(m/2) < \frac{\Phi(m/2+b) - \Phi(m/2)}{\Phi(m/2+b) - \Phi(m/2-b)},$$

where $b = (1/m)\log\left[(1-\alpha)/\alpha\right] > 0$. For fixed α, this can occur only if m is sufficiently large. For fixed m, it can occur only if α is sufficiently large (b sufficiently small).

Thus, we see that for small m (and/or small α) the decision function of Equation (15) of Chapter 4 actually gives worse results than a rule that does not account for context. Hence, in such a situation, the use of the sequential Bayes decision function of Equation (13) or (14) of Chapter 4 is mandatory. However, if m is large (and α is not too small), the decision rule of Equation (15) (Chapter 4) is quite adequate. For m > 5 (signal to noise ratio > 14 db), $e_2^\dagger$ would be indistinguishable from e_c for any practical value of α.

5.2 THE SECOND-COMPONENT SEQUENTIAL COMPOUND BAYES RISK

For a sequential compound decision rule, $t_k(j/\underline{\theta}_N, \underline{x}_N) = t_k(j/\underline{x}_k)$, we obtain from Equation (18) of Chapter 4,

$$\begin{aligned} e_k &= 1 - \sum_{\underline{\theta}_N} \int_{S^N} t_k(\theta_k/\underline{x}_k)\, p(\underline{x}_N, \underline{\theta}_N)\, dx^N \\ &= 1 - \sum_{\theta_k} \int_{S^k} t_k(\theta_k/\underline{x}_k)\, p(\underline{x}_k, \theta_k)\, dx^k. \end{aligned} \qquad (8)$$

By Equation (13) of Chapter 4,

$$p(\underline{x}_k, \theta_k) = p(x_k/\theta_k) \sum_{\theta_{k-1}} p(\underline{x}_{k-1}/\theta_{k-1})\, G(\theta_{k-1}, \theta_k) \tag{9}$$

for the case of a Markov chain. For k = 2 Equations (8) and (9) become

$$e_2 = 1 - \sum_{\theta_2} \int_{S^k} t_2(\theta_2/\underline{x}_2)\, p(\underline{x}_2, \theta_2)\, dx^2 \tag{10}$$

and

$$p(\underline{x}_2, \theta_2) = p(x_2/\theta_2) \sum_{\theta_1} p(x_1/\theta_1)\, G(\theta_1, \theta_2) \;. \tag{11}$$

These are Equations (39) and (38) of Section 4.4. For a stationary Markov chain $G(\theta_1) = G(\theta_2)$,so that

$$G(\theta_1/\theta_2) = G(\theta_2/\theta_1)\, G(\theta_1)/G(\theta_2) = G(\theta_2/\theta_1)$$

and

$$G(\theta_1, \theta_2) = G(\theta_2/\theta_1)\, G(\theta_1) = G(\theta_1/\theta_2)\, G(\theta_2) = G(\theta_2, \theta_1) \;.$$

The following calculations give us, as e_2, both the second-component sequential compound Bayes probability of error for a stationary Markov chain (Section 4.3) and the compound Bayes probability of error for the case N = 2 with $G(\theta_1, \theta_2) = G(\theta_2, \theta_1)$ (Section 4.4). We make the assumptions of Section 5.1: $\Omega = \{0, 1\}$ and $G(\theta_1, \theta_2)$ is given by $G(0, 1) = G(1, 0) = \alpha/2$, $G(0, 0) = G(1, 1) = (1-\alpha)/2$; i.e. $G(\theta_1 = 1) = G(\theta_2 = 1) = 1/2$, $G(\theta_2 = 1/\theta_1 = 0) = \alpha$, and $G(\theta_1, \theta_2) = G(\theta_2, \theta_1)$. Letting

$$\Gamma_2(1) = \left\{\underline{x}_2 : t_2(j/\underline{x}_2) = 1\right\} = \left\{\underline{x}_2 : p(\underline{x}_2, \theta_2 = 1) > p(\underline{x}_2, \theta_2 = 0)\right\} ,$$

we have by (10),

$$e_2 = G(\theta_2=0)\int_{\Gamma_2(1)} p(\underline{x}_2/\theta_2=0)\,dx^2 + G(\theta_2=1)\int_{\Gamma_2(0)} p(\underline{x}_2/\theta_2=1)\,dx^2$$

$$= \Pr\left\{\theta_2=0\right\} \Pr\left\{d_2=1/\theta_2=0\right\} + \Pr\left\{\theta_2=1\right\} \Pr\left\{d_2=0/\theta_2=1\right\}.$$

By (11) and the assumptions

$$\Gamma_2(1) = \left\{(x1, x2) : p_1(x2)\left[\alpha p_0(x_1) + (1-\alpha)p_1(x_1)\right] > p_0(x_2)\left[(1-\alpha)p_0(x_1) + \alpha p_1(x_1)\right]\right\}$$

and

$$e_2 = \int_{\Gamma_2(1)} p(\underline{x}_2/\theta_2=0)dx^2 = \int_{\Gamma_2(1)} p(x_2/0) \sum_{\theta_1} p(x_1/\theta_1)G(\theta_1/\theta_2=0)dx^2$$

$$= \iint_{\Gamma_2(1)} p_0(x_2)\left[(1-\alpha)p_0(x_1) + \alpha\, p_1(x_1)\right] dx_1\, dx_2 .$$

