SEQUENTIAL DECISION SCHEMES FOR STATISTICAL PATTERN RECOGNITION PROBLEMS WITH DEPENDENT AND INDEPENDENT HYPOTHESES

by

ABUL BASHAR SHAHIDUL HUSSAIN

B. Sc., University of Engineering and Technology, Dacca, Bangladesh, 1966

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

in the Department of Electrical Engineering

We accept this thesis as conforming to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

June 1972

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the Head of my Department or by his representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Electrical Engineering

The University of British Columbia
Vancouver 8, Canada

Date Aug. 1, 1972

ABSTRACT

Sequential decision schemes for the purpose of both pattern classification and feature ordering are investigated.

An optimum compound sequential probability ratio test (OCSPRT) for recognition problems with memory in the source as well as the observation medium is developed. The results of theoretical analysis and computer simulation of the OCSPRT for a two-class problem with first order Markov dependence among the pattern classes are presented.

For multiclass recognition problems the suitability of Bayes sequential decision schemes based on one-state ahead truncation approximation, with and without on-line feature ordering, is assessed from the points of view of computational complexity and expected cost of a terminal decision. The Bayes sequential decision scheme for dependent hypothesis problems is formulated and its performance is compared with that of the optimum compound nonsequential decision scheme.

For dependent hypothesis recognition problems, compound sequential pattern recognition schemes (CSPRS) are formulated. In CSPR schemes the required additional feature is observed either on the pattern to be decided, as in the classical sequential schemes, or on any one of the neighbouring patterns. The pattern selected and the feature actually observed are the ones which provide the maximum amount of additional information. The results of analytical and experimental evaluation of the CSPR schemes are presented.

The suitability of the suboptimal sequential decision scheme with on-line ordering of features as a feature evaluation and ordering criterion is discussed. A modified on-line sequential (MOLS) decision scheme based on limited length of search is proposed as a compromise

between the additional computational complexity and improvement in the recognition performance resulting from the on-line ordering of features. The advantage of incorporating such limited length of search over available features into sequential decision schemes using a set of preordered features is also examined.

For the purpose of experimental evaluation of the various decision schemes, recognition of handprinted English text as a particular example of a pattern recognition problem was simulated on a digital computer. The handprinted characters were obtained from Munson's multiauthor data file prepared at Stanford Research Institute.

TABLE OF CONTENTS

LIST OF PRINCIPAL SYMBOLS

Vectors

Symbol		Dimension
$X^k=(x_1^k,\ldots,x_{n_k}^k)$	k th feature vector	n_k
$\overline{X}^k=(X^k,\ldots,X^1)$	Set of feature vectors	k
$\overline{X}^k(t)=\begin{cases}(X^k,\ldots,X^{k-t+1}), & k\geq t\\ (X^k,\ldots,X^1), & k\leq t\end{cases}$	Set of t features	t
$\tilde{X}^k=(x_{n_k+1}^k,\ldots,x_D^k)$	Set of features yet to be observed at state n_k	$(D-n_k)$
$\overline{Y}^k=\begin{cases}(X^k,\ldots,X^{k-r+1}), & k\geq r\\ (X^k,\ldots,X^1), & k<r\end{cases}$	Set of feature vectors with r th order Markov dependence	r
$\Omega_W=(\omega_i,\ i=1,2,\ldots,M)$	Set of pattern classes	M
$\Lambda^k=\{\lambda^k,\ldots,\lambda^{k-t}\}$	Set of likelihood vectors	ξ^k-1
$\overline{W}=\{W_i^k,\ i=1,2,\ldots,\xi^k\}$	Set of state vectors	ξ^k
$\eta(N_i^t;\ \zeta^k)$	Set of states to which transition from N_i^t is possible	ζ^k
$\Delta^{t_o}=\begin{cases}(\delta^k,\ldots,\delta^{k-t_2}), & k\geq t_2\\ (\delta^k,\ldots,\delta^1), & k<t_2\end{cases}$	Set of optimum terminal decisions	t_o
$\{I_u\}=(i,j,\ldots,\ell)$	Set of indices of the state vectors combining the pattern class ω_u^k	ξ^k
$\Pi_b(q)=(i,j,\ldots,\ell)$	Set of q indices of feature components	q
$\overline{\Pi}_b(q)$	Set of (D-q) indices of features yet to be observed at state q	(D-q)
$\overline{\Pi}_b'(q)\varepsilon\overline{\Pi}_b(q)$	Set of indices of first $j\leq(D-q)$ features yet to be observed at state q	j

Symbol		Dimension
$\bar{d}=\{d_o,d_1,\ldots,d_M\}$	Set of decisions	M+1

Discrete Variables

Symbol	
D	Dimensionality of a feature vector
M	Total number of pattern classes
ω_i	i th pattern class
r	Order of Markov dependence among the feature vectors
m	Order of Markov dependence among the pattern classes
t	Number of feature vectors under observation
t_1	Amount of dependency on the future patterns
γ_{n_k}	Compound probability ratio
λ_i^k	Likelihood ratio
W_i^k	State vector
$W_1^k=(\omega_1^k,\ldots,\omega_1^{k-m})$	State vector with element belonging to pattern class ω_1 only
n_k	Decision state
N_i^t	Decision state
$N_i^t(j)$	State to which transition from state N_i^t is possible through the observation of $x^j_{n_j+1}$ feature
η_i	Saving in the average number of features obtained using the OCSPRT relative to SPRT
$c(n_k)$	Cost of n_k th feature component
C_{n_k}	Total cost of feature observation in continuing the decision process to state n_k

Symbol	
d_i^k; $i=1,2,\ldots,M$	A terminal decision denoting $X^k \sim \omega_i^k$
d_o^k	A nonterminal decision
$\ell(\omega_i^k;d_j^k)$	Loss involved in making the decision d_j^k when actually $X^k \sim \omega_i$
$\langle a \rangle$	Largest integer not exceeding a
$N_f(i)$	Average number of features required per pattern for a terminal decision based on the testing of i th text
N_F	Expected value of the average number of features required per pattern for a terminal decision based on all seven texts
$\hat{x}$	Best feature to be observed at any decision state
σ_f	Standard deviation of the features' variance
$S(n_{k-1})$	Number of states in which a feature from the previous pattern is preferable to the one from the pattern under consideration
$f_j(i)$	j th feature component (according to some convenient labelling system) at i th state
U_{ij}	Total usage of the j th feature at i th state
$I[x_i;\Omega_\omega]$	Mutual information between the i th feature and set of pattern classes Ω_ω
$\overline{J}(x_\ell)$	Expected divergence for ℓ th feature component
LS	Length of search
$S_c(\alpha)$	Computational complexity involved due to on-line ordering of features in sequential decision schemes
$S(\alpha,\beta)$	Complexity associated with MOLS decision schemes due to limited length of search

Probabilities

$P(x)$	Unconditional probability of event x
$P(x\|y)$	Conditional probability of event x conditioned on event y
$p_{ji}=p(\omega_i^{k+1}\|\omega_j^k)$	State transition probability
$P_e(i)$	Error probability based on the i th text
P_E	Expected error probability based on all seven texts
$P_f(i)$	Probability of forced decision based on the first text
P_F	Expected probability of forced decision based on all seven texts
$\epsilon_{ij}=P(X\sim\omega_j\|X\sim\omega_i)$	Error probability of type I and type II
$p_i(\varepsilon)$	Bayes error probability
$E(\alpha,\beta)$	Increase in error probability due to limited length of search relative to on-line sequential decision scheme
$R_e(\alpha)$	Reduction in error probability due to on-line ordering of features

LIST OF ILLUSTRATIONS

Figure Page

LIST OF TABLES

ACKNOWLEDGEMENT

I would like to express my sincere appreciation to Professor Robert W. Donaldson for his generous counsel and constant encouragement. I am grateful to my colleagues Fred Toussaint, Jim Yan and Don Allan for many stimulating discussions. I also wish to thank Dr. J. Munson of Stanford Research Institute for making his multiauthor hand-printed data set available.

I express my appreciation to Ms. Linda Morris for assuming the onerous burden of typing the mathematical thesis and Ms. Beverly Harasymchuk and Ms. Norma Duggan for helping me prepare the final manuscript. The patience and perseverence of Ms. Margaret Lockwood, who did a wonderful job of proofreading through several sleepless nights, is greatly appreciated. I also wish to thank Mr. Herb Black for his technical assistance.

Support of this study by the Association of University and Colleges of Canada in the form of a Commonwealth Scholarship, the National Research Council of Canada under Grant NRC A-3308 and the Defence Research Board of Canada under Grant DRB 2801-30 is gratefully acknowledged.

CHAPTER I

INTRODUCTION

1.1 Statistical Pattern Classification and Sequential Decision Schemes

Given an unknown pattern (e.g., noisy waveform, English character, blood cell), which belongs to one of M classes (categories, population, or source states), the problem of a pattern classifier is to "categorize" it correctly. A set of D features (measurement, attributes, or descriptors) from a pattern is a D-tuple of numbers (a point in a D-dimensional feature space) and is called the feature vector for that pattern. The classification process is the partitioning of the feature space into M mutually exclusive regions, ψ_i, i=1, 2, ..., M, such that if an unknown pattern point is within a region, ψ_i, the pattern is classified to be a member of class ω_i.

The pattern classification problem thus involves the consideration of three fundamental aspects, namely characterization, abstraction, and generalization [1]. The characterization involves the selection of the set of features that characterizes the pattern classes and is perhaps the most arduous work in the area of pattern recognition. Once the features are available, it is necessary to choose a decision rule (or decision scheme) using all the information in hand in order to be able to classify the pattern samples of unknown classes. This process of obtaining the decision rule is defined as abstraction. The ability of the decision rule to correctly classify the patterns is termed generalization.

A major portion of this thesis is concerned with the abstraction problem and the generalization ability of the corresponding schemes. A set of features heuristically selected is assumed available. The

available information for the abstraction part is a set of patterns whose true "identities" are known, the so called design or training set. For the generalization portion of the problem there is another set of patterns with unknown identities, the so called testing set.

Depending on how the features from a pattern sample are processed, the decision rule can be of two types -- nonsequential and sequential. In nonsequential decision schemes a set of features is processed all at one time and the pattern sample of unknown class is then classified. The number of features to be used is known in advance. In sequential decision schemes the features are observed at different stages of the classification process rather than all at one time. At every stage an additional feature is observed only if the previously observed feature components are not enough (in some sense) to classify the pattern with sufficient confidence. Thus, in a sequential decision scheme just enough features are observed to enable a pattern sample to be "identified" with desired confidence.

Cost is always associated with the features in pattern recognition problems. If this cost is considerable due to elaborate equipment, excessive time, involved computation, or risky or dangerous operation (in biomedical and industrial applications), one has to resort to the sequential decision scheme, since a trade-off between the system error (misclassification) and the number of features required for classification is always possible in sequential decision schemes. In some recognition problems (detection of radar signals, communication networks) the features are available only sequentially, in which case a sequential scheme may be particularly appropriate. The human perception of patterns in a person's surroundings is considered to be based on sequential process. We look at

a pattern, and based on what information we have, we look for another "cue" and this process continues until we can identify the pattern with sufficient accuracy. A sequential decision scheme seems to be a natural way to classify patterns.

In sequential decision schemes the optimality cannot be reckoned only in terms of the probability of misclassification, as in the nonsequential case. One has to consider the cost of feature observation as well. The generalization ability is, therefore, stated in terms of the total expected cost of classification which includes the cost of feature observations and the cost (or loss) of terminal decision-error (misclassification).

If some relative costs are associated with each type of misclassification, then an optimum decision scheme is obviously the one which minimizes the total expected cost of classification. If, for example, the costs of each type of misclassification are equal, the optimum decision scheme minimizes the probability of error. Such a decision scheme minimizing the cost of classification is a realization of Bayes decision rule, and will be defined the optimum in the sequel.

1.2 Previous Research

Sequential decision theory appears in many different forms in a variety of disciplines. Even within the areas of pattern recognition and signal detection there is ample application of the sequential schemes in feature selection, parameter estimation, adaptive learning and classification. This present outline of previous research is restricted to what seems most relevant and essential to our present investigation involving pattern recognition.

Most previous work in sequential decision theory is based on Wald's sequential probability ratio test (SPRT) [2] and on Blackwell and Girshick's work on the theory of games and statistical decisions [3]. Application of SPRT has been limited to the binary hypotheses case. With this restriction the decision rule consists of determining two stopping boundaries partitioning the feature space into three mutually exclusive regions; the region of acceptance of the first (null) hypothesis, region of acceptance of second (alternate) hypothesis and the region of indecision. The process of feature observation is continued as long as the pattern point is in the region of indecision. It has been shown [4] that Wald's SPRT is optimum for two hypothesized pattern classes. In other words there is no other procedure with at least as low an error probability and with a smaller average number of features required for the terminal decision than SPRT.

Much work based on both SPRT and deferred decision theory has been carried out in the area of signal detection. Bussgang and Middelton [5] applied Wald's [2] sequential analysis to the detection of signal in noise. Birdshell and Robert [6] made an extensive study of signal detectability and sequential decision theory under the constraint of bounded observation time. The composite hypotheses case and design and evaluation of optimum receiver for signals of unknown amplitude were considered by Roberts [7].

Fu [8] applied Wald's SPRT to pattern recognition. Chen and Fu [9]-[11] used the sequential decision approach to both pattern recognition and machine learning. Recognition of handprinted English characters using SPRT was also examined by Chen and Fu [9]-[11]. A modified SPRT using time-varying stopping boundaries proposed by Anderson, Bussgang,

and Marcus [12], [13], to have a finite sequential decision scheme, was successfully applied to the recognition of patterns by Fu and Chien [14]. The SPRT has also been considered for the recognition of shapes using an integral geometry approach by Wong and Steppe [15].

Attempts have been made to generalize Wald's SPRT to multi-category problems. Based on the theory of SPRT, Reed [16] formulated the GSPRT (generalized SPRT) for multiclass recognition problems. The GSPRT and its modified versions have been investigated by Fu, Chien and Chen [9], [14]. An optimum sequential test for multi-hypothesis case has also been proposed by Robert and Mullis [17], which is based on the property called subspace separability. Recently, Palmer [18] derived a multi-hypothesis sequential test from the measure associated with the SPRT. The resultant stopping rules are simpler than those of the former multiclass sequential test procedures. Aglintsev and Ter-Saakov [19] have also proposed a method for distinguishing multiple hypotheses, using sequential analysis.

The optimum (Bayes) multiclass sequential decision scheme is essentially a backward procedure [3] and it minimizes the expected cost of a terminal decision. The use of dynamic programming as a feasible computational technique for the class of optimal, finite sequential decision schemes was first proposed by Bellman, Kalaba, and Middleton [20], [21]. Its application in the area of pattern recognition is mainly due to Fu, Cardillo and Chien [11], [22]-[25]. These authors used computer-simulated character recognition experiments to demonstrate the advantage of the sequential decision schemes over the nonsequential ones. Both on-line ordering of features and off-line selection of feature subsets using the sequential decision schemes were also considered by Fu, Chien, Cardillo,

and Nikolic [25], [27]. Lasker [28] considered a sequential decision procedure which is heuristic in nature and a suboptimal version of the proposed scheme of Fu et al. [11], [22]-[25].

