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. Oestreich, PhD, Chief, Mathematics and Analysis Branch, Biodynamics and Bionics Division, Biomedical Laboratory, Aerospace Medical Research Laboratories.


This technical report has been reviewed and is approved.

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

Approved for Public Release
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 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. 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

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.2</td>
<td></td>
</tr>
<tr>
<td>Pattern Classification</td>
<td>1</td>
</tr>
<tr>
<td>1.3</td>
<td></td>
</tr>
<tr>
<td>The Simple Decision Problem</td>
<td>4</td>
</tr>
<tr>
<td>1.4</td>
<td></td>
</tr>
<tr>
<td>The Compound Decision Problem</td>
<td>8</td>
</tr>
<tr>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>Asymptotic Solutions</td>
<td>11</td>
</tr>
<tr>
<td>II</td>
<td></td>
</tr>
<tr>
<td>NONPARAMETRIC CLASSIFICATION</td>
<td>17</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td>The Fix-Hodges Procedure and &quot;Window Carpentry&quot;</td>
<td>17</td>
</tr>
<tr>
<td>2.2</td>
<td></td>
</tr>
<tr>
<td>Nearest-Neighbor Classification of Binary Patterns</td>
<td>24</td>
</tr>
<tr>
<td>2.3</td>
<td></td>
</tr>
<tr>
<td>Markov-Chain Classification of Binary Patterns</td>
<td>28</td>
</tr>
<tr>
<td>2.4</td>
<td></td>
</tr>
<tr>
<td>The Distribution-Free Compound Decision Problem</td>
<td>37</td>
</tr>
<tr>
<td>III</td>
<td></td>
</tr>
<tr>
<td>ADAPTIVE THRESHOLD ADJUSTMENT</td>
<td>41</td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>3.1</td>
<td></td>
</tr>
<tr>
<td>The Two-Threshold Method</td>
<td>41</td>
</tr>
<tr>
<td>3.2</td>
<td></td>
</tr>
<tr>
<td>The Bayesian Approach</td>
<td>44</td>
</tr>
<tr>
<td>IV</td>
<td></td>
</tr>
<tr>
<td>COMPOUND PROCEDURES FOR DEPENDENT STATES OF NATURE</td>
<td>47</td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td></td>
</tr>
<tr>
<td>The Use of Context in Print Reading</td>
<td>47</td>
</tr>
<tr>
<td>4.2</td>
<td></td>
</tr>
<tr>
<td>The Sequential Compound Bayes Procedure for Dependent States of Nature</td>
<td>49</td>
</tr>
<tr>
<td>4.3</td>
<td></td>
</tr>
<tr>
<td>Sequential Rules for Markov-Chain Dependence</td>
<td>54</td>
</tr>
<tr>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>Non-Sequential Compound Rules</td>
<td>61</td>
</tr>
</tbody>
</table>

iv

Approved for Public Release
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>THE TWO-CLASS PROBLEM WITH NORMAL DISTRIBUTIONS</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>5.1 Sequential Rules for Markov-Chain Dependence</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>5.2 The Second-Component Sequential Compound Bayes Risk</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>5.3 Non-Sequential Rules for Markov-Chain Dependence</td>
<td>80</td>
</tr>
<tr>
<td>VI</td>
<td>CONCLUSIONS</td>
<td>88</td>
</tr>
</tbody>
</table>
## LIST OF ILLUSTRATIONS

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Average Risks and the Bayes Envelope</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>An Open Space-Filling Curve — the First Four Steps</td>
<td>36</td>
</tr>
<tr>
<td>3</td>
<td>Error Probabilities for Normal Alternatives Displaced 1 Standard Deviation (S/N = 0 db)</td>
<td>71</td>
</tr>
<tr>
<td>4</td>
<td>Error Probabilities for Normal Alternatives Displaced 2 Standard Deviations (S/N = 3 db)</td>
<td>72</td>
</tr>
<tr>
<td>5</td>
<td>Error Probabilities for Normal Alternatives Displaced 3 Standard Deviations (S/N = 6 db)</td>
<td>73</td>
</tr>
<tr>
<td>6</td>
<td>Error Probabilities for Normal Alternatives Displaced 4 Standard Deviations (S/N = 12 db)</td>
<td>74</td>
</tr>
<tr>
<td>7</td>
<td>Error Probabilities vs Displacement for Normal Alternatives with a Transition Probability of 0.2</td>
<td>75</td>
</tr>
<tr>
<td>8</td>
<td>Error Probabilities vs Displacement for Normal Alternatives with Zero Transition Probability</td>
<td>81</td>
</tr>
<tr>
<td>9</td>
<td>Error Probabilities for Normal Alternatives Displaced 1 Standard Deviation (S/N = 0 db)</td>
<td>84</td>
</tr>
<tr>
<td>10</td>
<td>Error Probabilities for Normal Alternatives Displaced 2 Standard Deviations (S/N = 6 db)</td>
<td>85</td>
</tr>
<tr>
<td>11</td>
<td>Error Probabilities for Normal Alternatives Displaced 4 Standard Deviations (S/N = 12 db)</td>
<td>86</td>
</tr>
<tr>
<td>12</td>
<td>Minimum Error Probabilities vs Displacement for Normal Alternatives with a Transition Probability of 0.2</td>
<td>87</td>
</tr>
</tbody>
</table>
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^+ = k^{th} component error probability for the "decision-directed"
      rule.

e^+ = limit of e_k^+.

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

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

e_c^* = minimum error probability for a non-sequential compound
      rule.
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 imbode 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 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 $(x_1, x_2, \ldots, x_d)$, 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 $\Omega$ be an index set for the probability distributions $P_\omega$ 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 $\Omega$ are called the states of nature [1].

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 $\Omega$ is the set of letters in the English alphabet, action $a_0$ may be to decide that $\omega$ 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 $\Omega$ and $A$ are finite, we write $L_{ij}$ for $L$ and the loss function $L$ becomes a rectangular matrix with elements $L_{ij}$. We shall identify the states of nature with the classes $i = 1, 2, \ldots, r$ 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.
<table>
<thead>
<tr>
<th>Pattern Classification</th>
<th>Statistical Decision Theory</th>
<th>Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Pattern space</td>
<td>Outcome</td>
<td></td>
</tr>
<tr>
<td>2. Set of observables</td>
<td>Vector random variable</td>
<td>( X(x_1, \ldots, x_n) )</td>
</tr>
<tr>
<td>3. Pattern vector</td>
<td>Value of random variable</td>
<td>( x(x_1, \ldots, x_d) )</td>
</tr>
<tr>
<td>Measurement space</td>
<td>Induced sample space</td>
<td>( S = {x} )</td>
</tr>
<tr>
<td>4. Class</td>
<td>State of nature</td>
<td>( \omega )</td>
</tr>
<tr>
<td>Parameter space</td>
<td>( \Omega = {\omega_1, \omega_2, \ldots, \omega_r} )</td>
<td></td>
</tr>
<tr>
<td>5. Classification</td>
<td>Action</td>
<td>( a )</td>
</tr>
<tr>
<td>Action space</td>
<td>( A = {a_1, a_2, \ldots, a_k} )</td>
<td></td>
</tr>
</tbody>
</table>

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 of states of nature \( \theta = (\theta_1, \ldots, \theta_w) \) and a corresponding vector \( x = (x_1, \ldots, x_k) \) of (n-dimensional) random variables (patterns). In the \( k \)th component problem, we assume that for a given \( \theta_k \in \Omega \), \( x_k \) is independent of the other \( x \)'s and \( \theta \)'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.

Approved for Public Release
1.2 THE SIMPLE DECISION PROBLEM

A finite statistical decision problem involves a set of states of nature $\mathcal{N} = \{1, 2, \ldots, r\}$, and a set of actions $A = \{1, 2, \ldots, s\}$. For every $i \in \mathcal{N}$, $j \in A$, 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 $x \in S$ of an observable random variable, and we assume that the distribution of $x$ is $P_i(x)$ when the true state of nature is $i \in \mathcal{N}$. 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 $j = ji(x)$, $t$ is said to be non-randomized. The risk function $R(t, i)$ is the expected loss incurred by the use of $t$,

$$R(t, i) = \sum_{j=1}^{s} \int_{S} L_{ij}(j|x) p_i(x) \, dx.$$  \hspace{1cm} (1)

Let $q(i) = q_i$ be an a priori probability distribution on $\mathcal{N}$.

The average risk, or Bayes risk of $t$ with respect to $q$ is

$$R(q, t) = \sum_{i=1}^{r} R(t, i) q_i = \int_{S} \sum_{j=1}^{s} \sum_{i=1}^{r} L_{ij} t(j|x) p_i(x) \, q_i \, dx.$$  \hspace{1cm} (2)

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

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

$$R(q) = \min_t R(q, t) = \min_t R(q, t')$$

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 $x$:

$$q(i/x) = \frac{p_i(x) q_i}{p(x)} = \frac{p_i(x) q_i}{\sum_{j=1}^{J} p_{ij}(x) q_j} .$$

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

$$\sum_{i=1}^{I} L_{ij} q(i/x) .$$

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 $p(s, i) = p(x) q_i$.

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 $p_i(x)$ under the hypothesis $H_0$. Let $L_{ij}$ be the loss incurred if $H_j$ is chosen when $H_i$ is true ($i, j = 0, 1$) let $q_i = q$ 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 choosing $H_1$, given $x$, and $t(0/x) = 1-t(1/x)$.