Let $p_i(x)$ be univariate normal with mean μ_i and variance σ, and with $\mu = \mu_1 - \mu_0 > 0$. Letting

$$y(x_1) = \frac{(1-\alpha)\, p_0(x_1) + \alpha p_1(x_1)}{\alpha p_0(x_1) + (1-\alpha)\, p_1(x_1)} ,$$

$$\Gamma_2(1) = \left\{(x_1, x_2) : \frac{p_1(x_2)}{p_0(x_2)} > y(x_1)\right\} = \left\{(x_1, x_2) : x_2 > c(x_1)\right\} ,$$

where

$$c(x_1) = (\sigma^2/\mu) \log \left[y(x_1)\right] + \mu/2.$$

Then

$$e_2 = \int_{x_1} \left|(1-\alpha)p_0(x_1) + \alpha\, p_1(x_1)\right| \left[\int_{x_2>c(x_1)} p_0(x_2)\, dx_2\right] dx_1.$$

Since

$$\int_{x_2>c(x_1)} p_0(x_2)\, dx_2 = 1 - \Phi\left[c(x_1)/\sigma\right] = 1-\Phi\left[(\sigma/\mu)\log y(x_1) + \mu/2\sigma\right]$$

$$= \Phi\left\{(1/m)\log\left[1/y(x_1)\right] - m/2\right\},$$

where $m = \mu/\sigma$, we have

$$e_2 = \int_{-\infty}^{\infty} \left[(1-\alpha)\phi(x) + \alpha\phi(x-m)\right]\Phi\left(\frac{1}{m} \log \frac{\alpha\phi(x)+(1-\alpha)\,\phi(x-m)}{(1-\alpha)\,\phi(x) + \alpha\phi\,(x-m)} - \frac{m}{2}\right) dx. \tag{12}$$

For $\alpha = 1/2$, $e_2 = \Phi(-m/2) = 1-\Phi(m/2)$ as in Section 5.1. For $\alpha = 0$,

$$e_2 = \int_{-\infty}^{\infty} \phi(x)\ \Phi\left(\frac{1}{m} \log \frac{\phi(x-m)}{\phi(x)} - \frac{m}{2}\right) dx.$$

Since

$$\log \frac{\phi(x-m)}{\phi(x)} = \frac{2mx - m^2}{2},$$

$$e_2 = \int_{-\infty}^{\infty} \phi(x)\ \Phi(x-m)\ dx = \Phi(-m/\sqrt{2}) = 1-\Phi(m/\sqrt{2})$$

for $\alpha = 0$.

Considering the problem of Section 4.4 where $N = 2$, for $\alpha = 1/2$, both the compound and the sequential compound Bayes procedure give the same results as a simple Bayes procedure, namely an average error probability of $e_1 = \Phi(-m/2)$. At the other extreme, $\alpha = 0$, the compound Bayes rule for $N = 2$ gives an average probability of error of $e_2 = \Phi(-\sqrt{2}m/2)$, an increase in signal-to-noise ratio of 3 db. The sequential Bayes rule gives an average probability of error equal to the average of the two, $<e> = (e_1+e_2)/2$. Figure 8 shows e_1 and e_2 as a function of m for $\alpha = 0$.

5.3 NON-SEQUENTIAL RULES FOR MARKOV-CHAIN DEPENDENCE

Consider the Markov-chain problem of Section 4.3. We bounded the probability of error for a sequential rule by

$e_c \leq e_k \leq e_l$, where

$$e_c = \sum_{\theta_{k-1}} G(\theta_{k-1}) \min \Pr\left\{d_k \neq \theta_k/\theta_{k-1}\right\} \qquad (13)$$

$$= \sum_{i \in \Omega} q(i)\, R(\alpha_i)$$

with $\alpha_i = G(\theta_k/\theta_{k-1} = i)$. With the assumptions $\Omega = \{0, 1\}$, $\alpha = G(1/0) = G(0/1)$ we obtained

$$e_c = \frac{1}{2}\; R(\alpha) + \frac{1}{2}\; R(1-\alpha)\ .$$

For the normal case treated in Section 5.1, $R(1-\alpha) = R(\alpha)$ and hence $e_c = R(\alpha)$.