1.3 Scope of the Thesis

Statistical dependence of some form among the pattern classes and feature vectors of a sequence of pattern samples is frequently observed in recognition problems and has previously been considered in designing recognition schemes [29]- [41]. The consideration of such dependence, however, is prevalent only in the area of nonsequential classification process and practically nonexistent in the area of sequential classification process.

A sequential test based on Wald's theory of SPRT [2] and Bayes compound decision theory [42], [43] is proposed in Chapter II. The optimum compound sequential probability ratio test (OCSPRT) thus developed considers memory both in the source states and the observation medium. Formulation of the decision function in general form and its solution in a recursive manner are presented in Chapter II. For the purpose of detail analysis of the OCSPRT, a two-class problem with first order Markov dependence among the pattern classes is considered in Chapter III. Also, included in Chapter III are the results and discussions of computer-simulated recognition experiments.

In most situations a Bayes sequential decision scheme (based on dynamic programming) which minimizes the total expected cost of terminal decisions are impracticable from the points of view of computation and storage requirement. However, a suboptimal decision algorithm based on one-state ahead truncation approximation is easy to implement and can be quite attractive as a finite sequential decision scheme, especially

for multiclass problems. Various sequential schemes based on one-state ahead approximation for the purpose of both classification and feature ordering are formulated in this thesis. Recognition of handprinted English texts as a particular example of a pattern recognition problem was simulated on a digital computer in order to evaluate the sequential schemes. The description of the data set, the feature extraction scheme, the preprocessing operation, and the experimental methodology are included in Chapter IV [44].

Formulation of the suboptimal sequential decision scheme based on the one-state ahead truncation approximation, for recognition problems with dependence among the pattern classes as well as the feature vectors is included in Chapter V. In order to assess the suitability of the suboptimal sequential decision scheme over the optimal nonsequential one, simple (statistically independent pattern classes) sequential and non-sequential schemes were simulated on the digital computer. The results are presented in Chapter V. The performance of compound sequential decision scheme with first order Markov dependent pattern classes are also discussed in Chapter V.

Chapter VI extends the decision scheme developed in Chapter V by allowing required additional features to be observed either on the pattern sample to be classified or on any one of the neighbouring patterns. Thus, in the presence of dependence among the pattern classes, only the best features from each pattern sample are observed with the result that observation process terminates more quickly. For the purpose of analytical evaluation of this decision scheme a two-class, first order Markov problem is considered. The derived results and their implications are presented in Chapter VII. Also shown in Chapter VII are the results of experimental simulation of three special cases of the proposed multiclass

sequential decision schemes based on simple but realistic assumptions.

In sequential decision schemes the order in which the available features are observed is usually predetermined. However, it is possible to formulate a sequential decision scheme which would always observe a feature which, along with the already observed features, would provide the maximum amount of additional information. The suboptimal sequential decision scheme with such on-line ordering of features is considered in Chapter VIII. Its admissibility from the points of view of recognition performance and computational complexity is carefully assessed. The suitability of the suboptimal sequential decision scheme with on-line ordering of features as a feature evaluatuion and ordering criterion is also discussed in Chapter VIII. A modified on-line sequential (MOLS) decision scheme based on limited length of search is proposed as a compromise between the additional computational complexity and the improvement in the recognition performance resulting from the on-line ordering of features. The advantage of incorporating such limited length of search over available features into the sequential decision schemes using a set of perordered features is also examined in Chapter VIII.

Finally, a few concluding remarks and a brief indication of possible future research are presented in Chapter IX.

CHAPTER II

OPTIMUM COMPOUND SEQUENTIAL PROBABILITY RATIO TEST FOR DEPENDENT HYPOTHESIS PROBLEMS

2.1 Introduction

In this chapter the theory of Wald's SPRT [2] and optimum (Bayes) compound decision theory [34], [42], [43] are combined to develop the optimum compound sequential probability ratio test (OCSPRT) for recognition problems with memory in the source as well as the observation medium. The process of feature observation is continued as long as the information from the pattern to be classified and the finite number of previous patterns are not enough for a final decision with desired confidence. A binary hypothesis case is first considered for the formulation of the OCSPRT. The test is then generalized for the case of multihypothesis problems at the end of the chapter.

2.2 Statement of the Problem

Let $\bar{X}^L = (X^1, X^2, \ldots, X^L)$ represent a sequence of L feature vectors available at instant 1, 2, ..., L, respectively, of the classification process. Let n_k denote the dimensionality of the k th feature vector; thus

$$X^k = \begin{bmatrix} x_1^k \\ x_2^k \\ \cdot \\ \cdot \\ \cdot \\ x_{n_k}^k \end{bmatrix}, \quad n_k = 1, 2, \ldots, D,$$

where x_i^j, $i = 1, 2, \ldots, D$, corresponds to the ith feature of the j th feature vector. A maximum of D features are assumed to be available from each pattern sample. Corresponding to these feature vectors is a series of source states or pattern classes for each k, $k = 1, 2, \ldots, L$. Each input pattern, therefore, lies in the set $\Omega_\omega = \{\omega_i;\ i = 1, 2, \ldots, M\}$. At any instant k of the sequential decision process, a <u>decision state</u> is defined in terms of the number of features observed on the pattern under consideration at that instant. Thus, for a pattern requiring n_k features for a final decision, the sequential decision process has to go through n_k discrete decision states. Let $\bar{d} = \{d_o^k, d_1^k, \ldots, d_M^k\}$ be a set of decisions at instant k, where each d_i^k, $i = 1, 2, \ldots, M$, represents that the hypothesis H_i: $X^k \sim \omega_i$ (the pattern represented by feature vector X^k is in class ω_i) is accepted. Associated with each state of the deci-

sion process is an optimum decision δ^k, such that $\delta^k = d_i^k$, $i = 1, 2, \ldots, M$ if a terminal decision is possible at that state of the process or $\delta^k = d_o^k$, a nonterminal decision*, if additional features are necessary before a terminal decision is possible.

The recognition problem under consideration is assumed to have dependency structure of the following forms:

1. Markov dependence of any finite order m may exist among the pattern classes such that

$$P(\omega_i^k | \omega_j^{k-1}, \ldots, \omega_p^1) = \begin{cases} P(\omega_i^k | \omega_j^{k-1}, \ldots, \omega_\ell^{k-m}), & k > m \\ P(\omega_i^k | \omega_j^{k-1}, \ldots, \omega_p^1), & k \leq m \end{cases}$$

$$i, j, \ldots, \ell, \ldots, p = 1, 2, \ldots, M, \quad (2.1\text{-a})$$

where $P(\cdot|\cdot)$ and $P(\cdot)$ are used to denote the conditional and the unconditional probabilities.

2. For any finite r, the feature vectors may have r th order Markov dependence among themselves, such that at any instant k

$$P(X^k | X^{k-1}, \ldots, X^1) = \begin{cases} P(X^k | X^{k-1}, \ldots, X^{k-r}), & k > r \\ P(X^k | X^{k-1}, \ldots, X^1), & k \leq r. \end{cases} \quad (2.1\text{-b})$$

3. The feature vectors may also be statistically dependent on the identities of neighbouring patterns, and therefore

$$P(X^k | \omega_i^k, \omega_j^{k-1}) \neq P(X^k | \omega_i^k). \quad (2.1\text{-c})$$

Statistical dependence of these types are not uncommon in pattern recognition problems and have been considered in the past in one form or the other [32]-[41].

*A nonterminal decision d_o is either null or consists of rejects in case of nonsequential decision schemes [45].

2.3 Derivation of the Decision Function

In this section the OCSPRT is developed for the binary pattern-class case and is based on the assumption that the dependencies extend only into the finite past. In a fixed sample size (nonsequential) compound classification scheme a pattern sample at any instant k is classified on the basis of the information available from all the pattern samples of the sequence. However, when the decision on the k th pattern cannot be delayed for additional information from the subsequent patterns, only the first k patterns of the sequence are useful for the k th decision. In the absence of all forms of dependencies in the recognition system, the compound classification scheme reduces to the simple classification scheme, where the information from the pattern under consideration alone is useful.

If cost (or loss) can be assigned to each type of misclassification, that is, if $\ell(\omega_i^k;\ d_j^k)$ is an element of an $M \times M$ loss matrix and denotes the loss due to the decision $X^k \sim \omega_i^k$ when $X^k \sim \omega_j^k$ is true, where $\ell(\omega_i^k;\ d_j^k) \geq \ell(\omega_i^k;\ d_i^k)$, then a Bayes decision rule is one which makes a decision δ^k to minimize the average cost of misclassification. Let R_k be the expected cost or risk of making the optimum decision at any instant k. Then

$$R_k = \int \sum_{i=1}^{M} \ell(\omega_i^k;\ d_j^k)\ P(\bar{X}^k | \omega_i^k)\ P(\omega_i^k) d\bar{X}^k \tag{2.2}$$

where the integration is over the k-fold cartesian product of the feature space. If $\rho_k(j)$ is defined as

$$\rho_k(j) = \sum_{i=1}^{M} \ell(\omega_i^k;\ d_j^k)\ P(\bar{X}^k | \omega_i^k) P(\omega_i^k),\ j=1,2,\ldots M, \tag{2.3}$$

then the risk R_k is minimized by minimizing (2.3). Thus the Bayes decision

$$\delta^k = d_\ell^k \text{ when } \rho_k(\ell) = \underset{j}{\mathrm{Min}}\{\rho_k(j)\},\ \ell,\ j = 1,\ 2,\ \ldots,\ M, \tag{2.4}$$

and the corresponding risk involved is given by

$$R_k^* = \int \underset{j}{\mathrm{Min}}\{ \sum_{i=1}^{M} \ell(\omega_i^k;\ d_j^k)\ P(\bar{X}^k|\omega_i^k)\ P(\omega_i^k)\} d\bar{X}^k.$$

The decision rule defined in (2.4) is not unique, since there could be more than one pattern class satisfying (2.4). However, the alternatives are equivalent since they result in the same risk.

If the elements of the loss matrix are such that

$$\ell(\omega_i^k;\ d_j^k) = \begin{cases} 1, & \text{if } i \neq j \\ 0, & \text{if } i = j \end{cases}, \quad i,j=1,2,\ldots M \tag{2.5}$$

the Bayes decision rule is equivalent to

$$\delta^k = d_\ell^k \quad \text{if:} \quad P(\bar{X}^k|\omega_\ell^k)\ P(\omega_\ell^k) > P(\bar{X}^k|\omega_i^k)\ P(\omega_i^k), \tag{2.6}$$

$$\text{for all } i = 1,\ 2,\ \ldots,\ M$$

The decision rule (2.6) for the two-class case reduces to

$$\begin{cases} \delta^k = d_2^k: & \text{if } P(\bar{X}^k;\ \omega_2^k)/P(\bar{X}^k;\ \omega_1^k) < T_o \\ \delta^k = d_1^k: & \text{otherwise}, \end{cases}$$

where T_o is the threshold. The optimum nonsequential decision rule, therefore, reduces to the likelihood ratio test. Let γ_{n_k} be defined as the <u>compound probability ratio</u> at any state n_k of the sequential decision process. Then

$$\gamma_{n_k} = \frac{P(\bar{X}^k|\omega_2^k)P(\omega_2^k)}{P(\bar{X}^k|\omega_1^k)P(\omega_1^k)}, \quad n_k = 1,\ 2,\ \ldots,\ D. \tag{2.7}$$

In a sequential classification scheme two thresholds (called <u>stopping boundaries</u>) are required, defining a zone in the feature space corresponding to the nonterminal decision d_o. Thus, if T_1 and T_2 ($T_2 > T_1 > 0$) are the two stopping boundaries, then an optimum OCSPRT consists of forming

the compound probability ratio γ_{n_k} at decision state n_k and accept

$$\delta^k = \begin{cases} d_2^k & \text{if } \gamma_{n_k} \geq T_2 \\ d_o^k & \text{if } T_2 > \gamma_{n_k} > T_1 \\ d_1^k & \text{if } \gamma_{n_k} \leq T_1 \end{cases} \tag{2.8}$$

The two stopping boundaries T_1 and T_2 can be shown [2] to be related to the two types of error probabilities, ϵ_{12} (type-I) and ϵ_{21} (type-II), where

$$\epsilon_{ij} = P(X^k \sim \omega_j | X^k \sim \omega_i), \quad k = 1, 2, \ldots; \; i, j = 1, 2, \ldots, M.$$

The feature observation process is continued as long as $\delta^k = d_o^k$.

Note that the OCSPRT defined in (2.8) is nonfinite in nature, that is, the number of features required for a terminal decision is unbounded. In most recognition problems only a fixed number of features is available such that a terminal decision has to be made at the final state ($n_k = D$). One can, of course, define a finite OCSPRT by incorporating either the truncated scheme of Wald [2] or the time-varying stopping boundaries proposed by Anderson, Bussgang and Marcus [12]-[14]. In the truncated scheme, the decision rule at the final state of the process is modified to accept

$$\delta^k = \begin{cases} d_2^k, & \text{if } |\gamma_D - T_1| \geq |\gamma_D - T_2| \\ d_1^k, & \text{otherwise.} \end{cases}$$

In the latter scheme, the stopping boundaries are modified at every state of the decision process such that both T_1 and T_2 monotonically converge. Thus the observation process is always interrupted at the end of the desired state.

2.4 Recursive Solution of γ_{n_k}

Let at any instant k, $\bar{X}^k(t) = (X^k, X^{k-1}, \ldots, X^{k-t+1})$ be the set of t feature vectors available, on the basis of which the k th pattern has to be classified. Let

$$\bar{W} = \{W_i^k;\ i = 1, 2, \ldots, \xi^k\}$$

be a set of state vectors, where each element W_i^k of the set W for m th order dependence among the pattern classes ω_i, is given by

$$W_i^k = \begin{cases} (\omega_b^k, \ldots, \omega_q^{k-m}), & k \geq m \\ (\omega_b^k, \ldots, \omega_s^1), & k < m, \end{cases}$$

$$b, \ldots q, \ldots, s = 1, \ldots, M.$$

Thus there are M^m possible values of i if k > M and M^k possible values of i if k < M, with the result that

$$\xi^k = \begin{cases} M^m, & k \geq m \\ M^k, & k < m. \end{cases}$$

For any $m \geq 1$, it may be shown that the W_i^k's are first order Markov in the sense that

$$P(W_i^k | W_j^{k-1}, \ldots, W_\ell^1) = P(W_i^k | W_j^{k-1}).$$

Finally, let the element $W_1^k \in \bar{W}$ be always such that

$$W_1^k = \begin{cases} (\omega_1^k, \ldots, \omega_1^{k-m}), & k \geq m \\ (\omega_1^k, \ldots, \omega_1^1), & k < m. \end{cases}$$

The following two relations, the detailed derivation of which is presented in Appendix-A, are assumed to be available.