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

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

(6)

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

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

(7)

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

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

<table>
<thead>
<tr>
<th>Action</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>State of 1</td>
<td>0</td>
<td>w</td>
</tr>
<tr>
<td>Nature 0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

the minimum risk (the Bayes envelope) $R(q, t^\alpha)$ is the continuous concave function

$$R(q) = qw \left[ 1 - \frac{q}{1-q} \int_T p_1(x) \, dx \right]$$

(8)

shows 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

\[
\hat{R}(q, t^q) = q w \left[ 1 - P_1(T^q) \right] + (1 - q) P_0(T^q),
\]

(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 \( t_0 \), for which \( R(1, t_0) = R(0, t_0) = \max_q R(q) \). This "safe" procedure \( t_0 = t_{q_0} \) designs \( T_0 = T(q_0) \) against the maximum point \( q_0 \) so that the average risk is constant:

\[
\bar{R}(q, t_{q_0}) = w \left[ 1 - P_1(T_{q_0}) \right] = P_0(T_{q_0}).
\]

(10)

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Average Risks and the Bayes Envelope}
\end{figure}
1.3 THE COMPOUND DECISION PROBLEM

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

$$p(x_k/x_{k-1}, x_{k+1}, \ldots, x_N, \theta_N) = p(x_k/\theta_k) = P_{\theta_k}(x_k)$$

and hence

$$p(x_k/\theta_k) = \prod_{j=1}^{k} p(x_j/\theta_j).$$

We do not assume that the $\theta$'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 $x_N$ are at hand before the individual decisions must be made, one can use a compound decision rule $t_N = (t_1, \ldots, t_N)$, where $t_k = (j/x_N)$ is for each $x_N$ a distribution over $A$ according to which the $k$th action is chosen. If only the observations $x_k$ are at hand when the $k$th decision must be made, one can use a sequential compound decision rule, where $t_k = t_k(j/x_k)$. A simple rule is one where $t_k = t_k(j/x_k)$, that is, one where the decision about $\theta_k$ 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 $t_N$ is
\[ R(\theta_N, \mathbf{1}_N) = \frac{1}{N} \sum_{k=1}^{N} u(\theta_N, \mathbf{1}_k) \]  

(12)

where

\[ R(\theta_N, \mathbf{1}_k) = \int_{S_N} \prod_{j \neq k} \phi_k(\theta_j) \rho_{\theta_N}(\mathbf{x}_N) \, d\mathbf{x}_N \]  

(14)

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

\[ \rho_{\theta_N}(\mathbf{x}_N) = \prod_{k=1}^{N} \rho_{\theta_k}(x_k) \]  

(14.1)

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

One can also talk about a compound Bayes risk \( \overline{R}(G, \mathbf{1}_N) \) with respect to an a priori distribution \( G(\theta_N) \) over \( \mathcal{S}_N \), the \( N \)-fold Cartesian product of \( \mathcal{S} \). The compound Bayes risk is

\[ \overline{R}(G, \mathbf{1}_N) = \sum_{\theta_N \in \mathcal{S}_N} R(\theta_N, \mathbf{1}_N) G(\theta_N) = \frac{1}{N} \sum_{k=1}^{N} \overline{R}(G, \mathbf{1}_k) \]  

(15)

where

\[ \overline{R}(G, \mathbf{1}_k) = \sum_{\theta_N \in \mathcal{S}_N} R(\theta_N, \mathbf{1}_k) G(\theta_N) \]  

(16)

is the \( k \)-th component Bayes risk. A procedure is compound Bayes against \( G \) if it minimizes \( \overline{R}(G, \mathbf{1}_N) \). Thus, the compound Bayes procedure \( \mathbf{1}_N^G \) is one that, for every \( k \), minimizes

Approved for Public Release
\[ R(G, t_k) = \int \sum_{j=1}^{N} \sum_{\theta_N} L_{\theta k} t_k(j/\theta_N) \text{p}(\theta_N/\theta_N) \ G(\theta_N) \ d\theta_N. \] (17)

Hence \( t_k^G(j/\theta_N) = 1 \) for that \( j \) which minimizes the quantity

\[ \sum_{\theta_N} L_{\theta k} \text{p}(\theta_N/\theta_N) = \sum_{\theta_k} L_{\theta k} \text{p}(\theta_N/\theta_N) \] (18)

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

\[ G(\theta_k/\theta_N) = \frac{\text{p}(\theta_N/\theta_k)}{\text{p}(\theta_N)} \]

Since the denominator \( \text{p}(\theta_N) = \sum_{\theta_N} \text{p}(\theta_N/\theta_N) G(\theta_N) \) is independent of \( \theta_k \).

When \( G(\theta_N) = \sum_{k=1}^{N} q_k(\theta_k) \), 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 \( G(\theta_N) = \sum_{k=1}^{N} q(\theta_k) \).

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

Approved for Public Release
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

helpful 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 $t_k = \ell(j|x_k)$ Equation (12) be-

comes

$$R(\Theta, N) = \frac{1}{N} \sum_{k=1}^{N} R(\Theta_k, t) = \sum_{i=1}^{r} q^N(i) R(i, t) = \tilde{R}(q^N, t),$$  \hspace{1cm} (19)

where $R(\cdot, t)$ and $\tilde{R}(\cdot, t)$ are defined by Equations (1) and (2) and $q^N(i)$
is the fraction of the $N \Theta$'s that are equal to $i$. Thus, the simple sym-

metric procedure which is Bayes against the empirical a priori distribution

$q^N(i)$, would minimize the compound risk $R(\Theta, N)$. If we knew $q^N$ in

advance, we could, by using a simple symmetric rule $l_t(q(j|x_k))$ that

chooses action $j$ to minimize the quantity

$$\sum_{i=1}^{r} L_{ij} p_t(x_k) q^N(i),$$  \hspace{1cm} (20)

obtain the Bayes envelope $R(q^N)$. But $q^N$ is not known. To escape this

11

Approved for Public Release
predicament, we use the observations \( x_i \) to obtain an estimate \( \hat{\theta}_N \) of \( \theta \) and use the compound rule \( t_{k}(j/x_k) = t_{N}(j/x_k) \) to choose the \( k \)th action.

For known densities, but unknown \( \theta \), we may take one of two approaches. The empirical Bayes approach of Robbins [22, 26, 23] assumes that the \( \theta \)'s are independent and identically distributed according to an a priori distribution \( q(\theta) \) over \( \Omega \) and examines convergence of the \( N \)th component Bayes risk, \( R(G, t_N) \), 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, \( R(\hat{\theta}_N, t_N) \), to \( R(q_N) \) for any sequence \( \hat{\theta}_N \). Hannan and Robbins [10, 24] have shown that the rule \( \hat{\theta}_N = (\hat{t}_1, \ldots, \hat{t}_N) \), with

\[
\hat{t}_k = t_{N}(j/x_k)
\]

Bayes against the consistent estimate of \( q_N \) given in Equation (29) below, is "optimal in the limit", i.e., given any \( \varepsilon > 0 \), there exists \( N_\varepsilon \) such that for all \( N > N_\varepsilon \)

\[
R(\hat{\theta}_N, \hat{t}_N) - R(q_N) < \varepsilon
\]

uniformly for all \( \hat{\theta}_N \in N \), 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
\( R(\hat{\theta}_N, \theta_N) \) to \( R(\theta, \theta_N) \) is at least of the order \( \sqrt{N} \); for there exists a
constant \( c \), independent of \( \theta_N \) and of \( N \) such that
\[
R(\hat{\theta}_N, \theta_N) - R(\theta, \theta_N) \leq cN^{-1/2}.
\] (23)

With restrictions on \( p_j(x) \), even faster rates of convergence are
obtainable.

If only the first \( k \) observations are at hand when the \( k \)th decision
must be made, a sequential compound decision rule, \( \hat{\theta}^*_{k,N} \) introduced by
Samuel [27, 24, 28, 29] is used, where \( \hat{\theta}^*_{k,N}(j|x_k) = \hat{\theta}^*_{k-1}(j|x_k) \). This
rule is also optimal in the limit, and there even exists a constant \( c \)
independent of \( \theta_N \) and of \( N \) such that Equation (23) is satisfied [32, 33].

Since \( \hat{\theta}^*_{k,N}(j|x_k) \) does not depend on \( N \), the sequential compound decision
function \( \hat{\theta}^*_{k,N} = (\hat{\theta}^*_1, \ldots, \hat{\theta}^*_N) \) may be used without knowing \( N \) in advance.

Let us return to the case of the 2 \( \times \) 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 \( \theta \) and for any \( x_k \),
k \( > \) 0, let
\[
h_k(x_k) = \frac{1}{k} \sum_{i=1}^{k} h(x_i).
\] (24)

Letting
be the "truncated" estimate of
\[ \hat{q}_k(1/x_k) = \begin{cases} 
0 & \text{if } h_k < 0 \\
h_k & \text{if } 0 \leq h_k \leq 1 \\
1 & \text{if } h_k > 1 
\end{cases} \quad (25) \]
the rule
\[ \hat{g}_k(1/x_k) = \begin{cases} 
1 & \text{for } \frac{P_1(x_k)}{P_0(x_k)} > \frac{1 - \hat{g}_{N}(x_N)}{w \hat{g}_{N}(x_N)} \\
0 & \text{otherwise} 
\end{cases} \quad (27) \]
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 \(\hat{g}_N\) and of \(N\) [11].

In the sequential case we use at the \(k\)th decision rule [27, 28]
\[ \hat{t}_k(1/x_k) = \begin{cases} 
1 & \text{for } \frac{P_1(x_k)}{P_0(x_k)} > \frac{1 - \hat{g}_{k-1}}{w \hat{g}_{k-1}} \\
0 & \text{otherwise} 
\end{cases} \quad (28) \]
with \(\hat{g}_0 = 1/2\).

Using a constant threshold (corresponding to \(q^*\)), the compound risk as a function of \(\hat{g}_N\) lies along a straight line \(R(\hat{g}_N; q^*)\) as in
Figure 1 (see Equation (19)). With the procedure described above $R(\Delta N, \Delta N)$ approaches the Bayes envelope $R(\Delta N) \leq R(\Delta N, \Delta N')$, for any sequence $\Delta N$ (regardless of whether or not the $\Delta N$'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 $p(x)$ are unknown, b) the a priori distribution $C(\Delta N)$ 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.

15
In Section III the 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 \( G(\theta_N) \neq \prod_{k=1}^{N} q_k(\theta_k) \). 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 1961, 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

$$\frac{Q_1/N_1}{Q_2/N_2} > \zeta.$$  \hfill (1)

The threshold $\zeta$ 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 I, the Bayes procedure for known densities and known a priori probabilities is given by

$$t(1/x) = \begin{cases} 1 & \text{if } \frac{p_1(x)}{p_2(x)} > \frac{q_2}{q_1} \Rightarrow W \\ 0 & \text{if } \leq \end{cases}$$  \hfill (2)

or

$$t(1/x) = \begin{cases} 1 & \text{if } \frac{q(1/x)}{q(2/x)} > W \\ 0 & \text{otherwise} \end{cases}$$  \hfill (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 $N$) $x_N = (x_1, x_2, \ldots, x_N)$ 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

\[
\lim_{N \to \infty} K = \infty
\]

\[
\lim_{N \to \infty} K/N = 0
\]

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 \(f_N(x, X_{xN})\) is such that for each \(x\)

\[
f_N(x, X_{xN}) \xrightarrow{\text{a.s.}} q(1/x) - Wq(2/x) \text{ in probability}
\]

(where "in probability" refers to the distribution of \(X_{xN}\)), then the decision rule

\[
f_N(1/x) = \begin{cases} 
1 & \text{if } f_N(x, X_{xN}) > 0 \\
0 & \text{if } f_N(x, X_{xN}) \leq 0
\end{cases}
\]

has a Bayes risk that approaches the Bayes envelope,

\[
\lim_{N \to \infty} R(q, t_{xN}) = R(q).
\]

Thus the procedure of assigning \(x\) to \(P_1\) if and only if

\[Q_1/Q_2 > W\]

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

$$\sum_{i=1}^{r} L_{ij} Q_i.$$  \hspace{1cm} (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 $\hat{R}(q, t \hat{q}(x))$ 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 $L_{11} = L_{22} = 0$,

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

In general \cite{34},

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

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 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 \( F(x) = F(x_1, \ldots, x_i) \) and density function \( f(x) \).