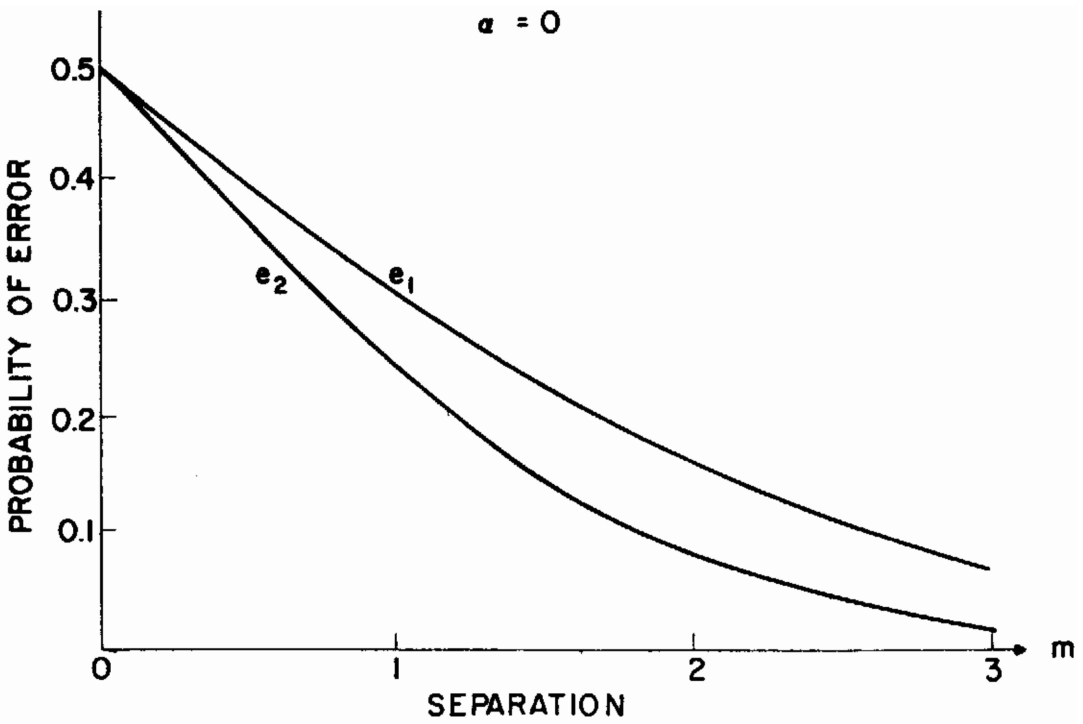

Figure 8 Error Probabilities Vs. Displacement for Normal Alternatives with Zero Transition Probability

Similarly, we bound the probability of error for a non-sequential compound rule for the case of a stationary Markov chain by $e_c^* \leq e_k^* \leq e_k$. In this case*

$$e_c^* = \sum_{(\theta_{k-1},\theta_{k+1})} G(\theta_{k-1},\theta_{k+1}) \min \Pr\left\{d_k \neq \theta_k/\theta_{k-1},\theta_{k+1}\right\}, \tag{14}$$

where

$$G(\theta_{k-1},\theta_{k+1}) = \sum_{i\in\Omega} G(\theta_{k+1}/i)\, G(i/\theta_{k-1})\, G(\theta_{k-1}) \tag{15}$$

and

$$\Pr\left\{d_k \neq \theta_k/\theta_{k-1},\theta_{k+1}\right\} = R\left[G(\theta_k/\theta_{k-1},\theta_{k+1})\right], \tag{16}$$

with $G(\theta_k/\theta_{k-1},\theta_{k+1}) = G(\theta_{k+1}/\theta_k)G(\theta_k/\theta_{k-1})/G(\theta_{k-1},\theta_{k+1})$. For the case $\Omega = \{0, 1\}$ and $\alpha = G(1/0) = G(0/1)$ we have $G(1) = 1/2$, and the quantities of interest are tabulated below:

$(\theta_{k-1},\theta_{k+1})$	$G(\theta_{k-1},\theta_{k+1})$	$G(1/\theta_{k-1},\theta_{k+1})$
(0, 0)	$\frac{\alpha^2+(1-\alpha)^2}{2} = \frac{1}{2} - \alpha(1-\alpha)$	$\frac{\alpha^2}{\alpha^2+(1-\alpha)^2}$
(0, 1)	$\alpha(1-\alpha)$	1/2
(1, 0)	$\alpha(1-\alpha)$	1/2
(1, 1)	$\frac{\alpha^2+(1-\alpha)^2}{2} = \frac{1}{2} - \alpha(1-\alpha)$	$\frac{(1-\alpha)^2}{\alpha^2+(1-\alpha)^2}$

* We use the fact proved in Section 2.3 that $G(\theta_k/\theta_1 \ldots, \theta_{k-1}, \theta_{k+1}, \ldots \theta_N) = G(\theta_k/\theta_{k-1},\theta_{k+1})$. Equation (14) follows from Equation (33) of Chapter 4 exactly as (13) followed from Equation (23) of Chapter 4.