If $\bar{X}^k(t)$ is the set of feature vectors available at instant k

and state n_k of the decision process, and dependence relations (2.1-a), (2.1-b) and (2.1-c) apply, then for any $r \geq m$

$$P(X^k|X^{k-1}, \ldots, X^{k-t+1}; W_i^k) = P(X^k|\bar{Y}^{k-1}; W_i^k) \tag{2.9-a}$$

where

$$\bar{Y}^k = \begin{cases} (X^k, X^{k-1}, \ldots, X^{k-r+1}), & k \geq r \\ (X^k, X^{k-1}, \ldots, X^1), & k < r \end{cases}$$

and

$$P(W_i^k|W_j^{k-1}; \bar{Y}^{k-1}) = P(W_i^k|W_j^{k-1}). \tag{2.9-b}$$

Let $\{I_j\}$ be the set of indices of W_i^k such that

$$i\varepsilon\{I_j\} \quad \text{iff} \quad \omega_j^k \varepsilon W_i^k, \; i = 1, 2, \ldots, M.$$

At any decision state n_k, we have the compound probability ratio γ_{n_k}, based on the finite set of t feature vectors

$$\gamma_{n_k} = \frac{P(X^k, X^{k-1}, \ldots, X^{k-t+1}; \omega_2^k)}{P(X^k, X^{k-1}, \ldots, X^{k-t+1}; \omega_1^k)}.$$

The above equation can be rewritten as follows:

$$\gamma_{n_k} = \frac{\sum_{i\varepsilon\{I_2\}} P(X^k, X^{k-1}, \ldots, X^{k-t+1}; W_i^k)}{\sum_{u\varepsilon\{I_1\}} P(X^k, X^{k-1}, \ldots, X^{k-t+1}; W_u^k)},$$

where the use of (2.9-a) yields

$$\gamma_{n_k} = \frac{\sum_{i\varepsilon\{I_2\}} P(X^k|\bar{Y}^{k-1}; W_i^k) P(\bar{X}^{k-1}(t-1); W_i^k)}{\sum_{u\varepsilon\{I_1\}} P(X^k|\bar{Y}^{k-1}; W_u^k) P(\bar{X}^{k-1}(t-1); W_u^k)}. \tag{2.10}$$

Since W's are first order Markov and (2.9-b) applies, (2.10) can be rewritten as follows:

$$\gamma_{n_k} = \frac{\sum_{i\varepsilon\{I_2\}} [P(X^k|\bar{Y}^{k-1};W_i^k) \quad \sum_{j=1}^{\xi^{k-1}} P(X_{(t-1)}^{k-1};W_j^{k-1})P(X_i^k|W_j^{k-1})]}{\sum_{u\varepsilon\{I_1\}} [P(X^k|\bar{Y}^{k-1};W_u^k) \quad \sum_{v=1}^{\xi^{k-1}} P(X_{(t-1)}^{k-1};W_n^{k-1})P(W_u^k|W_v^{k-1})]} \cdot \quad (2.11)$$

Equation (2.11) after suitable rearrangements of terms reduces to

$$\gamma_{n_k} = \frac{\sum_{i\varepsilon\{I_2\}} P(X^k|\bar{Y}^{k-1};W_i^k) \cdot \alpha_i^k}{\sum_{u\varepsilon\{I_1\}} P(X^k|Y^{k-1};W_u^k) \cdot \alpha_u^k} ,$$

where

$$\alpha_i^k = P(W_i^k|W_1^{k-1}) + \sum_{j=2}^{\xi^{k-1}} P(W_i^k|W_j^{k-1}) \cdot \beta_j[\bar{X}^{k-1}(t-1);W] \quad (2.12)$$

and

$$\beta_j[\bar{X}^{k-1}(t-1);W] = \begin{cases} \dfrac{P(X^{k-1},\ldots, X^{k-t+1};W_j^{k-1})}{P(X^{k-1}, \ldots, X^{k-t+1};W_j^{k-1})}, & k \geq t \\[2ex] \dfrac{P(X^{k-1},\ldots,X^1;W_j^{k-1})}{P(X^{k-1},\ldots,X^1;W_j^{k-1})}, & k < t;\ j=2,3,\ldots,\xi^{k-1}. \end{cases}$$

The probability density $P(X^k|\bar{Y}^{k-1};W_i^k)$ and the state transition probability $P(W_i^k|W_j^{k-1})$ are all assumed known. Therefore the compound probability ratio γ_{n_k} is computable from the known quantities only if $\beta_j[\bar{X}^{k-1}(t-1);W]$ is available at every instant k of the decision process. It will now be shown that $\beta_j[X^{-k-1}(t-1);W]$ is recursively computable from known information and that the required computation increases only linearly with t. The results are stated in the following theorem:

<u>THEOREM 2.1</u>: If at any state n_k of the sequential decision process $\bar{X}^k(t) = (X^k, X^{k-1}, \ldots, X^{k-t+1})$ is the set of t feature vectors available and the dependence relations stated in (2.1-a), (2.1-b) and

(2.1-c) apply, then for any $r \geq m$

$$\beta_i[\bar{X}^{k-q}(t-q);W] = \lambda_i^{k-q} \cdot \frac{P(W_i^{k-q}|W_i^{k-q-1}) + \sum_{\ell=2}^{\xi^{k-q-1}} P(W_i^{k-q}|W_j^{k-q-1})\beta_j[\bar{X}^{k-q-1}(t-q-1);W]}{P(W_i^{k-q}|W_1^{k-q-1}) + \sum_{\ell=2}^{\xi^{k-q-1}} P(W_1^{k-q}|W^{k-q-1})\beta_\ell[\bar{X}^{k-q-1}(t-q-1);W]}$$

$$\text{for,} \quad q > \begin{cases} (t+1), & k \geq t \\ 1, & k < t \end{cases}$$

and

$$\beta_i[\bar{X}^{k-q}(t-q);W] = \lambda_i^{k-q}$$

$$\text{for } q = \begin{cases} (t+1), & k \geq t \\ 1, & k < t \end{cases}$$

where $$\lambda_i^{k-q} = \frac{P(X^{k-q}|\bar{Y}^{k-q-1};W_i^{k-q})}{P(X^{k-q}|\bar{Y}^{k-q-1};W_1^{k-q})}, \quad i = 2, \ldots, \xi^{k-q}.$$

<u>PROOF</u>: for the sake of simplicity consider $k < t$; thus

$$\beta_i[\bar{X}^{k-q}(t-q);W] = \frac{P(X^{k-q}, X^{k-q-1},\ldots,X^1;W_i^{k-q})}{P(X^{k-q}, X^{k-q-1},\ldots,X^1;W_1^{k-q})},$$

$$q = 1,2,\ldots,(t+1); \; i = 2, \ldots, \xi^{k-q}$$

which can be rewritten in the following form:

$$\beta_i[\bar{X}^{k-q}(t-q);W] = \frac{P(X^{k-q}|X^{k-q-1},\ldots,X^1;W_i^{k-q}) \sum_{j=1}^{\xi^{k-q-1}} X^{k-q-1},\ldots,X^1;W_i^{k-q},W_j^{k-q-1})}{P(X^{k-q}|X^{k-q-1},\ldots,X^1;W_1^{k-q}) \sum_{\ell=1}^{\xi^{k-q-1}} P(X^{k-q-1},\ldots,X^1;W_1^{k-q},W_\ell^{k-q-1})}.$$

The W_i's form a first order Markov chain and (2.9-a) and (2.9-b) apply. Thus

$$\beta_i^{k-q}[\bar{X}(t-q);W] = \frac{P(X^{k-q}|\bar{Y}^{k-q-1};W_i^{k-q})}{P(X^{k-q}|\bar{Y}^{k-q-1};W_1^{k-q})} \cdot \left[\frac{\sum_{j=1}^{k-q-1} P(X^{k-q-1},\ldots,X^1;W_j^{k-q-1})P(W_i^{k-q}|W_j^{k-q-1})}{\sum_{\ell=1}^{\xi^{k-q-1}} P(X^{k-q-1},\ldots,X^1;W_\ell^{k-q-1})P(W_1^{k-q}|W_\ell^{k-q-1})}\right] . \tag{2.13}$$

After rearrangement of terms (2.13) can be expressed as follows:

$$\beta_i^{k-q}[\bar{X}(t-q);W] = \frac{P(X^{k-q}|\bar{Y}^{k-q-1};W_i^{k-q})}{P(X^{k-q}|\bar{Y}^{k-q-1};W_1^{k-q})} \cdot \left[\frac{P(W_i^{k-q}|W_1^{k-q-1}) + \sum_{j=2}^{\xi^{k-q-1}} P(W_i^{k-q}|W_j^{k-q-1})\beta_j^{k-q-1}[\bar{X}(t-q-1);W]}{P(W_1^{k-q}|W_1^{k-q-1}) + \sum_{\ell=2}^{\xi^{k-q-1}} P(W_1^{k-q}|W_\ell^{k-q-1})\beta_\ell^{k-q-1}[\bar{X}(t-q-1);W]}\right] , \tag{2.14}$$

$$= \lambda_i^{k-q} \cdot \left[\frac{P(W_i^{k-q}|W_1^{k-q-1}) + \sum_{j=2}^{\xi^{k-q-1}} P(W_i^{k-q}|W_j^{k-q-1}) \cdot \beta_j^{k-q-1}[\bar{X}(t-q-1);W]}{P(W_1^{k-q}|W_1^{k-q-1}) + \sum_{\ell=2}^{\xi^{k-q-1}} P(W_1^{k-q}|W_\ell^{k-q-1}) \cdot \beta_\ell^{k-q-1}[\bar{X}(t-q-1);W]}\right] .$$

One can similarly obtain (2.14) for any $k \geq t$. Due to the recurrence nature of $\beta_j^{k-q}[\bar{X}(t-q):W]$, all that is needed to be stored at any any instant $k \geq t$ is a set $\{\Lambda^{k-q};\ q = 1, \ldots, (t-1)\}$ of likelihood ratio vectors, where each vector

$$\Lambda^{k-q} = \begin{pmatrix} \lambda_2^{k-q} \\ \vdots \\ \lambda_u^{k-q} \end{pmatrix} , \qquad u = \xi^{k-q}.$$

In a recognition problem the amount of memory needed to store the probability densities $P(X^k|\overline{Y}^{k-1};W_j^k)$ is dependent on the values of parameters m and r and the amount of dependence among the feature components, but not on t. However, the OCSPRT would require additional memory to store the likelihood ratio vectors and the state transition probabilities. For $k \geq t$, this amounts to a total of $(2^m-1)\cdot t + 1$ numbers. Also, $\beta_j[\overline{X}^{k-q}(t-q);W]$ has to be computed for each $q = 1, 2, \ldots, (t-1)$ and $j = 2, 3, \ldots, \xi^{k-q}$, which requires storage of an additional (2^m-1) numbers. Therefore, the total requirement is $(2^m+1)\cdot(t+1)+1$, and for a finite m, it increases only linearly with the increase in the value of t, the length of the available set of feature vectors.

2.5 Generalized CSPRT for Multiclass Problems

The OCSPRT developed in Section 2.3 is only for the binary hypothesis case, where each pattern class ω_i is a member of the set $\Omega_\omega = \{\omega_i; i = 1, 2\}$. It is possible however to develop such a test for the general case of M hypotheses, starting from the GSPRT (Generalized SPRT) of Reed [16] instead of the SPRT of Wald [2]. In the GSPRT, the generalized probability ratio for pattern class ω_i at state n_k is computed as follows:

$$G_{n_k}(i) = \frac{P(X^k|\omega_i^k)}{[\prod_{j=1}^{M} P(X^k|\omega_j^k)]^{1/M}}, \quad i = 1, 2, \ldots, M.$$

This probability ratio is compared with the stopping boundary of the i th pattern class T_i and the sequential decision rule is to reject the pattern class ω_i from further consideration; in other words, X^k does not

belong to class ω_i if

$$G_{n_k}(i) < T_i, \quad i = 1, 2, \ldots, M.$$

The stopping boundaries are related to the error probabilties as follows:

$$T_i = \frac{1-e_{ii}}{\left[\prod_{j=1}^{M} (1-e_{ij})\right]^{1/M}} .$$

After the rejection of pattern class ω_i, the total number of classes is reduced by one. The feature observation process is continued until all but one of the pattern classes are rejected.

Let the generalized compound probability ratio at state n_k based on the set $\bar{X}^k(t)$ of t feature vectors be

$$\Gamma_{n_k}(i) = \frac{P(\bar{X}^k(t)\,|\,\omega_i^k}{\left[\prod_{j=1}^{M} P(\bar{X}^k(t)\,|\,\omega_j)\right]^{1/M}} , \quad i = 1, 2, \ldots, M.$$

If the dependence relations (2.9-a), (2.9-b) and (2.9-c) apply, then the generalized compound probability ratio can be expressed as follows:

$$\Gamma_{n_k}(i) = \frac{P(\bar{X}^k(t);\omega_i^k)}{\left[\prod_{j=1}^{M} P(\bar{X}^k(t);\omega_j)\right]^{1/M}} \cdot \frac{\left[\prod_{f=1}^{M} P(\omega_f^k)\right]^{1/M}}{P(\omega_i^k)}$$

$$= \frac{\sum_{\ell\varepsilon\{I_i\}} P(\bar{X}^k(t);W_\ell^k)}{\left[\prod_{j=1}^{M} \sum_{q\varepsilon\{I_j\}} P(\bar{X}^k(t);W_q^k)\right]^{1/M}} \cdot \frac{\left[\prod_{f=1}^{M} P(\omega_f^k)\right]^{1/M}}{P(\omega_i^k)} .$$

By following the same steps as in (2.10) and (2.11), we obtain

$$\Gamma_{n_k}(i) = \frac{\sum_{\ell \varepsilon I_i} P(X^k|\bar{Y}^{k-1};W_\ell^k) \cdot \alpha_i^k}{\left[\prod_{j=1}^{M} \sum_{q \varepsilon \{I_j\}} P(X^k|\bar{Y}^{k-1};W_q^k) \cdot \alpha_q^k\right]^{1/M}} \cdot \frac{\left[\prod_{f=1}^{M} P(\omega_f^k)\right]^{1/M}}{P(\omega_i^k)},$$

where α_q^k is defined in (2.12) and can be expressed in terms of $\beta_j[\cdot]$. Since the $\beta_j[\cdot]$s' are available recursively from known information, as proved in Theorem 2.1, the generalized compound probability ratio for class ω_i is easily computable at every state of the decision process.