The sample distribution function

\[
F_N(x_1, \ldots, x_n) = \frac{1}{N} \{ \text{No. of observations } X_i, i = 1, 2, \ldots, N \text{ such that } X_{ij} \leq x_j, j = 1, 2, \ldots, n \text{ is a binomially distributed random variable} \}
\]

whose mean and variance are given by

\[
\begin{align*}
E[F_N(x)] &= F(x) \\
\text{Var}[F_N(x)] &= \frac{1}{N} \cdot F(x) \left(1 - F(x)\right).
\end{align*}
\]

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

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

\[
dF_N(y_1, \ldots, y_n) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{1}{Nh_1 \cdots h_n} \sum_{i=1}^{N} \ \mathbf{K} \left( \frac{x_1 - y_1}{h_1}, \ldots, \frac{x_n - y_n}{h_n} \right)
\]

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 \( N \to \infty \).

Approved for Public Release

21
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. \( |h_i| \geq |h_j| \) for all \( i = 1, \ldots, n \Rightarrow K(x_1) \leq K(x_2) \)
4. \( \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} K(x) \, dx_1 \cdots dx_n = 1 \).

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

\[
\lim_{N \to \infty} N \prod_{j=1}^{n} h_j \, \text{Var}\left\{f_N(x)\right\} = f(x) \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} K^2(y_1) \, dy_1 \cdots dy_n
\]

if in addition to

\[
\lim_{N \to \infty} \prod_{j=1}^{n} h_j(N) = 0,
\]

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

\[
\lim_{N \to \infty} N^{1/2} h_1(N) \cdots h_n(N) = \infty,
\]

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 $h$",

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

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 $\|x\| = (\sum x_i^2)^{1/2}$ is used, it is equivalent to a "circular" window of radius $h(= h_1 = h_2 = \ldots = h_n)$:

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

with $h$ as a random variable.

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

$$f_N(x) = \frac{1}{Rh^{11}} \sum_{i=1}^{N} K \left( \frac{x - X_i}{h} \right)$$

with $h \rightarrow 0$ and $Nh^2 \rightarrow \infty$ 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 \( R = R(q) \), and \( R^* \) is the limit of the nearest neighbor error probability as \( N \to \infty \), then for \( r \) classes and \( I_{ij} = 1 - \delta_{ij} \), 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 \( \mu \) 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

\[
R \leq R^* \leq R \left( 2 - (r/r-1) R \right) .
\]

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

$$\hat{\theta} = \frac{1}{2Nn} \sum_{j=1}^{N} \sum_{i=1}^{n} x_{ij} + 1 - y_{ij}.$$

We have

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

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

A sample $z = (z_1, \ldots, z_n)$ 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

- $r < n/2$ and $\hat{\theta} < 1/2$,
- $r > n/2$ and $\hat{\theta} > 1/2$,

and class 2 if

- $r < n/2$ and $\hat{\theta} > 1/2$,
- $r > n/2$ and $\hat{\theta} < 1/2$.
Since $q_1 = q_2 = 1/2$, the probability of misclassification is equal to the probability of misclassifying a sample from class 1:

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

For $n$ odd

$$V_1 \{ r < n/2 \} \sum_{r = n/2}^{n} v_r(r) = \sum_{r = n/2}^{n} \binom{n}{r} \alpha^r (1 - \alpha)^{n-r} = B \left( \frac{n-1}{2} ; n, \alpha \right)$$

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. The probability of rejection (a tie) for $n$ odd, is

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

When we are faced with the preceding 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 1 is a y, given that 1 is an x. Let $P_2(r)$ be the probability that the nearest neighbor to 1 is a y, given that 1 has $r$ ones. Then $P_{1e}$ is the expectation of $P_2(r)$ with $z$ distributed according to $P_1$.

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

Let $h_{1e}(\hat{d})$ be the probability that the Hamming distance from $z$ to an
is \( \delta \) (given that \( x \) has \( r \) ones). It is the probability that \( x \) is the same as \( z \) in \( n-\delta \) places and different in \( \delta \) places,

\[
h_{x}(\delta) = P\{d(x,z) = \delta\}.
\]

Similarly, we define

\[
k_{y}(\delta) = P\{d(y,z) = \delta\},
\]

\[
H_{x}(\delta) = \sum_{d=0}^{\delta} h_{x}(d)
\]

\[
K_{x}(\delta) = \sum_{d=0}^{\delta} k_{x}(d).
\]

Then

\[
P_{2}(r) = N \sum_{\delta=0}^{n} \left[ 1 - H_{x}(\delta) \right] \left[ 1 - K_{x}(\delta) \right]^{N-1} k_{x}(\delta),
\]

where

\[
h_{x}(\delta) = \sum_{s=0}^{\delta} b(s;r,1-s) b(\delta-s; n-r, s) = \sum_{s=0}^{\delta} b(s;r,\alpha) b(\delta-s; n-r, \alpha)
\]

\[
H_{x}(\delta) = \sum_{\delta=0}^{\delta} \sum_{s=0}^{\delta} b(s;r,\alpha) b(\delta-s; n-r, \alpha) = \sum_{s=0}^{\delta} \sum_{d=s}^{\delta} b(r-s,\alpha) b(d-s; n-r, \alpha)
\]

\[
= \sum_{s=0}^{\delta} b(r-s,\alpha) \binom{\delta-s}{s} b(r-n,\alpha) = \sum_{s=0}^{\delta} b(r-s,\alpha) B(\delta-s; n-r, \alpha),
\]

with \( k_{x}(\delta) \) and \( K_{x}(\delta) \) defined the same way but with \( \alpha \) replaced by \( \beta \).

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

\[
P_{E} = \sum_{r=0}^{3} b(r,n,\alpha) P_{2}(r),
\]

27
where

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

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

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 develops 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
formulations 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:

$$p(x_k | x_1, x_2, \ldots, x_{k-1}) = p(x_k | x_{k-1})$$

(11)

for $k = 1, 2, \ldots, n$. A well-known property of such a chain is that

$$p(x_j | x_1, x_2, \ldots, x_{j-1}) = p(x_j | x_{j-1})$$

(12)

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

$$p(x_k | x_1, x_2, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n) = \frac{p(x_1, \ldots, x_n)}{p(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n)}$$

(13)

$$= \frac{p(x_1)p(x_2 | x_1) \ldots p(x_k | x_{k-1})p(x_{k+1} | x_k) \ldots p(x_n | x_{n-1})}{p(x_1)p(x_2 | x_1) \ldots p(x_{k+1} | x_k) \ldots p(x_n | x_{n-1})}$$

$$= \frac{p(x_k | x_{k-1})p(x_{k+1} | x_k)}{p(x_k+1 | x_k)}$$

$$= \frac{p(x_{k-1} x_k x_{k+1})}{p(x_{k-1} x_{k+1})} = p(x_k | x_{k-1}, x_{k+1})$$
so that any point is dependent on only its two nearest neighbors, one on either side. Similarly, for the $r$th-order Markov chain,

$$p(x_{k}/x_{k-1}x_{k-2}\ldots x_{k-r}) = p(x_{k}/x_{k-r}\ldots x_{k-1})$$  \hspace{1cm} (14)

we have

$$p(x_{k}/x_{1}x_{2}\ldots x_{k-1}x_{k+1}\ldots x_{n}) = p(x_{k}/x_{k-r}\ldots x_{k-1}x_{k+1}\ldots x_{k+r})$$  \hspace{1cm} (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 $x_0 = 0$ or 1 and $p(000:11) = 1/2$ 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 $p(x_4 = 1/x_2 = 0) = 1/2$ 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

$$x_1 = p(x_1 = 1/x_{i-1} = 0)$$

$$x_2 = p(x_1 = 1/x_{i-1} = 1)$$  \hspace{1cm} (16)

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

31

Approved for Public Release
\[ \alpha_1 = p(x_1 = 1) \]

and \( \omega_{n+1} = \beta_{n+1} = 0 \). There are now \( 2n-1 \) parameters rather than \( 2^n - 1 \). If \( \omega_1 \) and \( \beta_1 \) were independent of \( i \) (which is not the case here), one would have a stationary chain.

Expanding the joint probability,

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

\[= \alpha_1 (1-\omega_1) \prod_{i=2}^{n} \left( \beta_{i-1} x_i \right) \left( 1-\beta_i \right) x_{i-1}(1-x_i) \]

\[= \alpha_1 (1-\omega_i) x_i (1-\omega_{i-1}) (1-\omega_{i-1}) (1-x_i) \]  

(17)

Taking logarithms and collecting terms, we obtain:

\[
\log p(x_1 x_2 \ldots x_n) = A_\emptyset + \sum_{i=1}^{n} A_i x_i + \sum_{i=2}^{n} B_i x_{i-1} x_i ,
\]

(18)

where

\[
A_\emptyset = \sum_{i=1}^{n} \log (1-\omega_i)
\]

(19)

\[
A_i = \log \frac{\omega_i}{1-\omega_i} + \log \frac{1-\beta_{i+1}}{1-\alpha_{i+1}}
\]

\[
B_i = \log \frac{\beta_i}{1-\beta_i} = \log \frac{\omega_i}{1-\omega_i}
\]

For independent variables, \( \omega_1 = \beta_1 \) and hence \( B_i = 0 \).
For the second-order Markov chain (dependence on the four nearest neighbors), let

\[ \alpha_1 = p(x_i = 1/x_{i-2} = 0, x_{i-1} = 0) \]
\[ \beta_1 = p(x_i = 1/x_{i-2} = 0, x_{i-1} = 1) \]
\[ \gamma_1 = p(x_i = 1/x_{i-2} = 1, x_{i-1} = 0) \]
\[ \delta_1 = p(x_i = 1/x_{i-2} = 1, x_{i-1} = 1) \] \hspace{1cm} (20)

Since \( x_i = 0 \) for \( i < 1 \) and \( i > n \), \( \alpha_1 = p(x_1 = 1) \), \( \beta_2 = p(x_2 = 1/x_1 = 0) \), \( \beta_2 = p(x_2 = 1/x_1 = 1) \) 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:

\[ p(x_1x_2 \ldots x_n) = p(x_1)p(x_2/x_1)p(x_3/x_1x_2) \ldots p(x_n/x_{n-2}x_{n-1}) \] \hspace{1cm} (21)

Taking logarithms and collecting terms, we obtain

\[ \log p(x_1x_2 \ldots x_n) = A_0 + \sum_{i=1}^{n} A_i x_i + \sum_{i=2}^{n} B_i x_{i-1} x_i \]
\[ + \sum_{i=3}^{n} C_i x_{i-2} x_i + \sum_{i=3}^{n} D_i x_{i-2} x_{i-1} x_i \] \hspace{1cm} (22)
where

\[
A_0 = \sum_{i=1}^{n} \log (1 - \alpha_i)
\]