Thus, for the normal case under discussion,

$$e_c^* = \gamma R(\alpha^2/\gamma) + (1-\gamma) R(1/2)$$

where $\gamma \equiv \alpha^2 + (1-\alpha)^2$. Hence, curves of $e_c^*(\alpha)$ can be calculated from the curves of $e_c(\alpha) = R(\alpha)$ plotted in Figures 3-7. These are shown in Figures 9 to 11. To be explicit,

$$e_c(\alpha) = \Phi\left[\frac{1}{m} \log \frac{1-\alpha}{\alpha} - \frac{m}{2}\right] + (1-\alpha)\left\{1-\Phi\left[\frac{1}{m} \log \frac{1-\alpha}{\alpha} + \frac{m}{2}\right]\right\}$$

and

$$e_c^*(\alpha) = \alpha^2 \Phi\left[\frac{1}{m} \log \frac{(1-\alpha)^2}{\alpha^2} - \frac{m}{2}\right] + 2\alpha(1-\alpha)\left[1-\Phi\left(\frac{m}{2}\right)\right]$$

$$+ (1-\alpha)^2 \left\{1-\Phi\left[\frac{1}{m} \log \frac{(1-\alpha)^2}{\alpha^2} + \frac{m}{2}\right]\right\} .$$

Figures 9, 10, and 11 show e_1, e_2 at $\alpha = 0$, $e_c(\alpha)$, and $e_c^*(\alpha)$ for $m = 1, 2,$ and 4. Figure 12 shows e_1, e_c, and e_c^* as a function of m for $\alpha = 0.2$.

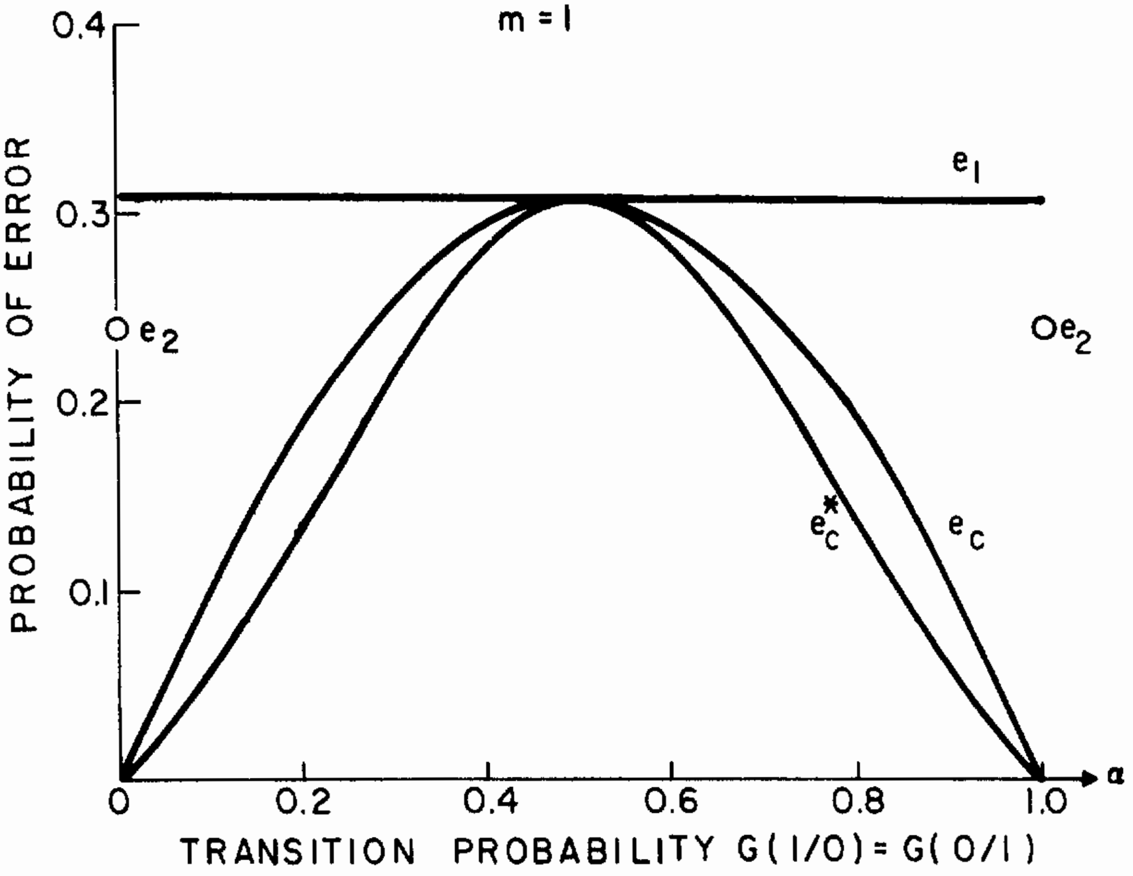

Figure 9 Error Probabilities for Normal Alternatives Displaced 1 Standard Deviation (S/N = 0 db)