Unfortunately, the GSPRT (consequently the GCSPRT) has not been shown to be optimal for $M > 2$. The designing of the stopping boundaries for proper convergence of the rejection criterion is sometimes difficult. Also the performance of the scheme deteriorates if some modification of the stopping boundaries is required to ensure termination of the feature observation process after a desired number of states (finite scheme). Therefore, for multiclass recognition problems Bayes sequential decision schemes (Chapter V) seem more suitable than those incorporating the GSPRT and the GCSPRT.

CHAPTER III

COMPOUND SPRT FOR FIRST ORDER MARKOV DEPENDENT HYPOTHESES

3.1 Description of the Model

To gain further insight into the optimum sequential probability ratio test a simpler recognition system is considered. The pattern classes in the set $\Omega_\omega = \{\omega_i,\ i = 1, 2\}$ are assumed to form a stationary first order Markov chain. Thus, at any instant k

$$P(\omega_i^k | \omega_j^{k-1}, \ldots, \omega_\ell^k) = P(\omega_i^k | \omega_j^{k-1}),\ i, j = 1, 2 . \tag{3.14}$$

At every instant k, k = 1, 2, ..., a set of t feature vectors is assumed to be available. The k th pattern has to be decided on the basis of this set of feature vectors and the available a priori information. The feature vectors are assumed conditionally independent of the neighbouring feature vectors and pattern identities. Thus

$$P(X^k | X^{k-1}, \ldots, X^1;\ \omega_i^k, \ldots, \omega_j^1) \tag{3.15}$$

$$= P(X^k | \omega_i^k)$$

Let

$$p_{ji} = P(\omega_i^{q+1} | \omega_j^q),\ q = 1, 2, \ldots;\ i, j = 1, 2, \tag{3.16}$$

therefore, the a priori probability of the pattern classes are given by

$$p_i = P(\omega_i^q) = \sum_{j=1}^{2} p_{ji} \cdot p_j,\ i = 1, 2.$$

A block diagram of the recognition system under consideration in this section is shown in Figure 3.1.

The compound probability ratio γ_{n_k} for the system under consideration now reduces to

$$\gamma_{n_k} = \lambda^k \cdot \frac{p_{12} + p_{22} \cdot \beta[\overset{k-1}{\bar{X}}(t-1);\omega]}{p_{11} + p_{21} \cdot \beta[\overset{k-1}{\bar{X}}(t-1);\ \omega]} \tag{3.17}$$

where,

$$\lambda^k = \frac{P(X^k|\omega_2^k)}{P(X^k|\omega_1^k)}$$

and,

$$\beta[\overset{k-1}{\bar{X}}(t-1);\omega] = \begin{cases} \dfrac{P(X^{k-1},\ldots,X^{k-t+1};\omega_2^{k-1})}{P(X^{k-1},\ldots,X^{k-t+1};\omega_1^{k-1})}, & \text{for } k \geq t \\[2ex] \dfrac{P(X^{k-1},\ldots,X^{1};\omega_2^{k-1})}{P(X^{k-1},\ldots,X^{1};\omega_1^{k-1})}, & \text{for } k < t\ . \end{cases}$$

The process of feature observation continues as long as γ_{n_k} is such that $T_2 > \gamma_{n_k} > T_1$ and an additional feature is available from the pattern under consideration.

It may be pointed out here that for k = t, that is, at every instant information from the entire past is assumed to be available, the compound probability ratio of (3.17) reduces to

$$\gamma_{n_k} = \lambda^k \frac{p_{12} + p_{22}\gamma_{n_{k-1}}}{p_{11} + p_{21}\gamma_{n_{k-1}}}\ . \tag{3.18}$$

However,

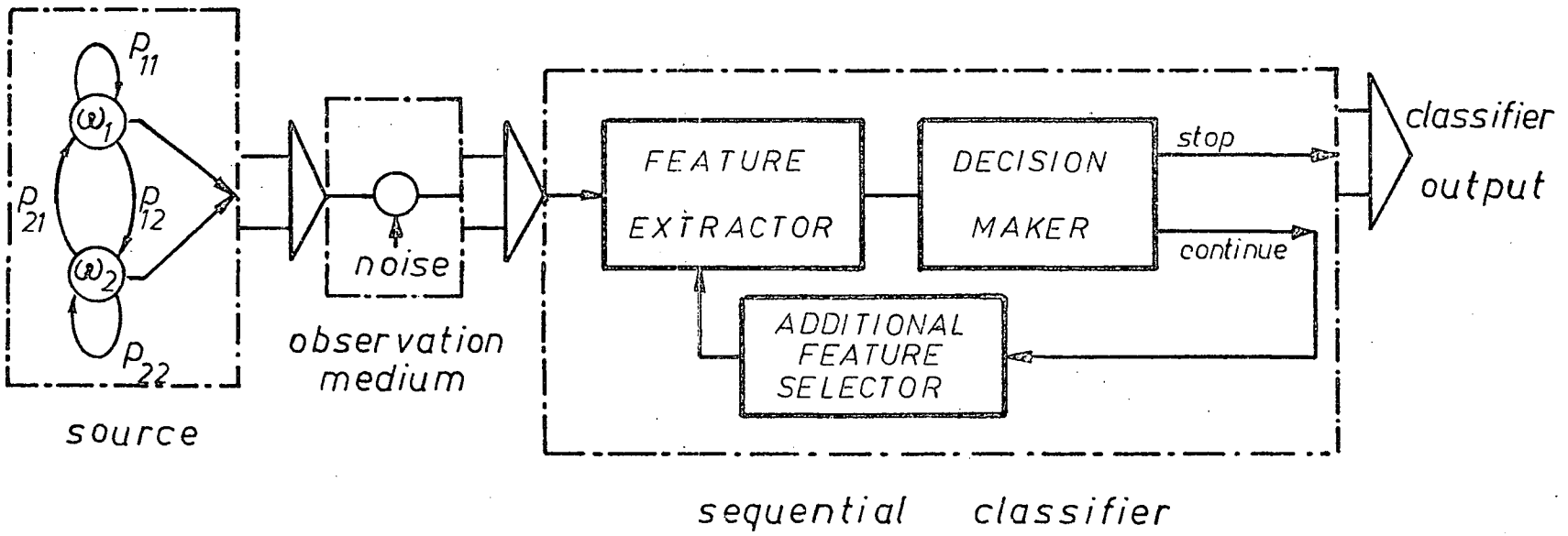

Figure 3.1 Block diagram of a sequential recognition system with two-class Markov dependent source.

$$\gamma_{n_{k-1}} \simeq \begin{cases} T_1, & \text{if at instant } (k-1) \quad \delta^{k-1} = d_1^{k-1} \\ T_2, & \text{if at instant } (k-1) \quad \delta^{k-1} = d_2^{k-1} . \end{cases} \tag{3.19}$$

Thus, the use of the entire past information in this case is equivalent to using the past decision alone. The storage and computation involved, for $t = k$ is thus less than what is involved when $t < k$.

3.2 Nature of the Stopping Boundaries

Without any loss of generality the Markov chain formed by the pattern classes can be assumed to be homogenous. Thus

$$p_{ii} = p_{jj} = p ,$$

$$p_{ij} = p_{ji} = (1-p), \; i, j = 1, 2, \quad i \neq j$$

and consequently $p_1 = p_2 = 0.5$.

Equation (3.17) therefore, reduces to

$$\gamma_{n_k} = \lambda^k \cdot \frac{(1-p) + p \cdot \beta[\bar{X}^{k-1}(t-1); \omega]}{p + (1-p) \cdot \beta[\bar{X}^{k-1}(t-1); \omega]} , \tag{3.20}$$

and can be considered to have two separate parts. At any state n_k, λ^k summarizes the information of n_k feature components from the k th pattern under consideration and the quantity within the braces summarizes the information of the past (t-1) patterns of the sequence. It is only the value of λ^k that changes from one decision state to another; therefore, the decision rule defined in (2.8) can be expressed as follows:

$$\delta^k = \begin{cases} d_2^k, & \text{if } \lambda^k \geq T_2'(k) \\ d_o^k, & \text{if } T_2'(k) > \lambda^k > T_1'(k) \\ d_1^k, & \text{if } \lambda^k \leq T_1'(k) \end{cases}$$

where the new set of stopping boundaries is given by

$$T_2'(k) = T_2 \cdot \frac{p + (1-p)\cdot\beta[\bar{X}^{k-1}(t-1);\ \omega]}{(1-p) + p\cdot\beta[\bar{X}^{k-1}(t-1);\omega]}$$

and

$$T_1'(k) = T_1 \cdot \frac{p + (1-p) \cdot \beta[\bar{X}^{k-1}(t-1);\omega]}{(1-p) + p \cdot \beta[\bar{X}^{k-1}(t-1);\omega]} .$$

The OCSPRT is now similar to the standard SPRT of Wald, where the decision rule is to accept

$$\delta^k = \begin{cases} d_2^k, & \text{if } \lambda^k \geq T_2 \\ d_o^k, & \text{if } T_2 > \lambda^k > T_1 \\ d_1^k, & \text{if } \lambda^k \leq T_1 \end{cases} \tag{3.21}$$

The set of stopping boundaries $T_1'(k)$ and $T_2'(k)$ in the case of the OCSPRT however, besides being dependent on the error probability ϵ_{ij}, are also dependent on the state transition probability p and the past observations. The stopping boundaries are nonlinear in nature except when p = 0 and p = 1. A set of such stopping boundaries, with $\epsilon_{12} = \epsilon_{21} = 0.1$ are presented in Figure 3.2. Regions of increase and decrease in feature requirement for a terminal decision relative to a standard SPRT, for various amount of information from the past observations, are also shown

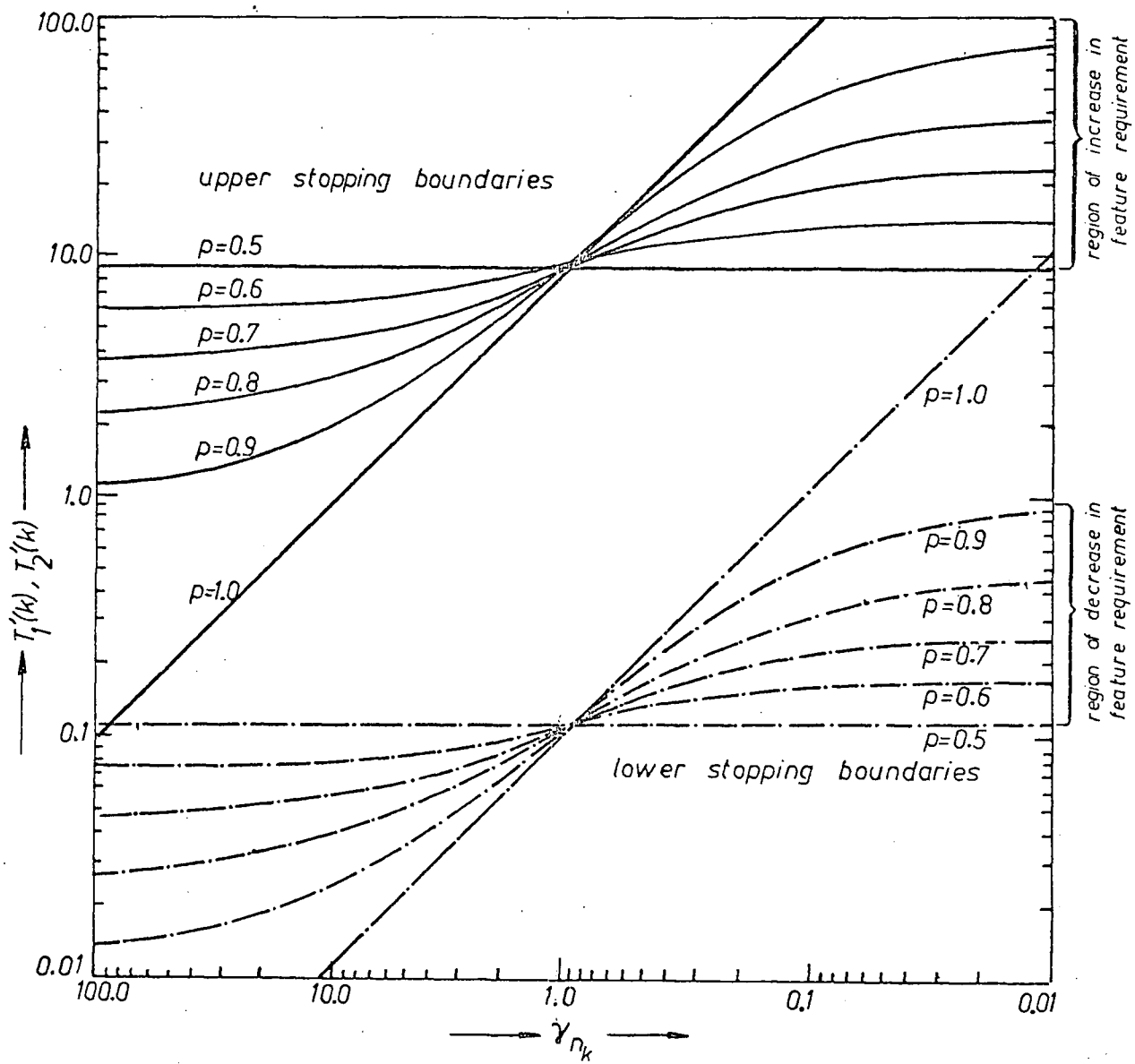

Figure 3.2 The stopping boundaries of OCSPRT, and the regions of saving and increase in the average number of features.

in Figure 3.2. For $p = 0.5$, the stopping boundaries are constant and the OCSPRT reduces to the Wald's [2] standard SPRT, which is only a special case of OCSPRT.

3.3 Expected Number of Features Per Pattern

The expected number of features required per pattern for a terminal decision using OCSPRT is derived in this section. The derivation is based on the assumption that the entire past information is available at every instant of the sequential classification process. Without any loss of generality it may be assumed that $\epsilon_{12} = \epsilon_{21} = \epsilon$, with the result that

$$T_2 = 1/T_1 = T$$
$$= (1-\epsilon)/\epsilon \, .$$

Taking logarithm of (3.18) yields

$$\ln \gamma_{n_k} = \ln \lambda^k + \ln \left[\frac{(1-p) + p\, \gamma_{n_{k-1}}}{p + (1-p)\, \gamma_{n_k}} \right] . \tag{3.22}$$

Taking the expected value of (3.22) yields

$$E[\ln \gamma_{n_k}] = E[\ln \lambda^k] + E[\ln \{\frac{(1-p) + p \cdot \gamma_{n_{k-1}}}{p + (1-p)\, \gamma_{n_{k-1}}}\}], \tag{3.23}$$

where $E[\cdot]$ is used to denote the expected value. For statistically independent feature components and for a large value of D, it is possible to show [2], [46] that

$$E[\ln \gamma_{n_k}] = \begin{cases} e_1(\gamma), & \text{when } X^k \sim \omega_1 \\ e_2(\gamma), & \text{when } X^k \sim \omega_2 \end{cases}$$

where $e_i(\gamma) = (-1)^i \cdot (1-2\epsilon)\ \ell n(\frac{1-\epsilon}{\epsilon})$; $i = 1, 2$, and

$$E[\ell n\ \lambda^k] = \begin{cases} e_2(n) \cdot e_2(f), & \text{if } X^k \sim \omega_2 \\ e_1(n) \cdot e_1(f), & \text{if } X^k \sim \omega_1 \end{cases}$$

where $e_i(n)$, $i = 1, 2$, denotes the expected number of features required per pattern for a terminal decision when $X^k \sim \omega_i$

and

$$e_i(f) = E[\ell n \frac{P(x_j^k|\omega_2)}{P(x_j^k|\omega_1)}], \quad i = 1, 2;\ j = 1, 2, \ldots, D.$$

The derivation of the above results appear in Appendix B.