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

\[
B_i = \log \frac{\beta_i}{1 - \beta_i} - \log \frac{\alpha_i}{1 - \alpha_i} + \log \frac{1 - \delta_{i+1}}{1 - \delta_{i+1}} - \log \frac{1 - \beta_{i+1}}{1 - \beta_{i+1}}.
\]

\[
C_i = \log \frac{\gamma_i}{1 - \gamma_i} - \log \frac{\alpha_i}{1 - \alpha_i}.
\]

\[
D_i = \log \frac{\delta_i}{1 - \delta_i} + \log \frac{\gamma_i}{1 - \gamma_i} - \log \frac{\beta_i}{1 - \beta_i} + \log \frac{\alpha_i}{1 - \alpha_i}.
\]

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 \( x_i \), \( x_{i-1}x_i \), \( x_{i-2}x_{i-1}x_i \), \( x_{i-3}x_{i-2}x_{i-1}x_i \), \( x_{i-4}x_{i-3}x_{i-2}x_{i-1}x_i \), and \( x_{i-5}x_{i-4}x_{i-3}x_{i-2}x_{i-1}x_i \) 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 logarithm 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 crimped as possible. An example of such a curve is illustrated in Figure 2. The limit of the curves as \( m \to \infty \) is a continuous curve, called a space-filling curve, that passes through every point of a given area. The curve 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 $2^r$ 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 $n \times n$ 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 $\theta_i$'s are independent and identically
distributed and examines convergence of the component Bayes risk to
$R(\theta)$ for the case $G(\theta_i) = \bigvee_{k=1}^N q(\theta_k)$. It has been assumed in the non-
parametric 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 $t_k = t_k(i/\theta_{k-1}, x_k)$ is used. One may use
$\bar{x}_{k-1}$ and $\bar{x}_{k-1}$ to obtain consistent estimates of $p_i(x)$ for all $i \in \mathcal{L}$
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 [34] uses "Window Caponery" [20, 18].

Van Ryzin [34] has shown that his above-mentioned procedure \( t_k(j/\theta_{k+1}, \theta_k) \) is also optimum in the limit in the sense that the compound risk converges to \( R(q^N) \) for any sequence \( \theta_N \). 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 \( \theta_k \) is given to us after each decision 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 \( \mathbf{x}_m = (x_1^m, \ldots, x_n^m) \) with known classification, i.e.,

\( \theta_m = (\theta_1^m, \ldots, \theta_n^m) \) is known. There is then a sequence of test samples.
\( x_{m+1}^N, \ldots, x_{m+N}^N \) with unknown \( \theta_{m+1}^N, \ldots, \theta_{m+N}^N \).

What really matters is convergence of the compound risk \( R(\theta_{m+1}^N, \bar{t}^N_N) \) 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 \( t^N_N = \{t_1^N, \ldots, t_N^N\} \) where \( t_k = t_k(j/\sum_{m=1}^{m+k} m^k) \) (its non-sequential counterpart has \( t_k = t_k(j/\sum_{m=1}^{m+k} m^k) \)). 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: \( R(\theta_{m+1}^N, \bar{t}^N_N) \rightarrow R(q^N) \). If we assume a cost of sampling \( C_i \) for each design sample from class \( i \), the compound risk for the rule \( \bar{t}^N_N \) with \( t_k = t_k(j/\sum_{m=1}^{m+k} m^k) \) is

\[
R_m(\theta_{m+1}^N, \bar{t}^N_N) = R(\theta_{m+1}^N, \bar{t}^N_N) + \frac{1}{N} \sum_{i=1}^{m} C_{i1}.
\]

For finite \( m \), \( R_m(\theta_{m+1}^N, \bar{t}^N_N) \rightarrow R(q^N) \) if and only if \( R(\theta_{m+1}^N, \bar{t}^N_N) \rightarrow R(q^N) \).
Thus, we wish to know under what circumstances it is possible to find a rule such that for some $\Theta_m$

$$R(\Theta_N^m, \Sigma_N) \rightarrow R(\eta^N)$$

uniformly for all $\Theta_N^m \in \Omega_N^N$. 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 TWC-THRESHOLD METHOD

We return to the problem of Section 1.4, page 15, where \( \Lambda = \mathbb{J} = \{0,1\} \) and only the a priori probability \( q = q_1 \) is unknown (\( q_0 = 1-q_1 \)). The sequential compound rule of Equation (28), Section 1.4, which chooses class 1 if

\[
\frac{p_1(x_k)}{p_0(x_k)} > \frac{L_{01}}{L_{10}} \frac{1 - \hat{q}_{k-1}}{\hat{q}_{k-1}}
\]

is asymptotically optimum when \( \hat{q}_k \) is a consistent estimator of \( q_1 \).

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.,

\[
\hat{q}_{k-1}(\bar{x}_{k-1}) = \left\{ h_{k-1}(\bar{x}_{k-1}) \right\}_{\text{Truncated}},
\]

\[
h_k(x_k) = \frac{1}{k} \sum_{i=1}^{k} h(x_i),
\]

and \( h(x_k) \) is a bounded unbiased estimator of \( \theta_k \). Since \( k\bar{h}_k = \bar{\theta}_k \) and \( \operatorname{Var} h_k \leq (1/k) \max_i \operatorname{Var} h(x_i) \), \( h_k \) is a consistent estimator of \( \theta_k \). 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

41

Approved for Public Release
Contraills

\[ h(x) = \frac{f_T(x)}{P_1(T)} P_0(T), \]  \quad (4)

where \( T \) is a set in \( S \) for which \( P_0(T) \neq P_1(T) \) and \( f_T(x) \) is the characteristic function of \( T \). Since

\[ E(f_T) = P_0(T) = \theta P_1(T) + (1-\theta) P_0(T), \]  \quad (5)

it follows that \( E(h(x)) = \theta \). (We have omitted the subscript \( k \) from the class \( \theta_k \) and the pattern \( x_k \), \( \theta \) 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:

\[ T = \left\{ x: \frac{p_1(x)}{p_0(x)} > \tau \right\}, \]  \quad (6)

Since \( E(f_T(x)) = P_0(T) \), we have \( E(h(x)) = \theta \), \( E(h_k) = \theta_k \), and

\[ \text{Var}(f_T(x)) = P_0(1-P_0) = \theta_1 P_1(1-P_1) + (1-\theta_1) P_0(1-P_0). \]  

Thus

\[ \sum_{i=1}^{k} \frac{\text{Var}(f_T(x_i))}{k^2 \left[ P_1 - P_0 \right]^2} = \frac{\theta_k P_1(1-P_1) + (1-\theta_k) P_0(1-P_0)}{k(P_1-P_0)^2}. \]  \quad (7)

Hence

\[ k \text{ Var } h = \frac{\theta_k (P_1-P_0)^2 + P_0(1-P_0)}{(P_1-P_0)^2}. \]
Let $T^+$ minimize $\text{Var } h_k$. 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

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

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 $k \text{Var } h_k^* \leq k \text{Var } h_k^+$.

But, since $\text{Var } h_k^+$ was assumed to be the minimum, $\text{Var } h_k^* = \text{Var } h_k^+$ and the set $T^*$ minimizes the variance.

From (7) we see that to minimize the maximum (over $\theta_k$) variance of $h_k$, we must choose $\tau$ 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 ($1_{ij} = 1 - \delta_{ij}$ 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, $\tau$, is set to give a minimax probability of error. Letting $\tau_{(S_{k-1})}$ be the fraction of times the fixed threshold was exceeded, we use

\[ u. \quad \text{This proof is due to T. J. Harley, Jr., Philco Corporation.} \]
\[
\hat{q}_{k-1}(x_{k-1}) = \frac{\mathbb{I}(x_{k-1}) - P_0(T)}{P_1(T) - P_0(T)}
\]

Truncated

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 \( k+1 \)st decision, given the first \( k \) patterns is

\[
\mathbb{E}_q \left[ R(q, \hat{t}_{k+1}) / f(q) \right] = \int_0^1 R(q, \hat{t}_{k+1}) f(q) \, dq.
\]

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

\[
\frac{P_1(x_{k+1})}{P_0(x_{k+1})} > \frac{L_{01} - L_{00}}{L_{10} - L_{11}} \frac{1 - \hat{q}_k(x_k)}{\hat{q}_k(x_k)}.
\]
where

\[ \hat{\theta}_k = E_q\left( \theta_k\right) = \int q \cdot f(q | x_k) \, dq. \]  

(11)

The density used in (11) to compute \( \hat{\theta}_k \) can be defined iteratively in terms of \( f(q | x_{k-1}) \) and \( \hat{\theta}_{k-1} \):

\[
f(q | x_k) = \frac{p(x_k, q | x_{k-1})}{p(x_k | x_{k-1})} = \frac{p(x_k/q) \cdot f(q | x_{k-1})}{\int p(x_k/q) \cdot f(q | x_{k-1}) \, dq}
\]

(12)

\[
= \frac{\left[ q \cdot p_1(x_k) + (1-q) \cdot p_0(x_k) \right] \cdot f(q | x_{k-1})}{\hat{\theta}_{k-1} \cdot p_1(x_k) + (1-\hat{\theta}_{k-1}) \cdot p_0(x_k)}
\]

where \( f(q | x_k) = f(q) \). The denominator in (12) is a normalization factor to insure \( \int f(q | x_k) \, dq = 1 \). In closed form,

\[
\hat{\theta}_k(x_k) = \frac{\int \int q \cdot f(q) \cdot f(x_{k-1} | x_k) \cdot \left[ q \cdot p_1(x_k) + (1-q) \cdot p_0(x_k) \right] \, dq}{\int \int q \cdot f(q) \cdot f(x_{k-1} | x_k) \cdot \left[ q \cdot p_1(x_k) + (1-q) \cdot p_0(x_k) \right] \, dq}
\]

(13)

The first procedure (Section 3.1) for obtaining \( \hat{\theta}_k \), though consistent, is not very efficient for small \( k \); \( \hat{\theta}_1(x_1) \), for example, can only take the values 0 or 1. Thus, \( t_2 \hat{\theta}_1(1/x_2) \) does not even depend on \( x_2 \). The estimate \( \hat{\theta}_k \) 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
\[ a_i = \frac{p_i(x_i)}{p_0(x_i)} - 1 \]

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

\[
\hat{q}_k = \frac{\int q \sum_{i=1}^{k} \left[ 1 + a_i q \right] dq}{\int q \sum_{i=1}^{k} \left[ 1 + a_i q \right] dq}.
\] (14)

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

\[
\frac{1}{k+2} \leq \hat{q}_k(x_k) \leq \frac{k+1}{k+2}.
\] (15)

The choice of a uniform a priori density for \( q \), leading to \( \hat{q}_0 = 1/2 \),

is reasonable when the Bayes envelope has its maximum at \( q = 1/2 \), since the first decision is then minimax.

In this chapter, the densities, \( p_i(a) \) were assumed to be known.