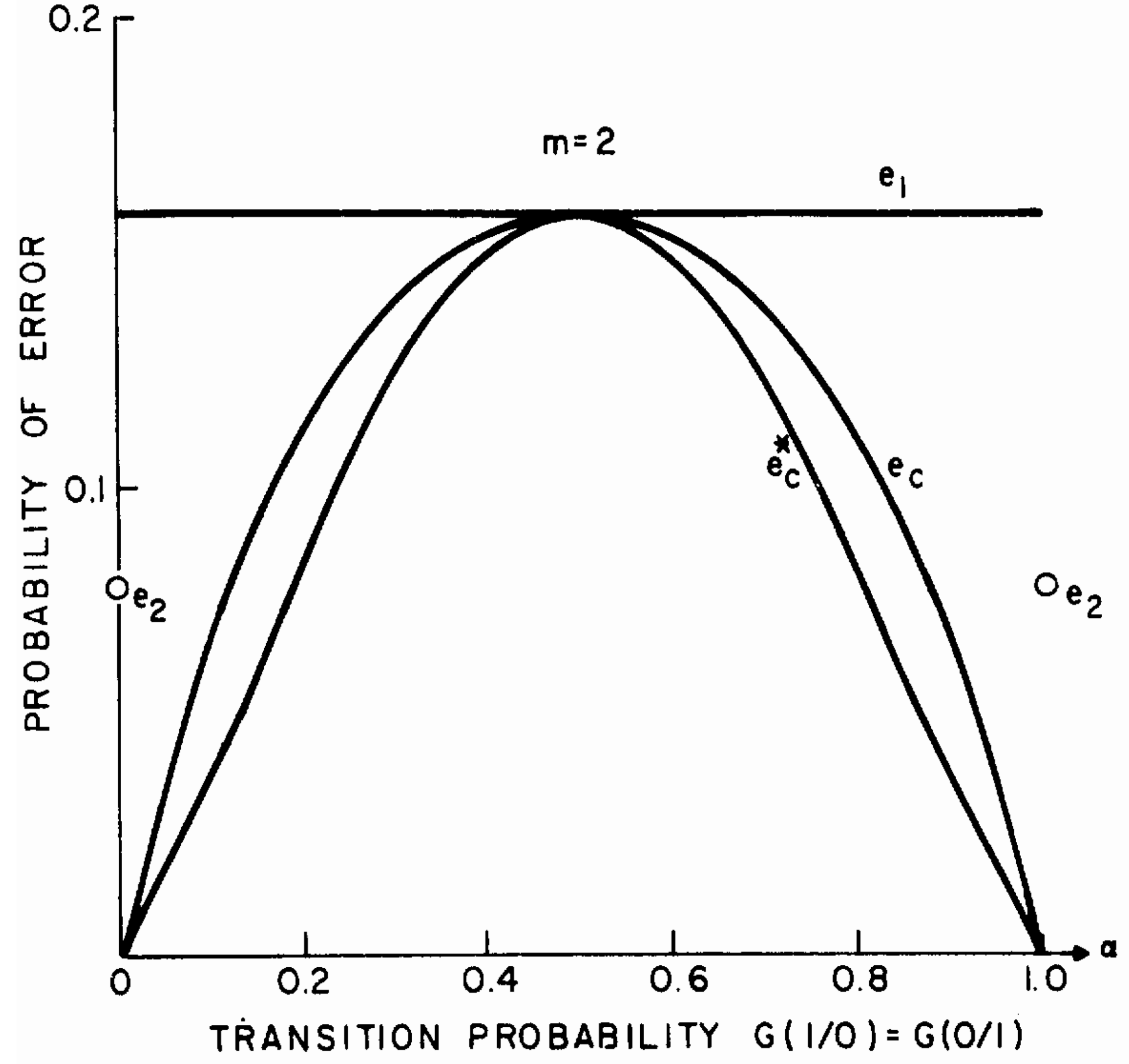

Figure 10 Error Probabilities for Normal Alternatives Displaced 2 Standard Deviations (S/N = 6 db)