As noted earlier, use of the entire past information in the case of OCSPRT, is equivalent to using the past decision. Thus, at any instant k (neglecting any excess over the stopping boundaries*)

$$\gamma_{n_{k-1}} = \begin{cases} T, & \text{if at instant } (k-1)\ \ \delta^{k-1} = d_2^{k-1} \\ 1/T, & \text{if at instant } (k-1)\ \ \delta^{k-1} = d_1^{k-1} \end{cases}$$

However,

$$\gamma_{n_{k-1}} = \begin{cases} T, & \text{with probability} \begin{cases} (1-\epsilon), & \text{if } X^{k-1} \sim \omega_2 \\ \epsilon, & \text{if } X^{k-1} \sim \omega_1 \end{cases} \\ 1/T, & \text{with probability} \begin{cases} \epsilon, & \text{if } X^{k-1} \sim \omega_2 \\ (1-\epsilon), & \text{if } X^{k-1} \sim \omega_1 \end{cases} \end{cases}$$

---

*This is assumed for the purpose of simplicity of derivation. If the expected number of features required per pattern is large, i.e. γ_{n_k} takes a long time to reach one of the stopping boundaries, then the increment $\Delta\gamma_{n_k}$ for each additional feature would be small and the assumption is valid.

Let

$$z = \ln \left\{ \frac{(1-p) + p\,\gamma_{n_{k-1}}}{p + (1-p)\,\gamma_{n_{k-1}}} \right\},$$

such that when $X^{k-1} \sim \omega_2$, one can write

$$e_2(z) = E\left[\ln \left\{ \frac{(1-p) + p\,\gamma_{n_{k-1}}}{p + (1-p)\,\gamma_{n_{k-1}}} \right\}\right]$$

$$= (1-\epsilon)\ \ln \left\{\frac{(1-p) + pT}{p + (1-p)T}\right\} + \epsilon\ \ln\left\{\frac{(1-P)T + p}{pT + (1-p)}\right\}$$

$$= (2\epsilon-1)\ \ln \left\{\frac{(1-p)T + p}{pT + (1-p)}\right\}.$$

Similarly, when $X^{k-1} \sim \omega_1$

$$e_1(z) = (1-2\epsilon)\ \ln \left\{\frac{(1-p)T + p}{pT + (1-p)}\right\}.$$

Let:

$e_k(n_{i,j,\ldots,\ell})$ be the expected number of features required per pattern when the sequence of input patterns is such that, $X^k \sim \omega_i$, $X^{k-1} \sim \omega_j$, ..., $X^1 \sim \omega_\ell$

and

$p^{(i)}_{j,\ldots,\ell}$ denote the probability $P(\omega_i^k | \omega_j^{k-1}, \ldots, \omega_\ell^1)$ of a sequence of patterns.

Equation (3.23) can now be rewritten in the following form

$$e_i(\gamma) = e_k(n_{k,j,\ldots,\ell}) \cdot e_i(f) + e_j(z)$$

or

$$e_i(n_{i,j,\ldots,\ell}) = [e_i(\gamma) - e_j(z)]/e_i(f), \qquad (3.24)$$

$$i,j,\ldots,\ell, = 1,\ 2$$

However,

$$e_i(n) = \sum_{j,\ldots,\ell} e_k(n_{i,j,\ldots,\ell}) \cdot p_{j,\ldots,\ell}^{(i)} \tag{3.25}$$

Combining (3.24) and (3.25) yields

$$\begin{aligned} e_i(n) &= \sum_{j,\ldots,\ell} \frac{e_i(\gamma) - e_j(z)}{e_i(f)} \cdot p_{j,\ldots,\ell}^{(i)} \\ &= \sum_{j=1}^{2} \left\{\frac{e_i(\gamma) - e_j(z)}{e_i(f)}\right\} \cdot p_{ji} \end{aligned} \tag{3.26}$$

where p_{ji} is defined in (3.16).

On substitution and simplification, (3.26) finally reduces to

$$e_i(n) = \{e_i(\gamma) - (-1)^i(2\tau - 1)\ln[\tau/(1 - \tau)]\}/e_i(f), \quad i = 1, 2 \tag{3.27}$$

where, $\tau = \epsilon + p - 2p\epsilon$.

It is intuitively satisfying that the expected number of features required per pattern for a terminal decision is a function of the probability distribution of the feature components, the error probability ϵ, and the state transition probability p. For p = 0.5, that is, when the source is memoryless, (3.27) reduces to the expression for the average number of features for standard SPRT and is given by

$$e_i'(n) = \frac{e_i(\gamma)}{e_i(f)}, \quad i = 1, 2, \tag{3.28}$$

where $e_i'(n)$ denotes the expected number of features required per pattern for standard SPRT, when $X^k \sim \omega_i$.

3.4 Comparison of the OCSPRT and the Standard SPRT

It is now possible to compare the performance of OCSPRT with that of an equally reliable standard SPRT. Combining (3.27) and (3.28) yields

$$\eta_i = (-1)^i \cdot (1 - 2\tau)\ \ell n[(1 - \tau)/\tau]/e_i(\gamma) \tag{3.29}$$

where $\eta_i = \{e_i'(n) - e_i(n)\}/e_i'(n)$.

Based on (3.29) we can state the following theorem.

THEOREM 3.2 For any $1 \geqslant p \geqslant 0$ and $1 > \epsilon > 0$, the OCSPRT using the entire past information requires no more features per pattern than the equally reliable standard SPRT.

The following lemmas need to be proved before proving the Theorem 3.2.

Lemma 3.1: The real valued function

$$f(x, y) = x + y - 2xy \tag{3.30}$$

for $1 \geqslant x \geqslant 0$ and $1 \geqslant y \geqslant 0$ is such that $1 \geqslant f(x,y) \geqslant 0$.

Proof: The proof of the lemma consists of two separate parts:

1. Lower Bound: the inequality of arithmetic and geometric means for any nonnegative numbers a and b is given by [47]

$$a + b \geqslant 2(ab)^{1/2}$$

on substitution in (3.30) we obtain

$$f(x,y) \geqslant 2[(xy)^{1/2} - xy].$$

However, for any $1 \geqslant x \geqslant 0$ and $1 \geqslant y \geqslant 0$

$$(xy)^{1/2} \geqslant xy.$$

Thus, $f(x,y) \geqslant 0$, where equality holds only when $x = y$.

2. Upper Bound: let the new variable $x' = (x - 1/2)$. Thus, for $1 \geqslant x \geqslant 0$, x' is such that

$$1/2 \geqslant x' \geqslant -1/2$$

On substitution in (2.30) we obtain

$$f(x', y) = x'(1 - 2y) + 1/2,$$

since for any real numbers a and b, $|a| + |b| \geqslant |a + b|$,

$$|f(x', y)| \leqslant |x'(1 - 2y)| + 1/2. \tag{3.31}$$

It is also possible to show that

$$\text{Max } |a \cdot b| \leqslant \text{Max } |a| \cdot \text{Max } |b|.$$

Therefore, (3.31) can be expressed as follows:

$\text{Max } |f(x',y)| \leqslant \text{Max } |x'| \cdot \text{Max } |1 - 2y| + 1/2.$

Since, $\text{Max } |x'| = 1/2$ and $\text{Max } |1 - 2y| = 1$

we obtain

$$\text{Max } |f(x', y)| \leqslant 1$$

or $\text{Max } |f(x, y)| \leqslant 1$ Q.E.D.

<u>Lemma 3.2</u>: For all values of $1 \geqslant x \geqslant 0$, the real valued function

$$f(x) = (1 - 2x) \ \ell n(\frac{1-x}{x}) \tag{3.32}$$

is always positive.

<u>Proof</u>: For any $1 \geqslant x \geqslant 0.5$, we have

$$0 \geqslant (1 - 2x) \geqslant -1,$$

and

$$1 \geqslant (\frac{1-x}{x}) \geqslant 0.$$

Equality sign holds only when $x = 1$ or $x = 0.5$.

Therefore, $f(x) \geqslant 0$.

For any $0.5 \geqslant x \geqslant 0$,

We have $1 \geqslant (1 - 2x) \geqslant 0$

and

$$\infty \geqslant \left(\frac{1-x}{x}\right) \geqslant 1,$$

again the equality sign holds only when $x = 0.5$ or $x = 0$

Therefore, $f(x) \geqslant 0$ Q.E.D.

In order to prove Theorem 3.2, one has to show that $\eta_i \geqslant 0$, $i = 1, 2$.

<u>Proof</u>: We have from (3.29)

$$\eta_2 = \frac{(1-2\tau)}{e_2(\gamma)} \ln \left(\frac{1-\tau}{\tau}\right)$$

where

$$e_2(\gamma) = (1 - 2\epsilon) \ln \left(\frac{1-\epsilon}{\epsilon}\right)$$

and

$$\tau = \epsilon + p - 2p\epsilon.$$

Since $1 \geqslant \epsilon \geqslant 0$, we have from Lemma 3.2, $e_2(\gamma) \geqslant 0$.

Also, the state transition probability $1 \geqslant p \geqslant 0$, therefore, from Lemma 3.1 we know τ is bounded by 1 and 0. It thus directly follows from Lemma 3.2 that $\eta_2 \geqslant 0$ for all values of $1 \geqslant p \geqslant 0$ and $1 \geqslant \epsilon \geqslant 0$.

It can similarly be shown that $\eta_1 \geqslant 0$.

Q.E.D.

In Figure 3.3 η_i is plotted for various values of p and ϵ. Except when $p = 0.5$, in which case the OSPRT reduces to the standard SPRT, the saving in the average number of features due to OCSPRT is always positive and reaches 100 percent when the channel is completely determinisitic. The higher the desired reliability of the recognition

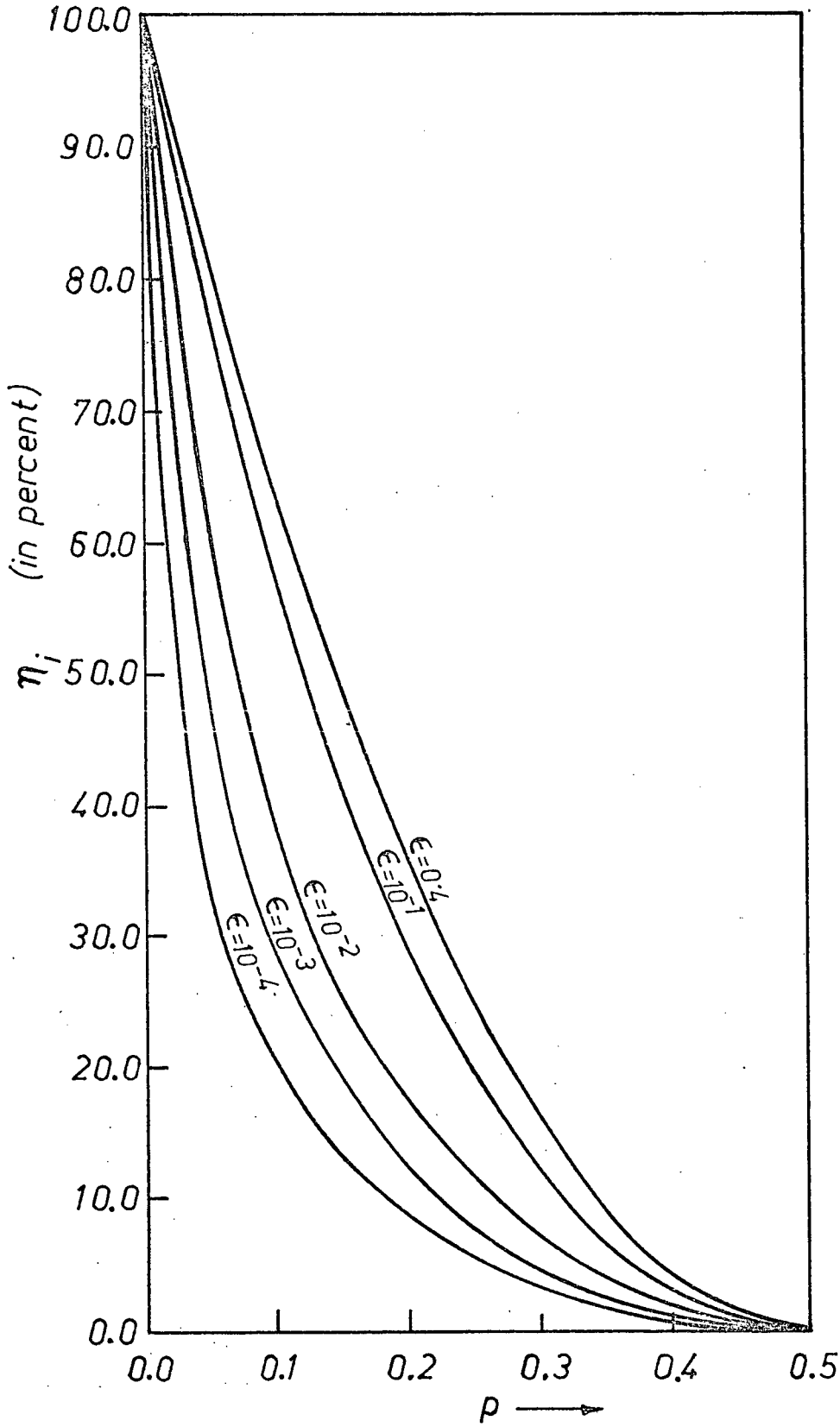

Figure 3.3 Saving in the average number of features per pattern obtained using OCSPRT relative to SPRT as a function of error probability ϵ and state transition probability p.

system, the lower is the amount of saving, since in such a situation both SPRT and OCSPRT require a relatively large number of features per pattern for terminal decisions.