In the following chapters, we treat the case where the a priori distribution, \( G(q_k) \), is also known.
4.1 USE OF CONTEXT IN PRINT READING

The objective here is to obtain an 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 $\theta$ be a character of the language (including a "space"), i.e., a pattern class or a point in the parameter space, $\mathcal{F}$. 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 $\theta$). For a given $x$, $p(x/\theta)$ is called the likelihood of $\theta$. 
Assuming a "constant" loss matrix \( L_{ij} = 1 - \delta_{ij} \), the Bayes decision (minimum risk) is to choose that character \( \theta \) which maximizes either

\[
p(\theta | x) = \frac{p(x | \theta) p(\theta)}{p(x)}
\]  

(1)

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

When the a priori probability of \( \theta \) depends on the context, \( c \), (assuming that \( x \) is independent of \( c \) when \( \theta \) is given, \( p(x | \theta, c) = p(x | \theta) \))

Equation (1) becomes

\[
p(\theta | x, c) = \frac{p(x | \theta) p(\theta | c)}{p(x | c)}
\]  

(2)

so that the decision is based on the product

\[
p(x | \theta) p(\theta | c)
\]  

(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 \( \theta_k \) denotes the \( k \)th character, and if a second-order chain is assumed, then
p(θ_k|all the other characters) = p(θ_k|θ_{k-2}, θ_{k-1}, θ_{k+1}, θ_{k+2}). \quad (4)

For sequential (rather than compound) processing

p(θ_k|the previous characters) = p(θ_k|θ_{k-2}, θ_{k-1}). \quad (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 preceding 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(θ_k|the previous 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 $\theta_k$, given the first $k$ observations
\(x_1, \ldots, x_k\), does not depend on the unknown values of \(\theta_1, \ldots, \theta_{k-1}\) nor on the decisions about them.

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

Let \(\Omega = \{1, 2, \ldots, r\}\) be a set of states of nature and \(A = \{1, 2, \ldots, s\}\) 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 \(\theta_N = (\theta_1, \ldots, \theta_N)\) of states of nature and a corresponding vector \(X_N = (x_1, \ldots, x_N)\) of random variables, where \(\theta_k\) denotes the state of nature in the \(k\)th component problem, and the probability density of \(x_k\) is \(p_{\theta_k}(x_k)\). For a given \(\theta_k\), \(x_k\) is independent of the other \(x_i\)'s and \(\theta_i\)'s.

\[
p(x_k/x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_N; \theta_k) = p(x_k/\theta_k) = p_{\theta_k}(x_k),
\]

(6)

and hence

\[
p(x_k/\theta_k) = \frac{1}{k} \sum_{i=1}^{k} p(x_i/\theta_i).
\]

We do not assume that the \(\theta_i\)'s are independent.

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

\[
R(\theta_N, L_N) = \frac{1}{N} \sum_{k=1}^{N} R(\theta_N, t_k) ,
\]

whereby Equation (13) of Chapter 1, the \( k \)th component risk is

\[
R(\theta_N, t_k) = \int \sum_{j=1}^{a} \log p_{j,k}^N(j/x_k) \ p_{j,k}^N(\theta_N) \ dx_N
\]

\[
- \int \sum_{j=1}^{a} \log p_{j,k}^N(j/x_k) \ p(\theta_j, \theta_N) \ dx^j = R(\theta_j, t_k) .
\]

We assume that \( p_i(x) \) is known for all \( i \in \mathcal{N} \), but that none of the \( \theta_i \)s is known.

The compound Bayes risk with respect to an a priori distribution \( C(\theta_N) \) over \( \mathcal{N}^N \) is

\[
\tilde{R}(G, L_N) = \sum_{\theta_N \in \mathcal{N}^N} \int_{\mathcal{N}^N} R(\theta_N, L_N) \ C(\theta_N) = \frac{1}{N} \sum_{k=1}^{N} \tilde{R}(G, t_k) ,
\]

where

\[
\tilde{R}(G, t_k) = \sum_{\theta_N \in \mathcal{N}^N} \int_{\mathcal{N}^N} R(\theta_N, t_k) \ C(\theta_N) .
\]

A procedure is compound Bayes against \( G \) when it minimizes \( \tilde{R}(G, L_N) \). Thus, the sequential compound Bayes procedure \( L_N^G \) is the one that minimizes the \( k \)th component Bayes risk, for every \( k \).
\[ \tilde{R}(G, t_k) = \sum_{\theta_N} R(\theta_N, t_k) \ G(\theta_N) = \sum_{\theta_k} R(\theta_k, t_k) \ G(\theta_k) \]

\[ = \int \sum_{j=1}^{m} \sum_{\theta_k} l_{\theta_k} \ p(x_k/\theta_k) \ G(\theta_k) \ t_k(j|x_k) \ dx_k, \]

where \( G(\theta_k) \) is the (marginal) a priori distribution over \( \Omega_k \). Hence \( t_k(j|x_k) = 1 \) for that \( j \) which minimizes the quantity

\[ Q = \sum_{\theta_k} l_{\theta_k} \ p(x_k/\theta_k) \ G(\theta_k) \ \cdot \]

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 \( p(x_k, \theta_k) = p(x_k/\theta_k) \ G(\theta_k) \)

\[ Q = \sum_{\theta_k} l_{\theta_k} \ p(x_k, \theta_k) = \sum_{\theta_k} l_{\theta_k} \ p(x_k, \theta_k) \ \cdot \]

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

\[ G(\theta_k|x_k) = \frac{p(x_k, \theta_k)}{p(x_k)} \]

52

Approved for Public Release
since the denominator, \( p(x_k, \theta_k) = \sum_{\theta_k} p(x_k/\theta_k) G(\theta_k) \), is independent of \( \theta_k \).

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

\[
p(x_k, \theta_k) = p(x_k/\theta_k) p(x_{k-1}, \theta_k) = p(x_k/\theta_k) \sum_{\theta_{k-1}} p(x_{k-1}, \theta_{k-1}) G(\theta_k)
\]

\[
= p(x_k/\theta_k) \sum_{\theta_{k-1}} G(\theta_{k-1}/\theta_{k-1}) p(x_{k-1}, \theta_{k-1}).
\]

If the states of nature form a Markov chain,

\[
G(\theta_k/\theta_{k-1}) = G(\theta_k/\theta_{k-1}) \quad (12)
\]

with known transition probabilities \( G(\omega/\nu) \), for \( \omega \) and \( \nu \) in \( \Omega \), then

\[
p(x_k, \theta_k) = p(x_k/\theta_k) \sum_{\theta_{k-1}} G(\theta_k/\theta_{k-1}) p(x_{k-1}, \theta_{k-1}) \quad (13)
\]

Note that \( p(x) \) and \( G(\omega/\nu) \) are known and \( p(x_{k-1}, \theta_{k-1}) \) is defined recursively in terms of \( p(x_1, \theta_1) = p(x_1, \theta_1) \). The \( k \)th decision \( \gamma_k G(j/x_k) \), depends on \( x_k \) and a function of \( \theta_{k-1} \) which had already been calculated in order to make the previous decision. It does not depend on the (unknown) value of \( \theta_{k-1} \) nor on the decision about it.

We incidentally note that if the states of nature are independent,

\[
G(\theta_k/\theta_{k-1}) = g(\theta_k),
\]

then
p(x_k, \theta_k) = p_{\theta_k}(x_k) q(\theta_k) p(x_{k-1}).

Since \( p(x_{k-1}) \) does not involve \( \theta_k \), we have a simple decision problem; the quantity to be minimized is
\[
\sum_{i=1}^f L_{ij} p_i(x) q(i).
\]

Example: For a \( 2 \times 2 \) loss matrix \( A = \mathcal{R} = \{0,1\} \) (a binary alphabet), with \( w = (L_{10} - L_{11})/\sqrt{L_{00}L_{11}} \), we decide \( \theta_k = 1 \) if \( p(x_k, 1) > p(x_k, 0) \), where (when the \( \theta_k \)'s form a Markov chain)
\[
p(x_k, 1) = p_i(x_k) \left[ G(i/0) p(x_{k-1}, 0) + G(i/1) p(x_{k-1}, 1) \right], i = 0, 1; k = 2, 3, \ldots
\]
and \( p(x_1, 1) = p_i(x_1) G(i) \).

4.3 SEQUENTIAL RULES FOR MARKOV-CHAIN DEPENDENCE

For a non-randomized decision, let \( d_k(x_k) \) denote the value of \( j \in A \) for which \( t_k(j|x_k) = 1 \); the decision function \( d_k(x_k) \) maps \( X_k \) into \( A \), i.e., \( d_k(x_k) = j \in A \). Such a function \( d_k \) determines (and can be considered as) a partition of \( X_k \) into mutually exclusive sets \( T_k(j) = \{x_k : d_k(x_k) = j\} \) whose union is \( X_k \). If \( x_k \) is an element of \( T_k(j) \), action \( j \) is taken in the \( k \)th problem.

For the square loss matrix \( L_{ij} = 1 - \delta_{ij} \) action \( j \) corresponds to deciding that \( \theta_k = j \) the \( k \)th component Bayes risk
\[
\bar{R}(G, t_k) = \int_{X_k} \sum_{j=1}^n \sum_{\theta_k} L_{ij} p(x_k, \theta_k) f_t(x_k) dx_k
\]
becomes the probability of error, \( e_k = \Pr \{d_k \neq \theta_k\} \).
\[ e_k = R(G, z_k) = 1 - \frac{k}{\sum_{j=1}^{k} \int_{\mathcal{T}_k(j)} p(x_k, j) \, dx_k}, \]

where \( p(x_k, j) \) for a first-order Markov chain is given by

\[ p(x_k, \theta_k) = p(x_k/\theta_k) \sum_{\theta_{k-1}} G(\theta_k/\theta_{k-1}) \, p(x_{k-1}, \theta_{k-1}). \] (13)

The error probability is minimized by choosing \( \theta_k(x_k) \) equal to the value of \( \theta_k \) that maximizes the expression in Equation (13), that is,

\[ \mathcal{T}_k(j) = \left\{ x_k : p(x_k, j) \geq p(x_k, \theta_k) \text{ for all } \theta_k \in \mathcal{H} \right\}. \]

Since \( p(x_k, \theta_k/x_{k-1}) = p(x_k, \theta_k)/p(x_{k-1}) \), this rule is equivalent to choosing \( \theta_k(x_k) \) equal to the value of \( \theta_k \) that maximizes

\[ p(x_k, \theta_k/x_{k-1}) = p(x_k/\theta_k) \, G(\theta_k/x_{k-1}) \] (14)

An alternate (sub-optimal) rule (suggested in Section 2.1) would be to choose \( \theta_k^*(x_k) \) as equal to the value of \( \theta_k \) that maximizes

\[ p(x_k, \theta_k/x_{k-1}) = p(x_k/\theta_k) \, G(\theta_k/d_k). \] (15)

Letting \( e_k \) and \( e_k^* \) represent the error probabilities for \( \theta_k \) and \( \theta_k^* \) respectively, we have \( e_k \leq e_k^* \). 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 \( \theta^*_1(x_k) \) equal to the value of \( \theta_k \) that maximizes

55

Approved for Public Release
\[ p(x_k, \theta_k) = p(x_k | \theta_k) \mathcal{C}(\theta_k). \]  

(16)

We note that for \( k = 1 \), \( d_1(x_1) = d_1(x_1) = d_1(x_1) \) and hence \( e_1^+ = e_1 \).