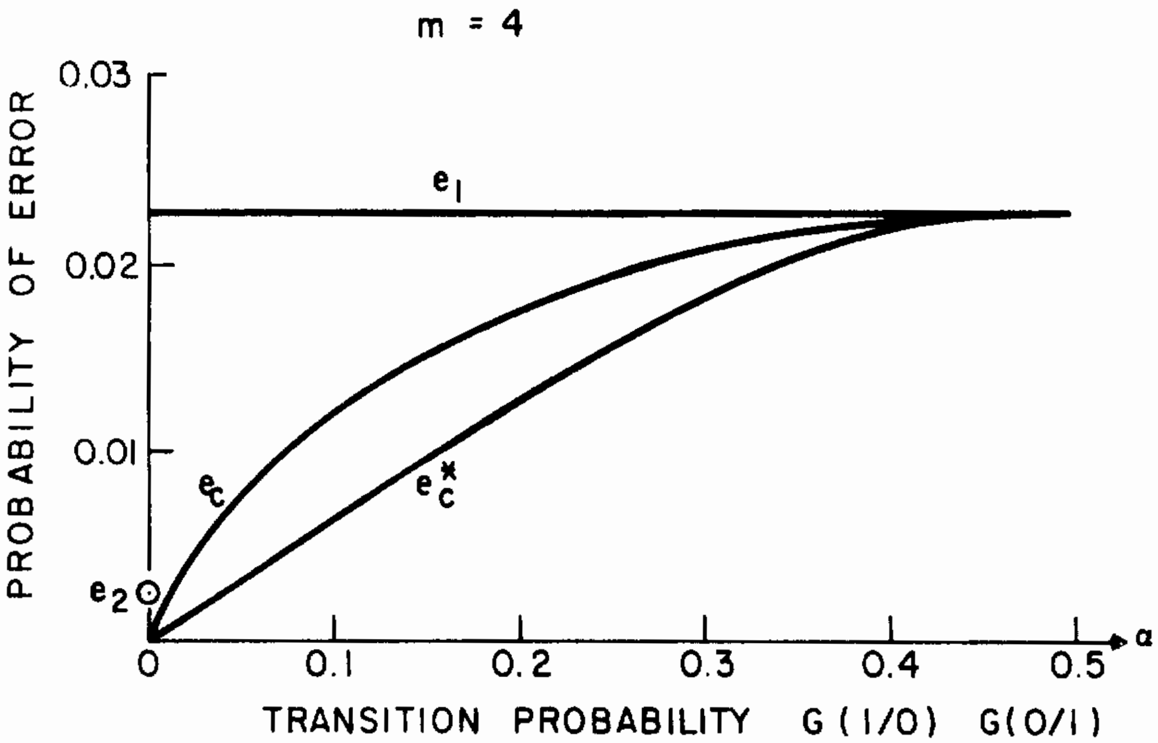

Figure 11 Error Probabilities for Normal Alternatives Displaced 4 Standard Deviations (S/N = 12 db)

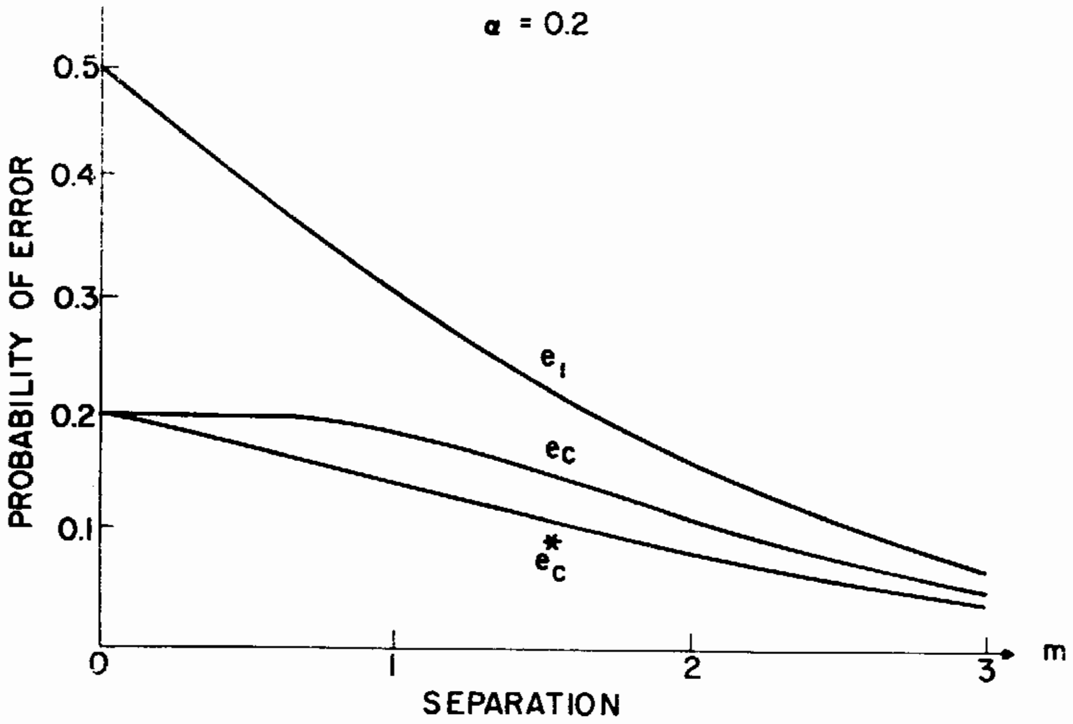

Figure 12 Minimum Error Probabilities Vs. Displacement for Normal Alternatives with a Transition Probability of 0.2

Section VI

CONCLUSIONS

The preceeding chapters represent a theoretical study of pattern recognition as a compound decision problem. Chapter 1 is mostly tutorial while Chapters 2 through 5 are mostly original. Some of the results of the original investigations are listed below:

- The evaluation of the finite sample-size performance of the Fix-Hodges nonparametric procedure is very cumbersome for multivariate distributions (Section 2. 2).
- The difficulty that an excessively large number of parameters need be estimated to represent general probability functions of many variables can be overcome by assuming that the variables are drawn from a Markov chain (Section 2. 3).
- The most general nonparametric pattern-recognition problem, which includes learning without a teacher, may be formulated as a distribution-free compound decision problem (Section 2. 4).
- Two specific procedures for adaptive adjustment of the threshold on the likelihood ratio have been prescribed. Using the sequential procedures of Chapter 3, the threshold rapidly approaches its optimum setting.
- Compound decision theory provides the optimal sequential procedure for taking context into accound in a classification

problem. The procedure described in Section 4.2 is almost as easily implemented as the simplest (non-optimal) procedure accounting for context.