3.5 Performance Evaluation by Simulation Experiments

In order to assess the performance of the OCSPRT, a recognition problem similar to the one discussed in section 3.1 was simulated on an IBM system /360 Model 67 digital computer. A sequence of 5000 patterns, with approximately* 2500 of them being a member of each one of the two pattern classes, was generated. The sequence was generated in such a way that at any instant k, k = 1, 2, ..., the state transition probability $\{p(\omega_i^{k+1}|\omega_i^k),\ i = 1, 2\} = p \pm \Delta p,\ 0.002 \geq \Delta p \geq 0$, and consequently $p(\omega_1) \simeq p(\omega_2)$. Each pattern sample of the sequence was represented by a feature vector of 40 statistically independent feature components generated in the computer using a random number generator. The feature components were generated in such a fashion that $P(x_i|\omega_j)$, i = 1, 2, ..., 40; j = 1, 2, was a univariate normal distribution with mean m_j and variance σ^2. Thus

$$P(x_i|\omega_j) = (2\pi)^{-1/2}\ \sigma^{-1}.\ \exp[-(x_i - m_j)^2/2\sigma^2]\ .$$

The dimensionality of the generated feature vectors was found to be large enough to be able to avoid the necessity of any form of truncation scheme in the sequential decision processes for the range of threshold values used in the experiments.

Besides the OCSPRT, three other parametric classifiers were

*There were 2504 pattern samples under pattern class ω_2 and 2496 pattern samples under pattern class ω_1.

simulated to classify the generated sequence of patterns. The classifiers were as follows:

1. Simple Nonsequential Bayes Classifier: with statistically independent feature components. The pattern classes were also assumed to be statistically independent. Thus, at any instant k the decision rule was

$$\delta^k = \begin{cases} d_2^k, & \text{if } P(X^k|\omega_2^k) \geq P(X^k|\omega_1^k) \\ d_1^k, & \text{otherwise} \end{cases} \tag{3.33}$$

for normally distributed feature components the decision rule (3.33) was equivalent to

$$\delta^k = \begin{cases} d_2^k, \text{ if } & \sum_{i=1}^{D} x_i^k \geq 0.5\ D.\mu \\ d_1^k, & \text{otherwise} \end{cases}$$

where

$$\mu = (m_2 - m_1), \qquad m_2 > m_1$$

and D is the dimensionality of the feature vector.

2. Compound Nonsequential Bayes Classifier: with statistically independent feature components. The pattern classes were assumed to form a first order stationary homogenous Markov chain with p = 0.85. Thus

$$\delta^k = \begin{cases} d_2^k, & \text{if } P(\bar{X}^k|\omega_2) \geq P(\bar{X}^k|\omega_1) \\ d_1^k, & \text{otherwise} \end{cases}$$

where

$$P(\bar{X}^k|\omega_i) = P(X^k|\omega_i) \sum_{j=1}^{2} P(\bar{X}^{k-1}|\omega_j^{k-1}) P(\omega_i^k|\omega_j^{k-1}),$$

and was obtained iteratively at every instant of the decision process.

3. Standard SPRT: with constant stopping boundaries. The error probabilities ϵ_{12} and ϵ_{21} were assumed to be equal such that $T_2 = 1/T_1 = T$. The decision rule is presented in (3.21) and for normally distributed features reduces to

$$\delta^k = \begin{cases} d_2^k, \text{ if } \sum_{i=1}^{n_k} x_i^k \geq \frac{\sigma^2}{\mu} \ln T + 0.5\, n_k \mu \\ d_o^k, \text{ if } \frac{\sigma^2}{\mu} \ln T + 0.5\, n_k \mu > \sum_{i=1}^{n_k} x_i^k > -\frac{\sigma^2}{\mu} \ln T + 0.5 n_k \mu \\ d_1^k, \text{ if } \sum_{i=1}^{n_k} x_i^k \leq -\frac{\sigma^2}{\mu} \ln T + 0.5 n_k \mu \end{cases}$$

where $T = (\frac{1-\epsilon}{\epsilon})$.

In case of OCSPRT the pattern classes were assumed to form a first order stationary, homogenous Markov chain with $p = 0.85$. Also the error probabilities ϵ_{12} and ϵ_{21} were assumed equal.

In the first set of experiments $t = 2$ was considered for the OCSPRT case, thus information was assumed to be available from the immediate past pattern only. Figures 3.4 and 3.5 show the error probability of the four different classification schemes as a function of the average number of features required per pattern. Two sets of curves are presented in order to compare the classification schemes at various ranges of error probability. It is observed that the OCSPRT for a specified recognition accuracy requires a significantly fewer number of features (on the average) per pattern than the other three classification schemes. In comparison with an equally reliable standard SPRT, savings in the

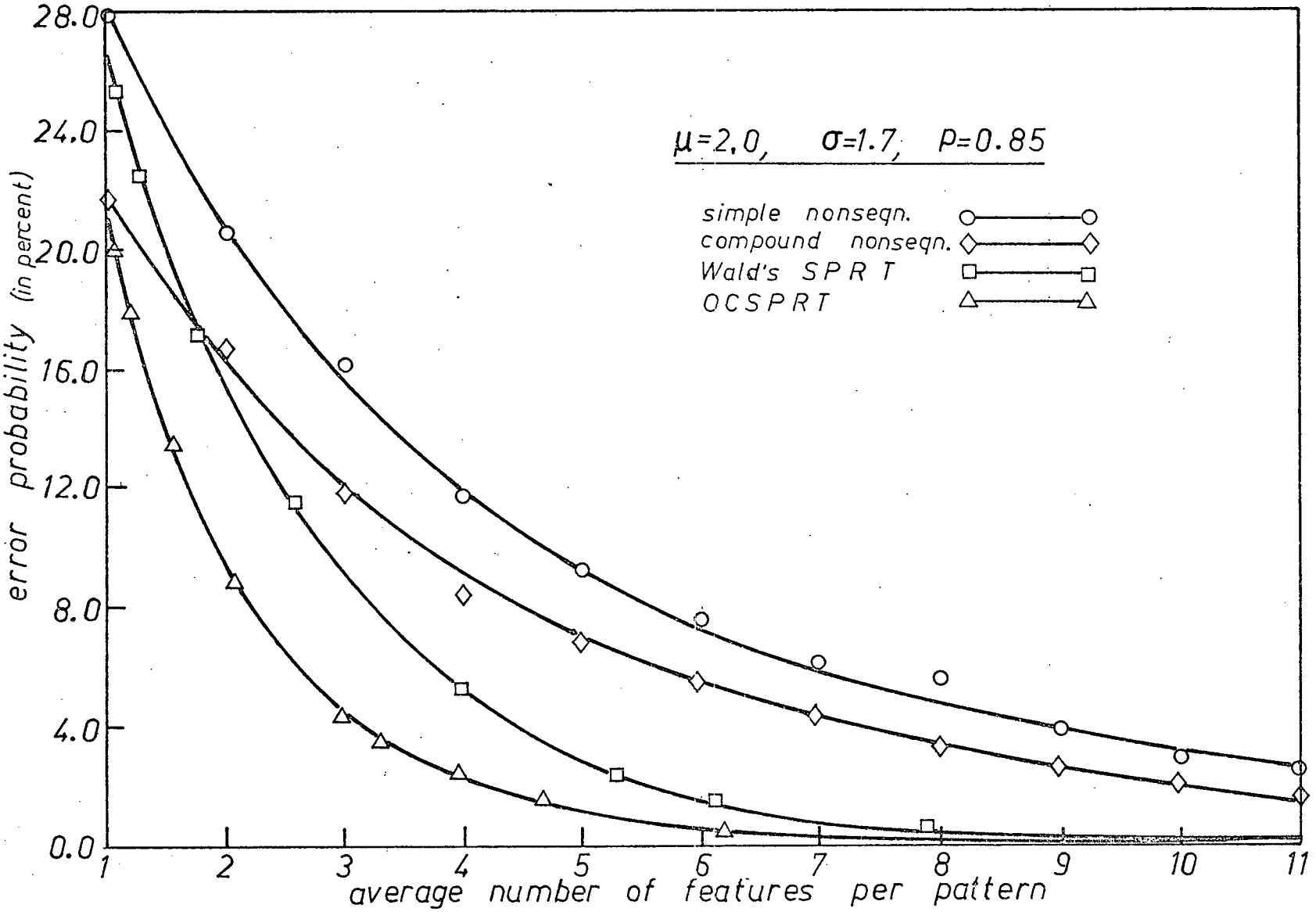

Figure 3.4 Comparison of OCSPRT, SPRT, optimal simple nonsequential, and optimal compound nonsequential schemes. $\sigma = 1.7$.

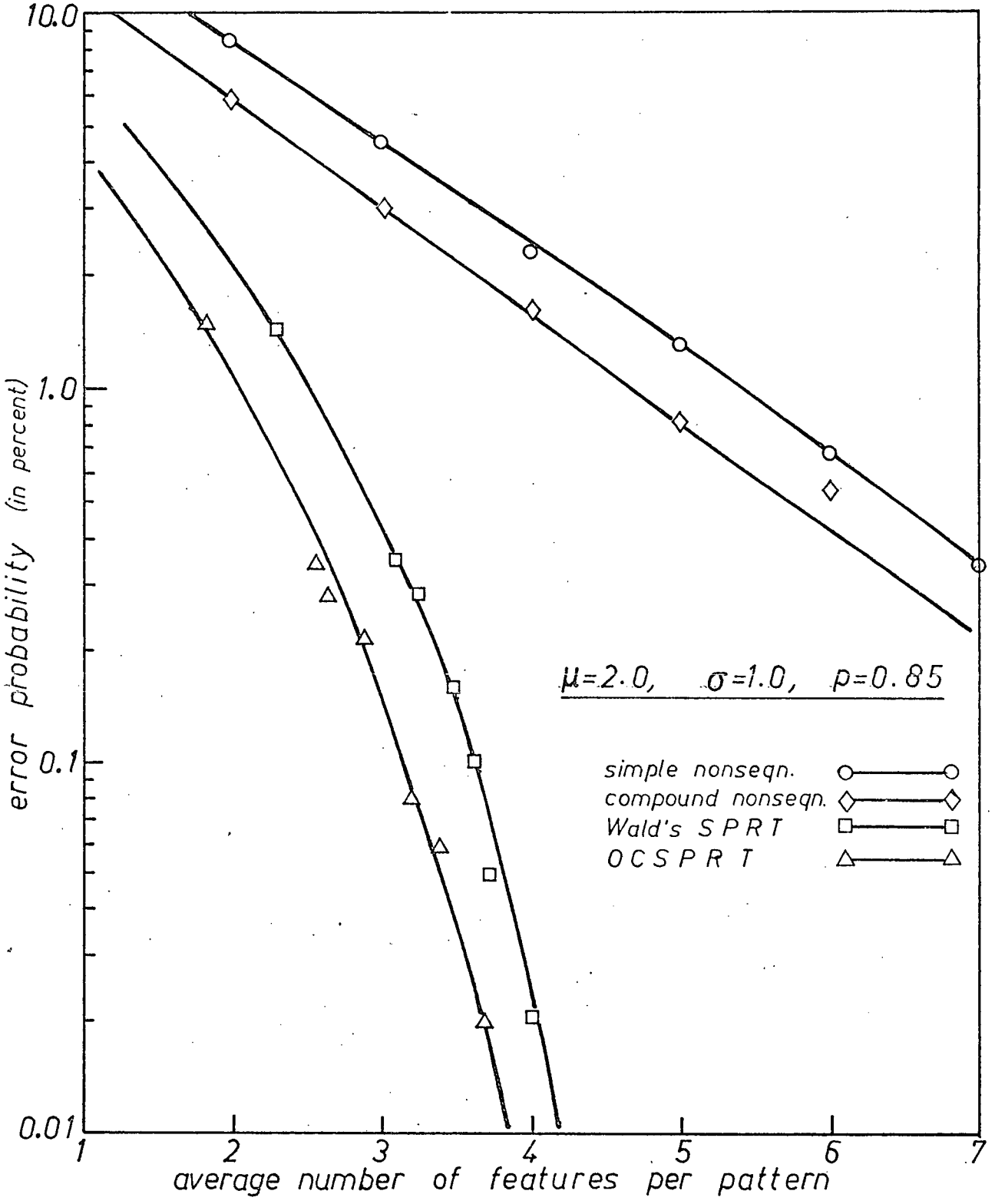

Figure 3.5 Comparison of OCSRPT, SPRT, optimal simple nonsequential, and optimal compound nonsequential decision schemes. $\sigma = 1.0$.

average number of features as high as 31 percent are observed using the OCSPRT. The OCSPRT requires approximately 60 percent fewer features per pattern than the equally reliable compound nonsequential classification scheme.

A second set of experiments using the same data set was conducted in order to observe the effect of varying t, the length of the set of feature vectors used in the OCSPRT. Four different values of t, including t = k (entire past) were considered. The results are presented in Figure 3.6. As shown in the figure, the performance of the OCSPRT improves with the increase in value of t. However, the recognition accuracy increases so little beyond t = 2 that this slight improvement does not justify the introduction of additional storage and computational complexity. An OCSPRT with a relatively small value of t would be satisfactory.

We have pointed out before that using the information from the entire past (t = k) in the OCSPRT is equivalent to using the past decision. For a recognition system with overlapping class clusters in the feature space, the method of using the decision alone tends to propagate error from one instant to the next. Thus, one would expect the OCSPRT with t = k to perform worse than the OCSPRT with t < k (using past information) and in fact it does perform worse as we have shown in Figure 3.6. For a certain range of average numbers of features per pattern, the OCSPRT with t = k requires more features than the equally reliable OCSPRT with t = 2,3 and 4 and vice versa. However, for highly reliable systems, that is, when the class clusters are quite disjoint in the feature space, the use of the previous decision would be as efficient as using the past information as far as the expected cost of terminal decisions is concerned [34].