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 \( \theta_k \) as well as \( x_k \), the minimum error probability of a rule \( d_k(\theta_k, x_k) \) is zero, as is easily seen by letting \( d_k(\theta_k, x_k) = \epsilon_k \).

We define the general \( k \)th component risk as

\[
\overline{R}(G, t_k) = \int_{\mathcal{S}^N} \sum_{j=1}^{a} \sum_{\theta_N} L_{\theta_N} t_k(j/\theta_N, x_N) p(x_N, \theta_N) \, dx^N
\]

(17)

where \( t_k(j/\theta_N, x_N) \) is a general decision rule. For the case where \( A = \mathcal{I} \) and \( L_{\theta} = 1 \), \( \overline{R}(G, t_k) \) is the probability of error for the \( k \)th problem:

\[
\overline{R}(G, t_k) = \int_{\mathcal{S}^N} \sum_{\theta_N} \sum_{j=1}^{a} \delta_{\theta_N} t_k(j/\theta_N, x_N) p(x_N, \theta_N) \, dx^N
\]

\[
- \int_{\mathcal{S}^N} \sum_{\theta_N} \sum_{j=1}^{a} \delta_{\theta_N} t_k(j/\theta_N, x_N) p(x_N, \theta_N) \, dx^N
\]

\[
= \int_{\mathcal{S}^N} \sum_{\theta_N} t_k(\theta_N/\theta_N, x_N) p(x_N, \theta_N) \, dx^N.
\]

(18)

For a non-randomized decision rule, \( t_k(\theta/\theta_N, x_N) \) is the characteristic function of the set of \( (\theta_N, x_N) \in \mathcal{J}^N \times \mathcal{S}^N \) for which \( \theta_N = \theta_k \). Since the

56

Approved for Public Release
The average value of the characteristic function of a set is the probability of the set

$$\tilde{R}(G, t_k) = 1 - \Pr \{ d_k = \theta_k \}. \quad (19)$$

By identifying a randomized decision rule $t_k$ with an equivalent probability distribution over the class of decision functions $d_k(\theta_N, x_N)$,

$$t_k(j/\theta_N, x_N) = \Pr \{ d_k \neq j/\theta_N, x_N \},$$

the same result:

$$\tilde{R}(G, t_k) = \Pr \{ d_k \neq \theta_k \} \quad (20)$$

is obtained.

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

$$\tilde{R}(G, t_k) = \int_{\mathbb{S}^k} \sum_{j=1}^{N_k} \sum_{\theta_k} t_{\theta_k} j \cdot P(\theta_k, \theta_{k-1}) t_k(j/\theta_{k-1}, x_k) \, dx_k. \quad (21)$$

Obviously,

$$\min_{t_k(j/\theta_{k-1}, x_k)} \tilde{R}(G, t_k) \leq \min_{t_k(j/\theta_{k}, x_k)} \tilde{R}(G, t_k), \quad (22)$$

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

\[ R(G, t_k) = \int_{S_k} \theta_k \sum_{\theta_{k-1}} G(\theta_{k-1}) \sum_{j=1}^{\theta_k} L_{\theta_k,j} p(x_k | \theta_k / \theta_{k-1}) t_k(j / \theta_{k-1}, x_k) \, dx_k \]

\[ = \sum_{\theta_{k-1}} G(\theta_{k-1}) \int_{S_k} p(x_k / \theta_{k-1}) \sum_{j=1}^{\theta_k} \sum_{\theta_k} L_{\theta_k,j} p(x_k / \theta_{k-1}) \]

the sequential Bayes rule for known \( \theta_{k-1} \) has \( t_k(j / \theta_{k-1}, x_k) = \)

\[ t_k(j / \theta_{k-1}, x_k) = 1 \text{ for the value of } j \text{ that minimizes} \]

\[ \sum_{\theta_k} L_{\theta_k,j} p(x_k / \theta_k) \frac{G(\theta_k / \theta_{k-1})}{G(\theta_{k-1})} \cdot \] (23)

For the case of a Markov chain, \( t_k(j / \theta_{k-1}, x_k) = t_{k,j}(j / \theta_{k-1}, x_k) = 1 \)

for the value of \( j \) that minimizes

\[ \sum_{\theta_k} L_{\theta_k,j} p(x_k / \theta_k) G(\theta_k / \theta_{k-1}) \cdot \] (24)

For this \( t_k \) (call it \( t_k^* \))

\[ \tilde{R}(G, t_k^*) = \sum_{\theta_{k-1}} G(\theta_{k-1}) \int_{S_k} \sum_{j=1}^{\theta_k} \sum_{\theta_k} L_{\theta_k,j} p(x_k / \theta_k) G(\theta_k / \theta_{k-1}) \]

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

\[ = \sum_{j=1}^{\theta_k} G(j) \int_{S_k} \sum_{l=1}^{\theta_k} \sum_{i=1}^{\theta_k} L_{ij} p_i(x) G(l / \gamma) \, t(j / \gamma, x) \, dx \cdot \] (25)

58
where \( f(y, x) = 1 \) for the value of \( j \) that minimizes

\[
\sum_i L_{ij} P_i(x) G(i/y)
\]

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

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

\[
c_k \leq c_k
\]

where \( c_k \) is the minimum probability of error for the \( k^{th} \) decision when \( \theta_{k-1} \) is known. The sequential Bayes rule for known \( \theta_{k-1} \) is nonrandomized, with \( d_k(\theta_{k-1}, \theta_k) = d_k(\theta_{k-1}, x_k) \) independent of \( x_{k-1} \) and \( \theta_{k-2} \) and equal to the value of \( \theta_k \) that maximizes

\[
p(x_k/\theta_k) G(\theta_k/\theta_{k-1})
\]

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 \( e_1 \geq e_2 \geq \ldots \geq e_{k+1} \geq e_c \geq e \), 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 \( k-1^{st} \) decision is correct (i.e., when \( d_{k-1}^{T} = \theta_{k-1} \)) the probability of error is \( e_c \). The probability that \( d_{k-1}^{T} = \theta_{k-1} \) is \( 1 - e_{k-1} \). Let \( e_c \) be the probability of error.
when the \( k \)-th decision is wrong. Then

\[
\varepsilon_k^+ = \varepsilon_{k-1} + (1 - \varepsilon_{k-1}^+) \varepsilon_C ,
\]

(29)

and, since \( \varepsilon_C \) is the minimum error probability for known \( \theta_{k-1} \),

\( \varepsilon_C \leq \varepsilon_0 \).

Since \( 0 \leq e_C \leq e \leq 1 \), \( 0 \leq e_0 - e_C \leq 1 - e_C \). Thus \( e_0 - e_C = 1 \)
if and only if \( e_C = 0 \) and \( e_0 = 1 \). For this case \( e_k^+ = e_{k-1}^+ = \ldots = e_1 \).

Otherwise, \( 0 \leq e_0 - e_C < 1 \) and hence, letting \( a = e_0 - e_C \),

\[
e_{k+1}^+ = a e_k^+ + e_C \\
= a^k e_1 + (1 + a + a^2 + \ldots + a^{k-1}) e_C \\
= a^k e_1 + \frac{1 - a^k}{1 - a} e_C .
\]

Thus \( e_k^+ \) converges to a limit \( e^+ \) given by

\[
e^+ = \frac{e_C}{1 - (e_C - e^+)} .
\]

(30)

Since \( e_k \leq e_k^+ \) for all \( k \), \( e \leq e^+ \). Since

\[
e_{k+1} = (e_C - e_C) e_k^+ + e_C ,
\]

(31)

we obtain

\[
e_1 - e^+ = \left[ \frac{1 - (e_C - e_D)}{1 - (e_C - e^+)} \right] e_1 - e_C = \frac{e_1 - e^+}{1 - (e_C - e^+)} .
\]

(32)
Hence, $c^+ < c_1$ if and only if $c_2^+ < c_1$.

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

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

\[
d_2^+(x_k) = 1 \text{ if } \frac{p_1^+(x_k)}{p_0(x_k)} > \frac{1 - G(1/d_k^+)}{G(1/d_k^+)}, \tag{15a}
\]

and by (14)

\[
d_k(x_k) = 1 \text{ if } \frac{p_1(x_k)}{p_0(x_k)} > \frac{1 - G_k(1/x_k)}{G_k(1/x_k)}, \tag{14a}
\]

where $G_k(1/x_k) \equiv G(\theta_k = 1/x_k)$. For all $k > 0$ we have

\[
ee_c \leq c_k \leq c_1
\]

and

\[
ee_c \leq c^+_k \leq c_1^+.
\]

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 $c_k^+ > c_1$ for all $k > 1$, while when they are well separated $c_k^+ < c_1$.

4.4 NON-SEQUENTIAL COMPOUND RULES

A procedure is compound Bayes if it minimizes the $k$th-compoment Bayes risk.
\[
\tilde{R}(G, t_k) = \sum_{k} R(\theta_N, t_k) G(\theta_N)
\]

\[
= \int \sum_{j=1}^{N} \sum_{k} \theta_{kj} t_k(j | x_N) p(x_N | \theta_N) G(\theta_N) \, dx_N .
\]  

(33)

for each \(k\). In Section 4.2, we derived the Sequential Compound Bayes procedure. The compound Bayes procedure chooses \(t_k(j | x_N)\) equal to one for that \(j\) which minimizes

\[
\sum_{k} \theta_{kj} p(x_N | \theta_N) G(\theta_N) = \sum_{k} \theta_{kj} p(x_N, \theta_N) = \sum_{k} \theta_{kj} p(x_N, \theta_k) .
\]  

(34)

We note that the minimum \(N\) component Bayes risk is the same in the sequential and non-sequential case. The minimum \(k\) component risk for \(s < N\) for the non-sequential case is less than or equal to the corresponding minimum for the sequential case.

Let us consider the case \(N = 2\), \(A = \mathcal{A}_1\), and \(L_{ij} = 1 - \delta_{ij}\). The \(k\) component Bayes risk is equal to the probability of error

\[
\tilde{R}(G, t_k) = \varepsilon_k
\]  

(35)

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

\[
\tilde{R}(G, t_k) = \frac{\varepsilon_1 + \varepsilon_2}{2} = <e> .
\]  

(36)

62

Approved for Public Release
The compound Bayes procedure chooses $d_2$ equal to the value of $\theta_1$ that maximizes

$$p(x_2, \theta_1) = p(x_1/\theta_1) \sum_{\theta_2} p(x_2/\theta_2) G(\theta_1, \theta_2)$$ (37)

and $d_2$ equal to the value of $\theta_2$ that maximizes

$$p(x_2, \theta_2) = p(x_2/\theta_2) \sum_{\theta_1} p(x_1/\theta_1) G(\theta_1, \theta_2).$$ (38)

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

$$e_k = 1 - \sum_{\theta_k} \int_{S^2} t_k(\theta_k/x_2) p(x_2, \theta_2) \, dx^2.$$ (39)

In particular

$$e_2 = 1 - \sum_{\theta_2} \int_{S^2} t_2(\theta_2/x_2) p(x_2, \theta_2) \, dx^2.$$ (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 $G(\theta_N)$ 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, $G(\theta_N)$ is stationary, and as in Section 4.3, $e_k$ is

63
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

5.1 Sequential Rules for Markov-Chain Dependence

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

Thus

\[
p_1(s) = \phi \left( \frac{(s-\mu_1)}{\sigma} \right) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(s-\mu_1)^2}{2\sigma^2}},
\]

and there is no loss of generality in assuming \( \mu_0 = 0 \) and \( \mu_1 = \mu \).