- The non-optimal procedure, using the previous decision to account for context, performs very poorly when the populations are not well separated. Under such conditions, it performs even worse than a procedure that makes no attempt to account for context. When the populations are well separated, it can yield an improvement almost equal to that of the optimal sequential rule.
- Further improvement may be made through the use of non-sequential compound rules.

The following problems are suggested for further study:

- Develop procedures for estimating the parameters in the Markov-chain expansions of Section 2.3.
- Evaluate the Markov-chain procedure experimentally.
- Investigate the distribution-free compound decision problem defined in Section 2.4.
- Compare the two threshold-adjustment procedures prescribed in Chapter 3, by Monte-Carlo simulation.

Modify the procedures in Chapters 5 and 6 to take into account various factors peculiar to particular languages.

INDEX

REFERENCES

1 Abend, K., and Kanal, L.N.: "Classification of binary random patterns"(abstract), Ann. Math. Statist., vol. 36, no. 2, p. 730, April 1965.

2 Abend, K., Harley, T.J. and Kanal, L.N.: "Classification of binary random patterns," IEEE Trans. on Information Theory, vol. IT-11, no. 4, pp. 538-544, October 1965.

3 Blackwell, D., and Girshick, M.A.: Theory of Games and Statistical Decisions, New York: John Wiley and Sons, 1954.

4 Chow, C.K.: "An optimum character recognition system using decision functions," IRE Trans., vol. PGEC-6, pp. 247-254, December 1957.

5 Cooper, D.B., and Cooper, P.W.: "Non-supervised adaptive signal detection and pattern recognition," Information and Control, vol. 7, no. 3, pp. 416-444, September 1964.

6 Cover, T.M., and Hart, P.E.: "Nearest Neighbor Pattern Classification," Presented at the IEEE International Communications Conference, Philadelphia, July 1966.

7 Duda, R.O., and Fossum, H.: "Pattern classification by iteratively determined linear and piecewise linear discriminant functions," IEEE Trans. on Electronic Computers, vol. EC-15, no. 2, April 1966.

8 Fix, E., and Hodges, J.L.: Discriminatory Analysis; Nonparametric Discrimination: Consistancy Properties, USAF SAM Series in Statistics, Project No. 21-49-004, Report No. 4, School of Aviation Medicine, Randolph AFB, Texas, 1951.

9 Fix, E., and Hodges, J.L.: Discriminatory Analsyis; Nonparametric Discrimination: Small Sample Performance, USAF SAM Series in Statistics, Project No. 21-49-004, Report No. 11, School of Aviation Medicine, Randolph AFB, Texas, 1952.

10 Hannan, J. F., and Robbins, H.:"Asymptotic solutions of the compound decision problem for two completely specified distributions," Ann. Math. Statist., vol. 26, no. 1, pp. 37-51, March 1955.

11 Hannan, J. F., and Van Ryzin, J. R.:" Rate of convergence in the compound decision problem for two completely specified distributions," Ann. Math. Statist., vol. 36, no. 6, pp. 1743-1752, December 1965.

12 Harley, T. J., Kanal, L. N., and Abend, K.: "A note on the expected error in the probability of misclassification," Proc. IEEE, vol. 53, no. 11, pp. 1764-1765, November 1965.

13 Hilbert, D.: "Uber die stetige abbildung einer linie auf ein flachenstuck," Mathematische Annalen, vol. 38, pp. 459-460, Leipzig, March 1891.

14 Johns, M. V.: "An Empirical Bayes Approach to Non-parametric Two-way Classification," Studies in Item Analysis and Prediction (ed., H. Solomon), Stanford University Press, pp. 221-232, 1961.

15 Kanal, L., et al.: "Basic principles of some pattern recognition systems," Proc. National Electronics Conference, vol. 18, pp. 279-295, October 1962.

16 Kanal, L.: "Statistical Methods for Pattern Classification," Section 6 and Appendix H in Harley et al. Semi-Automatic Imagery Screening Research and Experimental Investigation, Report No. 2 and 3 by Philco Corporation on USAERDL Contract DA-36-039-SC-90742; 1963.

17 Murthy, V. K.: "Estimation of probability density," Ann. Math. Statist., vol. 36, no. 3, pp. 1027-1031, June 1965.

18 Murthy, V. K.: Nonparametric Estimation of Multivariate Densities with Applications, International Symposium on Multivariate Analysis, Aerospace Research Labs., Wright-Patterson AFB, Ohio, June 1965.