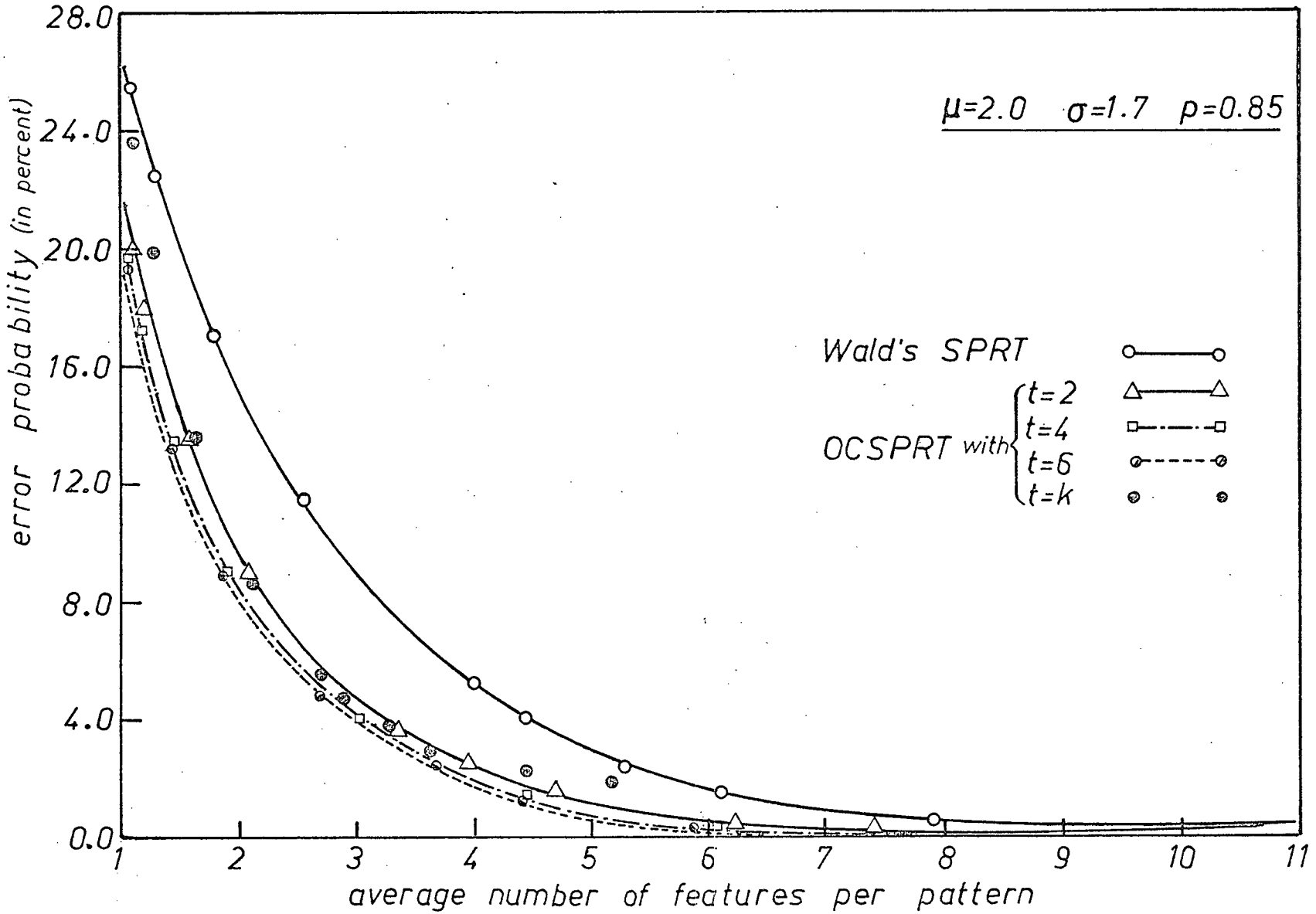

Figure 3.6 Effect of t, the length of the sequence of feature vectors available at any instant, on the performance of OCSPRT.

CHAPTER IV

DATA BASE, PREPROCESSING, FEATURE EXTRACTION, AND SYSTEM EVALUATION

4.1 Description of the Handprinted Character Set

In order to assess the performances of the various decision schemes considered in the rest of the thesis, a particular pattern recognition problem was simulated on the IBM systems /360 Model 67 digital computer. The pattern classes considered were handprinted, uppercase English alphabetic characters obtained from Munson's multi-author data file [51]. The original data file was prepared at Stanford Research Institute, and contains the entire 46 Fortran characters. The characters were collected from ordinary Fortran coding forms, prepared by 49 different authors, with each author contributing 3 sets of alphabet. Thus there were 147 samples of each Fortran character. Although the authors were instructed to slash the letter Z and to put crossbars on the letter I, we noticed during visual examination of the data set that several authors failed to follow these instructions. Some lowercase characters were also encountered in the data set, however they were not removed from the data set. Each character was available in the form of a 24 × 24 matrix of 0 and 1 corresponding to white and black parts of the character image.

There are two main reasons for using Munson's data set for the purpose of simulating our recognition experiments. Firstly, this was the only well prepared data set of adequate size that was readily

available during the period when the design of our experiments was in progress. Secondly, the I.E.E.E. Technical Committee on pattern recognition has organized a subcommittee on Reference Data sets [51] in the hope of convincing researchers to use standardized data sets in their experiments; Munson's multiauthor data file is one of these reference data sets and recognition results obtained by others have already been published using this data set [51]-[53]. Thus one can compare our experimental results with those of others.

4.2 Preprocessing of the Characters

Prior to the extraction of the features, each character was subjected to a simple size-normalization preprocessing operation. The purpose of the preprocessing was to reduce the sensitivity of the feature extraction scheme (described in Section 4.3) to the variation in the size of the character samples. The manner in which the characters were size-normalized is described in the following paragraphs.

The width W and height H of each character was first defined by locating the rectangle defined by the two rows and columns closest to the edges of the character array and containing at least two black points. The ratio H/W was then calculated. If $H/W \leqslant 3$ the original character was size normalized to 20×20 points by adding to, or deleting from, rows and/or columns at regular spatial intervals in the original 24×24 array. Let NHA and NHD equal the number of rows to be added and deleted, respectively, to make the final height equal 20. Then NHD = H-20 and NHA = 20-H. Let $\langle x \rangle$ denote the largest integer not exceeding x. Let i be any positive integer. If $H > 20$ then the NHD rows numbered $i \cdot \langle (H + NHD)/(NHD+1) \rangle$ from the top were deleted from the

original 24 rows. If $H < 20$ then $NHD = H/(NHA+1)$ was obtained and then set equal to the nearest integer to yield NDI. A new row was inserted immediately below row $i \cdot (NDI)$ in the original 24×24 array. Each inserted row was made identical to the row above it in the new array. The first 20 rows were retained as those defining the size-normalized character. A similar procedure was used to insert or delete columns.

If $H/W > 3$ then the original width W was not altered. Instead, the character was centered horizontally on a grid 20 columns wide. The height was normalized to 20 rows, as explained in the above paragraph. This procedure prevented characters such as "I", which may be printed either with or without serifs, from being converted into a virtually all - black 20×20 array when serifs were absent. Figure 4.1 shows two characters in their original and preprocessed forms.

4.3 Feature Extraction Scheme

A good set of features results in the well discrimination of the pattern classes and is, at the same time, relatively insensitive to normal variations among the patterns of the same class. The question of simplicity of the feature extraction algorithm and the dimensionality of the generated feature vector is of considerable importance. Various feature extraction algorithms, especially designed for character recognition, based on designer's ingenuity and intuition have been reported from time to time [51]-[55]. The few outstanding of which are n-tuple, contour tracing, topological and geometrical descriptors, characteristic loci, and mask matching. Most of these methods result in a feature vector of relatively high dimensionality, require extremely complicated operations, or suffer from both of the above disadvantages.

```
*************************
*                       *
*                       *
*                       *
*                       *
*        MMMMM   M M    *
*        MMMMMMMMMMMMM  *
*        MMMMMMMMMMMMM  *
*        MMMMMM  MMM    *
*       MMMM            *
*       MMMM            *
*       MMMMMMMMMMM     *
*      MMMMMMMMMMMMM    *
*      MMMMMMMMMMMMM    *
*     MMMMMM  M         *
*    MMMMM              *
*    MMMMM        MM    *
*    MMMMMMMMMMMMMM     *
*   MMMMMMMMMMMMMM      *
*   MMMMMMMMMMMMMM      *
*    MMMMMMMMMM         *
*                       *
*                       *
*                       *
*                       *
*************************
```

original

```
*********************
*     MMMMMM   M M  *
*     MMMMMMMMMMMMMM*
*     MMMMMMMMMMMMMM*
*     MMMMMMMMMMMMMM*
*     MMMMMMM  MMM  *
*    MMMMM          *
*    MMMMM          *
*    MMMMM          *
*    MMMMMMMMMMMMM  *
*   MMMMMMMMMMMMMMM *
*   MMMMMMMMMMMMMMM *
*   MMMMMMMMMMMMMMM *
*  MMMMMMM  M       *
* MMMMMM            *
* MMMMMM        MM  *
* MMMMMM        MM  *
* MMMMMMMMMMMMMMMM  *
*MMMMMMMMMMMMMMMM   *
*MMMMMMMMMMMMMMMM   *
*MMMMMMMMMMMMMMMM   *
*********************
```

size-normalized

```
***************************
*            MMMM         *
*          MMMMMMM        *
*       MMMMMMMMMM        *
*       MMMM   MMMMM      *
*      MMMM       MMMM    *
*      MMMM          MMM  *
*    MMMM            MMMM *
*    MMMMMMMM        MMMM *
*    MMM  MMMM       MMMM *
*    MMM  MMMMM       MMM *
*   MMMM   MMMM      MMMM *
*    MMMM  MMMM    MMMM   *
*    MMMMMMMMMMMMMM      *
*      MMMMMMMMMMMM      *
*       MMMMMMMMMM       *
*           MMMMM        *
*            MMMM        *
*            MMMM        *
*            MMMM        *
*             MMMM       *
*             MMMM       *
*             MMMM       *
*              MMM       *
*                        *
***************************
```

original

```
*********************
*          MMMMM    *
*       MMMMMMMMM   *
*   MMMMMMMMMMMMMM  *
*   MMMMMM   MMMMMMM*
* MMMMM          MMMMM*
*MMMMM              MMMM*
*MMMMMMMMMM         MMMM*
*MMMM  MMMMM        MMMM*
*MMMM  MMMMMM        MMM*
*MMMM   MMMMMM      MMMM*
*MMMMMMMMMMMMMMMMMMMM   *
* MMMMMMMMMMMMMMMMM    *
*  MMMMMMMMMMMMMMM     *
*         MMMMMM       *
*          MMMMM       *
*          MMMMM       *
*           MMMMMM     *
*           MMMMMM     *
*           MMMMMM     *
*              MMMM    *
*********************
```

size-normalized

Figure 4.1 The original and the size-normalized character samples

These schemes are unrealizable in many practical situations because of the prohibitive system complexity and because of the large amounts of training data needed for proper estimation of the required parameters. In this section of the thesis a feature extraction method is proposed which, in spite of its simplicity and relatively low dimensionality, has been found to yield performance comparable with that of other character recognition schemes.

Let $G(i, j)$, $i = 1, 2, \ldots, h$; $j = 1, 2, \ldots, w$ be a point on a $(h \times w)$ dimensional pattern grid on which each pattern image is placed as illustrated in Figure 4.2. Each point on the grid may be allowed to assume one of q values depending on the intensity of the image at that point. A mask is defined to be a grid of size $(h' \times w')$, $h' \leq h, w' \leq w$, such that through the mask only a portion of the pattern grid is observable. All possible configurations of the mask within the pattern grid are assumed to be valid. The points of the mask are so labelled that the lower leftmost point, called the origin, is subscripted (0, 0). Let a convenient weighting procedure associate weight $g(i, j)$, $i = 0, 1, \ldots, h' - 1$; $j = 0, 1, \ldots, w' - 1$, with each point (i, j) of the mask. A feature component is generated by superimposing the mask on the pattern grid and scanning only that portion of the pattern grid which is observable through the mask. Different feature components are generated for different configurations of the mask on the pattern grid. If one of the configurations is such that the origin of the mask is given by co-ordinate (ℓ, k) on the pattern grid, then the value of the corresponding feature component is expressed as follows:

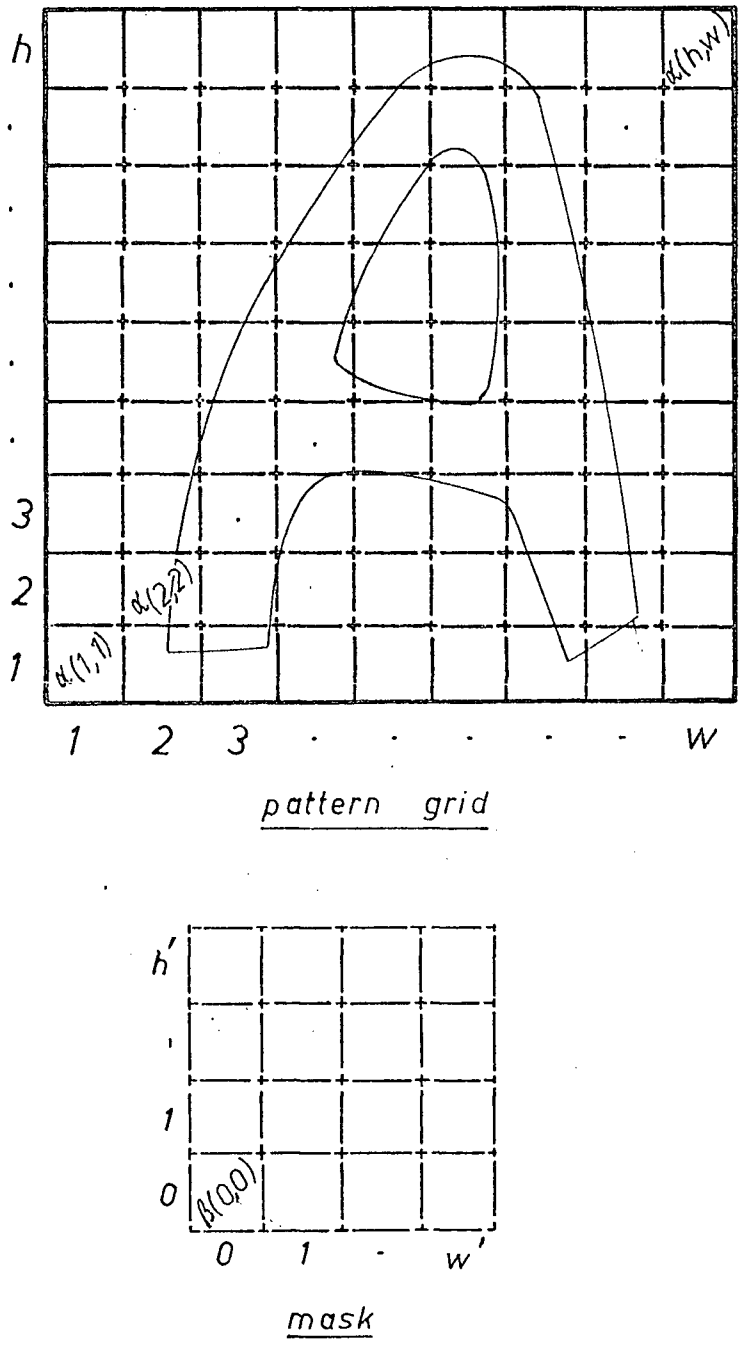

Figure 4.2 The (h × w) pattern matrix and the (h' × w') mask

$$x = \{ \sum_{i=0}^{h'-1} \sum_{j=0}^{w'-1} g(i, j) \cdot \{ G[(\ell + i), (k + j)] \}.$$

Even though the dimensionality of the feature vector is dependent on the size of the pattern grid and the mask, as well as on the allowable overlap between neighbouring mask positions, the actual number of components is governed by the number of allowable mask configurations. If N_{max} represents the maximum number of features available, then one can show that

$$N_{max} = \begin{cases} (h - h' + 1)(w - w' + 1); & \text{if the configurations of the mask are not mutually exclusive,} \\ \langle h/h' \rangle \cdot \langle w/w' \rangle; & \text{otherwise} \end{cases}$$

where $\langle x \rangle$ again denotes the largest integer not exceeding the value of x.