The set \( T(q) \) in Equation (7) of Chapter 1 is given by

\[
T(q) = \left\{ s: \frac{p_1(s)}{p_0(s)} > \frac{1 - q}{q} \right\} = \{ s: \Phi \left( \frac{s-\mu}{\sigma} \right) > c \},
\]

where

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

Then

\[
P_1(T) = \int_c^\infty \phi \left( \frac{(s-\mu_1)}{\sigma} \right) ds = 1 - \Phi \left( \frac{(c-\mu_1)}{\sigma} \right),
\]

\[
P_0(T) = 1 - \Phi \left( \frac{c}{\sigma} \right) = 1 - \Phi \left( \frac{\sigma \left( \frac{1-q}{q} \right) + \mu/2\sigma}{\sigma} \right),
\]

\[
P_1(T) = 1 - \Phi \left( \frac{(c-\mu_0)}{\sigma} \right) = 1 - \Phi \left( \frac{\sigma \left( \frac{1-q}{q} \right) - \mu/2\sigma}{\sigma} \right).
\]

Letting \( m = \mu/\sigma \), the probability of error given that \( q = 1 \) is

\[
\gamma(q) = 1 - P_1(T) = \Phi \left( \frac{1/m}{q} \right) \log \left( \frac{(1-q)}{q} \right) = \mu/m \}
\]

65

Approved for Public Release
and the probability of error given that \( \theta = 0 \) is

\[
P_0(T) = \hat{\rho} \left( \frac{1}{m} \log \left( \frac{q(1-q)}{1-q} \right) - m/2 \right) = \gamma(1-q) .
\]

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

\[
q = \frac{\alpha}{1 - \alpha + \beta - \beta q}
\]

and

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

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

\[
e_1 = \tilde{R}(q, t^0) = R(q)
\]

\[
e_c = (1-q) \tilde{R}(\alpha, t^0) + q \tilde{R}(\beta, t^0)
\]

\[
e_c = (1-q) \tilde{R}(\alpha, t^0) + q \tilde{R}(\beta, t^0)
\]

where

\[
\tilde{R}(q, t^0) = q \gamma(q^0) + (1-q) \gamma(1-q^0) .
\]

We shall assume \( \alpha = 1/\beta \) so that \( q = 1/2 \). Then

\[
e_1 = \gamma(1/2) \quad (1)
\]
\[ c_c = \Psi \gamma(\alpha) + (1-\alpha) \gamma(1-\alpha) \]  \hspace{1cm} (2)

and

\[ c_e = (1-\alpha) \gamma(\alpha) + \alpha \gamma(1-\alpha). \]  \hspace{1cm} (3)

From (2) and (3) we see that \( c_c(\alpha) = c_c(1-\alpha) \), \( c_e(\alpha) = c_e(1-\alpha) \), and

\[ c_e - c_c = (1-2\alpha) \left[ \gamma(\alpha) - \gamma(1-\alpha) \right], \]  \hspace{1cm} (4)

where

\[ \gamma(\alpha) = \frac{1}{(1/m) \log ((1-\alpha)/\alpha) - m/2}. \]  \hspace{1cm} (5)

and

\[ \gamma(1-\alpha) = \frac{1}{(1/m) \log ((1-\alpha)/\alpha) + m/2}. \]  \hspace{1cm} (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 \( m = \Theta \) defined by Blackwell and Girshick [2], page 158.

From (1) we calculate

\[ e_1 = 1 - \phi(m/2), \]  \hspace{1cm} (7)

and using (5) and (6) in (2) and (3) we compute \( e_c(\alpha) \) and \( e_e(\alpha) \).

We compute \( e^T(\alpha) \) and \( e^T_k(\alpha) \) from Equations (30) and (31) of Chapter 4 with \( e_1^T = e_1 \). For \( \alpha = 1/2 \) we have \( c_c = e_1 = e_k^T = e^T = c_e \).
while for \( \delta = 0 \) (or \( \delta > 1 \)) we have \( c_c = 0 \), \( c_1 = c_\delta = \delta \), and \( c_c > 1 \).

For all other \( \delta \), the results are highly dependent on the separation, \( m \).

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

\[
\frac{m}{2} < \left( \frac{1}{m} \right) \log \left( 1 - \frac{\delta}{m} \right),
\]

so that

\[
\gamma(\delta) \approx 1 - \gamma(1-\delta).
\]

Thus

\[
e_c \approx \delta + (1 - 2\delta) \gamma(1 - \delta)
\]

and

\[
e_c = \frac{e_1 + 1 - 2e_c}{1 - (1 - 2e_c)} \cdot \frac{1}{2}
\]

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

\[
e^+ \approx \frac{e_c}{1 - (1 - 2e_c)} \cdot \frac{1}{2}
\]

and

\[
e^+ \approx e_c + (1 - 2e_c) e_1 = e_1 + e_c(1 - 2e_c) = e_1 + \left( \frac{2\delta(1/2) - 1}{m} \right) e_c > e_1.
\]

The limit \( e^+ \) has a discontinuity at \( \delta = 0 \) (and \( \delta = 1 \)) with \( e^+(0) = e_1 \)

and \( e^+(\infty) = 0, 5 \). A proof of \( \lim_{\delta \to 0} e^+(\delta) = 1/2 \), using L'Hospital's rule, is given below:

Since \( e_c(\delta) = R(\delta) \) and

\[
\overline{R}(q; e_0) = R(q) + \frac{dR(q')}{dq'} (q - q'),
\]

68

Approved for Public Release
we have

$$e_c(\alpha) = e_0(\alpha) e(\alpha) + (1-2\alpha) e_c'(\alpha),$$

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

$$e_c'(\alpha) = \gamma(\alpha) - \bar{\gamma}(1-\alpha).$$

From (2)

$$e_c'(\alpha) = \gamma(\alpha) + \gamma'(\alpha) - \bar{\gamma}(1-\alpha) - (1-\alpha) \gamma'(1-\alpha)$$

and hence

$$\alpha \gamma'(\alpha) = (1-\alpha) \gamma'(1-\alpha).$$

From (4)

$$e_c'(\alpha) = e_c'(\alpha) = \frac{1}{(1-\alpha)(\bar{\gamma}'(1-\alpha))} \left[ \gamma(\alpha) + \gamma'(1-\alpha) \right]$$

$$- 2 e_c'(\alpha) - \frac{1}{1-\alpha} \gamma'(\alpha).$$

From Equation (30) of Chapter 4

$$\lim_{\alpha \to \beta} e_\beta(\alpha) = \frac{e\beta'(0)}{e\beta'(0) - e_\beta'(0)} = \frac{1}{2},$$

since $e_\beta'(0) = 1$ and $\gamma'(0) = 0$.

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

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

we have from Equation (4)
\[ e_c - e_c = (1 - \Delta \kappa) \left( \theta \left( \frac{m}{2} + \frac{1}{m} \log \frac{1 - \kappa}{\kappa} \right) - \theta \left( \frac{m}{2} - \frac{1}{m} \log \frac{1 - \kappa}{\kappa} \right) \right) < 1, \]

so that
\[ e^+ = \frac{e_c}{1 - (e_c - e_c)} \ll e_c \]

and even
\[ e_{2}^+ = e_c + (e_c - e_c) e_{1} \ll e_c. \]

In Figures 3 through 5 we see (roughly) that while \( e_c \leq e_1 \leq e_k^+ \leq e_c \) when \( m = 1 \), \( e_c \leq e^+ \leq e_k^+ \leq e_1 \leq e_c \) when \( m = 2 \) or 4.

We further note that for \( m = 4 \) convergence of \( e_k^+ \) to \( e^+ \) is extremely rapid, with \( e_3^+ \ll e^+ \) for \( \kappa > 0.02 \). The reversal of the inequality \( e_k^+ < e_1 < e_k^+ \) which occurs for \( \kappa < 0.2 \) 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 \( \kappa \ll 10^{-3} \)). Figure 7 illustrates the relationship between \( e_1 \) and \( e_k^+ \) as a function of \( m \) for a fixed \( \kappa \).

With 0 < \( \kappa \leq 1/2 \) and \( m > 0 \), suppose
\[ e_2^+ \ll e_1. \]

Then
\[
\begin{align*}
e_c + (e_c - e_c) e_1 & \leq e_1 \\
\omega & \gamma(\kappa) + (1 - \omega) \left[ \gamma(1 - \kappa) + \gamma(1/2) (1 - 2\omega) \left( \gamma(\kappa) - \gamma(1 - \kappa) \right) \right] \leq \gamma(1/2) \\
& \left[ \omega + (1 - \omega) \gamma(1/2) \right] \left[ \gamma(\kappa) - \gamma(1 - \kappa) \right] \leq \gamma(1/2) - \gamma(1 - \kappa) \\
& \omega + (1 - 2\omega) \gamma(1/2) \leq \\
& \frac{\theta(m/2^\omega (1/m) \log((1 - \kappa)/\kappa) - \theta(m/2)}{\theta(m/2^\omega (1/m) \log((1 - \kappa)/\kappa) - \theta(m/2^\omega (1/m) \log((1 - \kappa)/\kappa))}.
\end{align*}
\]
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 $\gamma(1/2) = 1 - \alpha(m/2) < 1/2$,

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

Hence

$$1 - \alpha(m/2) = \frac{\gamma(1/2) + \alpha}{\gamma(1/2) + \alpha} = \frac{\gamma(1/2) + \alpha}{\gamma(1/2) + \alpha},$$

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

Thus, we see that for small $m$ (and/or small $\alpha$) 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 $\alpha$ is not too small), the decision rule of Equation (15) (Chapter 4) is quite adequate. For $m > 5$ (signal to noise ratio > 14 dB), $e_k$ would be indistinguishable from $e_c$ for any practical value of $\alpha$.