19 Nilsson, N. J., Learning Machines, New York: McGraw Hill, 1965.

20 Parzan, E.:"On estimation of a probability density function and mode," Ann. Math. Statist., vol. 33, no. 3, pp. 1065-1076, September 1962.

21 Robbins, H., "Asymptotically Subminimax Solutions of Compound Statistical Decision Problems," Proc. Second Berkley Symposium on Math. Statist. and Prob., University of California Press, pp. 131-148, 1951.

22 Robbins, H.: "An Empirical Bayes Approach to Statistics," Proc. Third Berkley Symposium on Math. Statist. and Prob., University of California Press, pp. 157-163, 1956.

23 Robbins, H.: "The empirical bayes approach to statistical decision problems," Ann. Math. Statist., vol. 35, no. 1, pp. 1-20, March 1964.

24 Robbins, H., and Samuel, E.: "Testing Statistical Hypothesis - the 'Compound' Approach," Recent Developments in Information and Decision Processes (e.d., R.E. Machol and P. Gray), Macmilln, New York, pp. 63-70, 1962.

25 Rosenblatt, M.:"Remarks on some nonparametric estimates of a density function," Ann. Math. Statist., vol. 27, pp. 832-837, 1965.

26 Samuel, E.:"An empirical Bayes approach to the testing of certain parametric hypothesis," Ann. Math. Statist., vol. 34, no. 4, pp. 1370-1385, December 1963.

27 Samuel, E.: "On the Compound Decision Problem in the Nonsequential and the Sequential Case," Ph.D. Thesis, Columbia University, 1961.

28 Samuel, E.:" Asymptotic solutions of the sequential compound decision problem," Ann. Math. Statist., vol. 34, no. 3, pp. 1079-1094, September 1963.

29 Samuel, E.: "On simple rules for the compound decision problem," Journal of the Royal Statistical Society, Series B, vol. 27, no. 2, pp. 238-240, 1965.

30 Van Meter, D., and Middleton, D.:"Modern statistical approaches to reception in communication theory,"IRE Trans., vol. PGIT-4, pp. 119-145, September 1954.

31 Van Ryzin, J.R.: "Asymptotic Solutions to Compound Decision Problems," Ph.D. Thesis, Michigan State University, February 1964.

32 Van Ryzin, J.R.: "The sequential compound decision problem with m x n finite loss matrix," Ann. Math. Statist., vol. 37, no. 4, pp. 954-975, August 1966.

33 Van Ryzin, J.R.: "Non-Parametric Bayesian Decision Procedures for (Pattern) Classification with Stochastic Learning," Fourth Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes, September 1965.

34 Van Ryzin, J.R.:"Repetitive play in finite statistical games with unknown distributions," Ann. Math. Statist., vol. 37, no. 4, pp. 976-994, August 1966.

35 Wald, A.: Statistical Decision Functions, New York: John Wiley and Sons, 1950.

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*
Philco-Ford Corporation
System Sciences Laboratory
Blue Bell, Pennsylvania 19422

2a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

2b. GROUP
N/A

3. REPORT TITLE

COMPOUND DECISION PROCEDURES FOR PATTERN CLASSIFICATION

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Final Report, 1 August 1965 - 30 April 1967

5. AUTHOR(S) *(First name, middle initial, last name)*

Kenneth Abend

6. REPORT DATE
DECEMBER 1967

7a. TOTAL NO. OF PAGES
97

7b. NO. OF REFS
35

8a. CONTRACT OR GRANT NO. AF 33(615)-2966

b. PROJECT NO. 7233

c. Task No. 723305

d.

9a. ORIGINATOR'S REPORT NUMBER(S)

PR-345

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

AMRL-TR-67-10

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited. It may be released to the Clearinghouse, Department of Commerce, for sale to the general public.

11. SUPPLEMENTARY NOTES
Dissertation for the Degree of Doctor of Philosophy at the University of Pennsylvania.

12. SPONSORING MILITARY ACTIVITY
Aerospace Medical Research Laboratories, Aerospace Medical Div., Air Force Systems Command, Wright-Patterson AFB, O. 45433

13. ABSTRACT

Compound decision theory is shown to be powerful as a general theoretical framework for pattern recognition, leading to nonparametric methods, methods of threshold adjustment, and methods for taking context into account. The finite-sample-size performance of the Fix-Hodges nearest-neighbor nonparametric classification procedure is derived for independent binary patterns. The optimum (Bayes) sequential compound decision procedure, for known distributions and dependent states of nature is derived. When the states of nature form a Markov chain, the procedure is recursive, easily implemented, and immediately applicable to the use of context. A similar procedure, in which a decision depends on previous observations only through the decision about the preceding state of nature, can (when the populations are not well separated) yield results significantly worse than a procedure that does not depend on previous observations at all. When the populations are well separated, however, an improvement almost equal to that of the optimum sequential rule is achieved.

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Statistical Decision Theory						
Pattern Recognition						

Security Classification