5.2 THE SECOND-COMPONENT SEQUENTIAL COMPOUND BAYES RISK

For a sequential compound decision rule, $t_k(j/\theta_N, \theta_N) = t_k(j/\theta_k)$, we obtain from Equation (18) of Chapter 4,

$$e_k = 1 - \sum_{\theta_N} \int_{\theta_N} t_k(\theta_k/\theta_N) p(\theta_N, \theta_N) \, d\theta_N$$

$$= 1 - \sum_{\theta_k} \int_{\theta_k} t_k(\theta_k/\theta_k) p(\theta_k, \theta_k) \, d\theta_k.$$
By Equation (13) of Chapter 4,
\[ p(x_k, \theta_k) = p(x_k / \theta_k) \sum_{\theta_{k-1}} p(x_{k-1} / \theta_{k-1}) G(\theta_{k-1}, \theta_k) \]  
(9)
for the case of a Markov chain. For \( k = 2 \) Equations (8) and (9) become
\[ x_2 = 1 - \sum_{\theta_2} t_2(x_1 / \theta_2) p(x_2 / \theta_2) \]  
(10)
and
\[ p(x_2, \theta_2) = p(x_2 / \theta_2) \sum_{\theta_1} p(x_1 / \theta_1) G(\theta_1, \theta_2). \]  
(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 \( \sigma_2 \), both the second-component sequential compound Bayes probability of error for a stationary Markov chain (Section 4.1) 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: \( \mathcal{J} = \{0, 1\} \) 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.,
\[ G(\theta_1 = 1) = G(\theta_2 = 1) = 1 / 2, \ G(\theta_2 = 1 / \theta_1 = 0) = \alpha, \text{ and } G(\theta_1, \theta_2) = G(\theta_2, \theta_1). \] Letting

77
$$T_2^*(1) = \begin{cases} x_2^*; \tau_2^*(j/x_2^*) = 1 \end{cases} \cup \begin{cases} x_2^*; p(x_2^*, \theta_2^* = 1) > p(x_2^*, \theta_2^* = 0) \end{cases}$$

we have by (10),

$$e_2 = \int_{T_2^*(1)} x_2^* dG(\theta_2^* = 0) + \int_{T_2^*(1)} x_2^* dG(\theta_2^* = 1)$$

By (11) and the assumptions

$$T_2^*(1) = \{ (x_1, x_2); p_1(x_2) [\alpha p_0(x_1) + (1 - \alpha) p_1(x_1)] > p_0(x_2) [\alpha p_0(x_1) + (1 - \alpha) p_1(x_1)] \}$$

and

$$e_2 = \int_{T_2^*(1)} x_2^* dG(\theta_2^* = 0) + \int_{T_2^*(1)} x_2^* dG(\theta_2^* = 1)$$

Let $p_1(x)$ be univariate normal with mean $\mu_1$ and variance $\sigma^2$, and with $\mu = \mu_1 - \mu_0 > 0$. Letting

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

$$T_2^*(1) = \{ (x_1, x_2); \frac{p_1(x_2)}{p_0(x_2)} > y(x_1) \} \cup \{ (x_1, x_2); x_1 > c(x_1) \} \cup \{ x_2^*; \tau_2^*(j/x_2^*) = 1 \} \cup \begin{cases} x_2^*; p(x_2^*, \theta_2^* = 1) > p(x_2^*, \theta_2^* = 0) \end{cases}$$
where
\[ c'(x_1) = \left( \sigma^2 / \mu \right) \log \left[ y(x_1) \right] + \mu / 2. \]

Then
\[ e_2 = \int_{x_1} \left[ \left( 1 - \sigma \right) p_0(x_1) + \sigma p_1(x_1) \right] \left( \int_{x_1} p_0(x_2) \, dx_2 \right) \, dx_1. \]

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

where \( m = \mu / \sigma \), we have
\[ e_2 = \int_{-\infty}^{\infty} \left[ (1-x) \phi(x) + x \phi(x-m) \right] \left( \frac{1}{m} \log \frac{\alpha \phi(x) + (1-\alpha) \phi(x-m)}{(1-\alpha) \phi(x) + \alpha \phi(x-m)} - \frac{m}{2} \right) \, dx. \]

(12)

For \( \alpha = 1/2 \), \( e_2 = \Phi(-m/2) = 1 - \Phi(m/2) \) as in Section 5.1. For \( \alpha = 0 \),
\[ e_2 = \int_{-\infty}^{\infty} \phi(x) \phi(x-m) \, dx = \Phi(-m/\sqrt{2}) = 1 - \Phi(m/\sqrt{2}) \]

for \( \sigma = 0 \).
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 $e_1 = \mathcal{H}(m/2)$. At the
other extreme, $\alpha > 0$, the compound Bayes rule for $N = 2$ gives an
average probability of error of $e_2 = \mathcal{H}(\sqrt{2m}/2)$, 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, $\langle e \rangle = (e_1 + e_2)/2$.

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

$e_c \leq e_k \leq e_1$, where

$$e_c = \sum_{i \in \Omega} G(\theta_{k+1}) \min \Pr \left\{ d_k \neq d_{k} / \theta_{k-1} \right\}$$  \hspace{1cm} (13)

$$= \sum_{i \in \Omega} q[i] R(\alpha)$$

with $\alpha = G(\theta_{k} / \theta_{k-1} = i)$. 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

\[ c^* = \min \{ c_k^* \} \leq c_k. \]

In this case

\[ c_k^* = \sum_{(\theta_{k-1}, \theta_{k+1})} G(\theta_{k-1}, \theta_{k+1}) \min \{ \text{Pr}\{ \hat{a}_k \neq a_k / \theta_{k-1}, \theta_{k+1} \} \}, \quad (14) \]

where

\[ G(\theta_{k-1}, \theta_{k+1}) = \sum_{i \in \Omega} G(\theta_{k+1} / \theta_{k-1}) G(1 / \theta_{k-1}) G(\theta_{k-1}) \]

and

\[ \text{Pr}\{ \hat{a}_k \neq a_k / \theta_{k-1}, \theta_{k+1} \} = R\{ G(\theta_{k-1}, \theta_{k+1}) \}, \quad (15) \]

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

<table>
<thead>
<tr>
<th>((\theta_{k-1}, \theta_{k+1}))</th>
<th>( G(\theta_{k-1}, \theta_{k+1}) )</th>
<th>( G(1/\theta_{k-1}, \theta_{k+1}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>((0, 0))</td>
<td>( \frac{\alpha^2 + (1-\alpha)^2}{2} = \frac{1}{2} - \alpha(1-\alpha) )</td>
<td>( \frac{\alpha^2}{\alpha^2 + (1-\alpha)^2} )</td>
</tr>
<tr>
<td>((0, 1))</td>
<td>( \alpha(1-\alpha) )</td>
<td>( 1/2 )</td>
</tr>
<tr>
<td>((1, 0))</td>
<td>( \alpha(1-\alpha) )</td>
<td>( 1/2 )</td>
</tr>
<tr>
<td>((1, 1))</td>
<td>( \frac{\alpha^2 + (1-\alpha)^2}{2} = \frac{1}{2} - \alpha(1-\alpha) )</td>
<td>( \frac{(1-\alpha)^2}{\alpha^2 + (1-\alpha)^2} )</td>
</tr>
</tbody>
</table>

* We use the fact proved in Section 2, 3 that \( G(\theta_{k-1}, \theta_{k+1}), \theta_{k-1}, \theta_{k+1} = G(\theta_{k-1}, \theta_{k+1}), \) Equation (14) follows from Equation (13) of Chapter 4 exactly as (13) followed from Equation (23) of Chapter 4.
Thus, for the normal case under discussion,

\[ e_c^* = \sqrt{R(\alpha^2/\beta)} + (1-\beta)R(1/2) \]

where \( \beta = \alpha^2 + (1-\alpha)^2 \). 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,

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

and

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

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

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
CONCLUSIONS

The proceeding 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 account 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

Action, 2, 3, 54

Average risk, see Bayes risk

Bayes envelope, \( R(q) \), 5-7, 11, 12, 15, 19

sequential (compound), 57-59

Bayes procedure 4-6, 11

compound 9, 10

sequential 49, 51-53, 58, 59

Bayes risk, \( \bar{R}(q, t) \) 4-7

component, \( \bar{R}(G, t_k) \) 9, 12, 38, 51, 54, 56

compound, \( \bar{R}(G, t_{nk}) \) 9, 51

Bayes rule, see Bayes procedure

Binary patterns 24-37

Character recognition 47

Class 1-3

Component problem 3, 8

Component risk, \( R(N, t_k) \) 9, 51

See also: Bayes risk, component

Compound decision problem 3, 8, 38

Compound decision rule 8, 10, 56

distribution free 37-40

sequential 2, 13, 14, 37-39, 49-51

90
Compound risk, $\mathbb{R}^{0,\infty}_{0,\infty}$ 8, 9, 11, 12, 15, 38-40
Context 11, 16, 47-49, 88, 89
Convergence of risks 12-15, 37-40
Cover, T. M. 24

Decision, see Action
Decision function 54, 57
  randomized 4-6, 57
  see also, Decision rule
Decision procedure, see Decision rule
Decision rule 4, 56
  compound, see Compound decision rule
  sequential, see Sequential procedure
  simple 8, 10
  simple symmetric 8, 10

Dependence,
  between states of nature 49, 53
  spatial 30, 35

Empirical a priori distribution 11-14, 37
Empirical Bayes approach 10, 12, 19, 37, 38
Error probability 5, 26-28, 54-57, 59
Fix-Hodges procedure 17, 23, 38, 88
Hamming distance 24, 26
Hannan, J. F. 1, 12, 41
Hart, P. E. 24
Hilbert, D. 35
Johns, M. V. 37
Likelihood ratio 6, 29, 41-43
threshold on 6, 41-44, 83
Loss 2, 4, 8
Markov chain 28-37, 48, 53-55, 58, 59
Measurement 1-3
Minimax 7, 12
Murthy, V. K. 21
Nearest neighbor 24-28
Nonparametric techniques 17, 37-40
Pattern 1-3, 28
Pattern classification 1-3, 28, 36
Pattern recognition, see Pattern classification
Probability of error 5, 26-28, 54-57, 59
Randomized strategy, see Decision function, randomized

Approved for Public Release
Rate of convergence 13
Risk 4
Bayes, see Bayes risk
Component, see Component risk
Compound, see Compound risk
Robbins, H. 1, 12, 41
Samuel, E. 13, 41
Separation between normal distributions, m 65, 67
Sequential compound decision rule, see Sequential procedure
Sequential procedure 8, 13, 14, 37, 39, 49-61
States of nature 2-4, 50
dependent 49, 53
vector (or sequence) 3, 8
Simple decision rule 8, 10
Space filling curve 35, 36
Threshold 6, 18, 41-44
Van Rynin, J. R. 13, 20, 38
Window carpentry 21-23, 38
REFERENCES


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.
<table>
<thead>
<tr>
<th>KEYWORDS</th>
<th>L1R1</th>
<th>L1R2</th>
<th>L1R3</th>
<th>L1R4</th>
<th>L2R1</th>
<th>L2R2</th>
<th>L2R3</th>
<th>L2R4</th>
<th>L3R1</th>
<th>L3R2</th>
<th>L3R3</th>
<th>L3R4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical Decision Theory</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pattern Recognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>