For our experiments the mask in its simplest form was used. The mask used was of size 4 × 4 and the points on it were all equally weighted, such that $g(i, j) = 1$, $i, j = 0, 1, \ldots, 3$. The pattern grid of size 20 × 20 had only two levels of quantization, 0 and 1 depending on whether the particular point on the grid was white or black. All the mutually exclusive configurations of the mask on the pattern grid were allowed, thus 25 different feature components, each one taking on 17 discrete values, were obtained.

The feature extraction scheme described above is fairly general and may be useful for various other pattern recognition problems. For character recognition the scheme is found to have the following desirable properties:

1. The feature vector is obtainable from the pattern grid by a repetitive procedure involving only simple logical operations;

thus the method is very simple compared to most of the schemes used for handprinted characters.

2. The feature vector's dimensionality is significantly lower than that of many other schemes and the feature components themselves assume only a relatively small number of values.
3. Discontinuities in the character matrix including salt and pepper noise can be tolerated with this feature extraction scheme.
4. The scheme is especially suited for studying sequential recognition processes where an additional feature is observed only if further information regarding the unknown pattern is necessary. This scheme allows the parts of the character to be examined sequentially so that a new region of the character is examined only if those previously examined do not provide enough information for a final decision.

The proposed scheme, like many others, suffers from the drawback that it is not invariant under size, rotation, and thickness transformations on the patterns. One may circumvent the difficulties through simple preprocessing operations such as size-normalization and shearing of the patterns, and still maintain the important attributes of the feature extraction method. The potential sensitivity of the scheme to large deviations in size among the characters of any given class motivated us to use the size-normalization preprocessing.

The 25 feature components were serially numbered, starting from the top leftmost configuration of the mask on the pattern grid and moving to the right. Figure 8.3 shows the serial number of various

feature components. This ordering of the feature components is defined as the natural ordering of features.

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	25

Figure 4.3 The natural ordering of the 25 feature components

4.4 Estimation of Probabilities and System Evaluation Procedure

The important question that invariably arises in the statistical approach to pattern recognition is: what is the best way to use a set of finite number of pattern samples to design a classification scheme and evaluate its performance? The reliability of the performance estimate is influenced by the size of both the testing and the training sets of samples. "Sufficientcy" of the size of the sets for reliable estimation is a function of the number of parameters involved in the recognition system. It is usually an arduous job to obtain a "sufficient" size sample due to the limitations imposed by cost, difficulty, and computation. The problem essentially is to make "optimum" use of the available fixed data set to maximize the accuracy of the estimate.

The problem has received considerable attention in the last several years [44], [48]-[50], [56], [57]. In most of the earlier work in pattern classification, the same set of data was used both for the purpose of testing and training the classifiers. It is now well known that this method yields an over-optimistic estimate of the performance. Results thus obtained are, therefore, of limited value. In 1962 Highleyman [48] suggested partitioning of the data set into two disjoint sets and using one as the training set and the other as the testing set. This method is referred to as the H method (Holdout method) in the pattern recognition literature. Highleyman presented analytical results concerning the optimum partitioning to minimize the variance of the estimated error rate. Kanal and Chanrasekaron [50] however, showed that Highleyman's partitioning is valid only when the total sample size is large. His analysis breaks down in the small sample case. When the data set is small compared to the number of parameters that has to be estimated, the H method yields an overly pessimistic estimate of the system performance.

In 1968 Lachenbruch and Mickey [49] made an embirical comparison of various methods of estimating performances of recognition systems. They observed that one of their methods, referred to as theU method, yields the best estimate of performance when the size of the data set is small relative to the number of parameters involved, and when normality of the underlying probability distributions is questionable. For a data set with N samples, the U method consists of conducting N recognition experiments by successively testing each one of the N samples when the parameters are estimated with the remaining (N-1) samples. The d r a w - back with this method is that, unless the data set is relatively small,

a very large amount of computation is required.

In this thesis, a method earlier developed by Toussaint and Donaldson [58], [59], combining the important aspects of both the U and the H methods is used. The R (rotation) method, as we shall call it, would provide a better estimate of the performance than the H method but does not require as much computation as the U method. In this method the total data is partitioned into J sets, each consisting of (N/J) samples. Out of the total N samples, (N/J) of them are used for testing purposes, while the training operation is performed on the remaining $(J-1) \cdot N/J$ samples. The performance of the system is estimated in turn J times, each time using a different set for testing, and the final estimate of the performance is obtained by averaging the results of each testing operation. Both H and U methods can be considered to be the special cases of this R method.

In our experiments the values of N and D are 147 and 7 respectively. Set 1 contained alphabets 1 to 21. Set 2 contained alphabet 22 to 42, and so on, with the result that each of the seven disjoint sets contained three alphabets from each of the seven different authors.

The components of the 25-dimensional feature vector within each class were assumed to be statistically independent, in order to minimize the complexity of the problem*. Therefore, for each feature component,

---

* The features were assumed to be statistically independent on an ad hoc basis. Unfortunately, the process of checking the validity of such assumption involves laborious computation [60]. Recently this author [60] has proposed a sequential test to determine whether a set of binary valued feature components are first order Markov or statistically independent. The test requires only the feature values and therefore can be performed before computing the feature likelihoods and the actual classification of the patterns.

17 different parameters per pattern class, each one corresponding to the probability $P(x_i = j/\omega_\ell)$, $i = 1, 2, \ldots, 25$, $j = 1, 2, \ldots, 17$, $\ell = 1, 2, \ldots, 26$, had to be estimated. The Baye's estimate [61], with square error loss, was obtained for each parameter using the specific set of training data. Thus $P(x_i = j/\omega_\ell)$ was set equal to $(n_{ij} + 1)/(N + \alpha_i)$, where n_{ij} denotes the number of samples out of the total of N pattern samples from class ω_ℓ in which the i th component of the feature vector takes on the j th value and α_i is the number of values that the i th feature component can take.

It was also necessary to estimate the a priori statistics $P(\omega_i)$, and for classification schemes with contextual constraints, the bigram and the trigram statistics. These statistics were estimated from an English text [62] containing approximately 340,000 characters. The trigram statistics of the 27 characters (including blank) were obtained using a character handling routine developed by Dykama [63]. The bigram and the a priori statistics were then obtained as follows:

$$P(\omega_i^k, \omega_j^{k-1}) = \sum_{\ell=1}^{27} P(\omega_i^k, \omega_j^{k-1}, \omega_\ell^{k-2}), \quad i,j=1,2,\ldots,27,$$

and

$$P(\omega_i^k) = \sum_{j=1}^{27} P(\omega_i^k, \omega_j^{k-1}), \qquad k = 1, 2, \ldots.$$

The Bayes estimate was obtained in each case such that all the entries of the bigram and trigram tables were nonzero.

4.5 Preliminary Recognition Experiments

In order to assess the usefulness of the feature extraction and the size-normalization preprocessing algorithms, two sets of preliminary experiments were conducted. The experimental results also clearly indicate why we advocate the R method for significant estimation of the performance of any recognition system.

Experiment 4.1:

In this set of experiments, the entire 3822 characters were recognized using a simple Bayes nonsequential classifier, first assuming equiprobable characters ($P(\omega_i) = 1/26$) and again when the character probabilities equalled those of the English text. In each one of the J (J=7) testing operations, the probability of error $P(\varepsilon/\omega_i)$, conditioned on class ω_i was calculated and from these the total error probability $P(\varepsilon)$ was obtained as follows:

$$P(\varepsilon) = \sum_{i=1}^{26} P(\varepsilon/\omega_i)\, P(\omega_i).$$

Averaging the results of J experiments yielded an overall average error probability while the standard deviation provided a measure of confidence associated with it. Table 4.1 shows the error probabilities obtained using each of the seven disjoint sets for testing. Also shown are the mean value of the error probability and the corresponding standard deviation.

Table 4.2 summarizes results obtained by others who have used Munson's multiauthor data file. Exact comparisons between these results and those obtained using our recognition system are not possible because

A Priori Character Probabilities	Error Probabilities							Average Error Probability $P(\varepsilon)$	Standard Deviation of $P(\varepsilon)$
	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7		
Equiprobable Characters	22.8	25.6	30.9	24.5	24.5	33.3	30.9	27.5	3.7
English A Priori Probabilities	16.0	23.2	24.4	18.6	17.5	27.0	26.7	21.9	4.1

Table 4.1 Results of preliminary recognition experiments with Munson's multiauthor data file.

Description of Testing and Training Samples	Type of Classifier Used	Feature Extraction Method	Probability of Correct Recognition	Reject Probability	Error Probability
2340 Samples for training Testing set unspecified Munson [52]*	Table lookup	Contour Tracing	58	19	23
Alphabet sets 1-96 for training, alphabet sets 97-147 for testing Munson [51]	Adaptive Piecewise Linear	Feature-template (PREP.9-VIEW)	78	-	22
Same as above**	Same as above	Topological descriptors (TOPO)	77	-	23
Alphabets 1-30 for both training and testing [53]**	Table lookup	Characteristic loci	75.9++	11.4	12.7
Alphabets 31-60 for both training and testing [53]†	Same as above	Same as above	69.5++	15.1	15.4
Alphabets 1-60 for training and 1, 4, 7, ..., 58 for testing†	Same as above	Same as above	73.9++	13.0	13.1

*26 upper case letters only.
**Samples include the 10 numerals, the 26 upper case letters, and the 10 Fortran symbols.
†25 upper case samples - letter 0 is excluded.
++The results have been modified by us to include half of the ambiguities in [53] as correct recognition and half as error.

Table 4.2 Recognition results obtained by others using Munson's multiauthor data file.

of differences in performance evaluation procedures and partitioning of alphabet sets. Comparison of the performance capabilities of the various feature extraction schemes is even more difficult because of differences in classification procedures. Munson's results [51] were obtained using some of the alphabet sets for training and some or all of the remaining ones for testing. Knoll's results [53] were obtained by doing most of the training on alphabet sets 1-30. However, since human trial-and-error was the learning approach used for designing the table lookup contents, modifications were introduced after a preliminary recognition run was performed on alphabets 31-60. Recognition runs were then performed on all 60 alphabets to give the results shown in Table 4.2. Full 46-character FORTRAN alphabets were used by Munson to obtain recognition accuracies of 77% and 78%*, while the last three sets of results in Table 4.2 were obtained using 25-character alphabets which contained only upper case characters, the letter O being excluded. If one bears in mind that the full 46-character alphabet would normally be more difficult to recognize than the 26-character upper case alphabet, it would be natural to conclude that Munson's PREP. 9-VIEW and TOPO systems with adaptive piecewise classifiers would yield somewhat lower error probabilities than our own if all three systems were evaluated using identical test data and test procedures. Our system would likely outperform the contour tracing-table lookup system of Munson [51] and might outperform the table lookup-characteristic loci system of Knoll [53]. This latter system was evaluated using some of the same data for training and testing; if disjoint training and test data were used the

---

* Munson [51] achieved a recognition accuracy of 83 percent combining the classification responses of PREP 9-VIEW and TOPO.

recognition accuracies in lines 4-6 of Table 4.2 would likely decrease [48] - [50],[59]. One might argue that the recognition accuracies of the table lookup systems are lower than they would be if reject criteria were not used. Rejects, however, tend to be inherent in table lookup systems, and avoiding rejects is difficult.

In conclusion, our feature extraction scheme is much simpler than all those mentioned in Table 4.2 while the recognition results compare favourably with those obtained by others who have used Munson's data base.

Experiment 4.2:

In Experiment 4.2, the 42 samples of each upper case character which appeared to vary most in size and line thickness were first eliminated using visual examination. The remaining characters were used to form 105 alphabets, and the training and testing procedure described earlier was then used with N = 105 and M = 7 (see Table 4.3). In the first group of recognition tests an algorithm similar to the one described in Section 4.2 was used to size-normalize the original 24 × 24 character matrix to 20 wide × 21 high. The normalized matrix was divided into 28 non-overlapping regions 5 wide × 3 high, and a 28 dimensional feature vector was then obtained for use with the simple Bayes classifier. In the second group of tests, size-normalization was not used. The feature vector was obtained by the use of 36 non-overlapping 4 × 4 rectangles in the original character matrix. In both groups of tests in Experiment 4.2 the characters' probabilities equalled those of the English text. One concludes, from Table 4.3 that the size normalization pre-processing algorithm described in section 4.2 is very attractive for

Character Preprocessing	Error Probabilities							Average Error Probability $P(\varepsilon)$	Standard Deviation of $P(\varepsilon)$
	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7		
Size Normalized	20.8	25.8	21.7	22.3	25.0	38.1	28.0	26.0	5.4
Not Size Normalized	27.8	41.4	26.0	32.9	31.8	39.4	26.4	32.3	5.7

Table 4.3 Recognition results showing the advantage of size-normalization preprocessing.

the proposed feature extraction scheme. Table 4.1 also supporting the notion that a combination of heuristic and statistical procedures should be incorporated in pattern recognition systems [64], instead of over stretching either of them; resulting in the complicated or costly scheme.

Tables 4.1 and 4.3 show clearly that the recognition results depend quite strongly on both the way in which the data are partitioned into training and testing sets, and on the system evaluation procedure. If only one of the seven sets of alphabets had been used for testing and the other six for training, then the error probability in Experiment 4.1 would have varied between 16 and 27 percent depending on the test set. Even the relative comparison of two different systems, based on such arbitrary partitioning of the data set may be quite misleading. In Experiment 4.2, the improvement in system performance due to the size normalization preprocessing varied between 1 and 16 percent, again depending on the set of alphabets used for testing. In fact, from Table 4.3 one can observe that it is possible to arrive at entirely contradictory conclusions by basing the results on different partitionings. The advantage of using the rotation method for the estimation of system performance and the reason why the recognition results using standardized data set be accompanied by detailed description of how the data set was partitioned for training and testing purposes are clearly indicated.

4.6 Test Procedures and Organization of Results

The performance of various decision schemes was estimated on the basis of the recognition of seven different handprinted English

texts containing a total of 3570 characters. The texts were from seven different sources and were formed using the disjoint sets of alphabets put aside for the purpose of testing. Thus the first text was formed using characters from alphabets 1-21, the second text was formed using alphabets 22-43, and so on. Therefore, a total of 21 samples for each character-class (excluding blank, for which there was no restriction) were available for forming each Text. Unless some character-class had more than 21 occurrences in a text, each character sample was used only once for each piece of text, otherwise the samples were recycled. Punctuation marks other than the period were either removed or replaced by a period depending on their place of occurrence in the text. During the process of classification, the blanks and periods were assumed to be perfectly recognizable. Thus, in classification schemes with contextual constraint, the identity of a blank was always used in classifying the subsequent (and the previous sample if the decision was delayed) character sample. It follows, therefore, that at any instant k, the string of conditional feature vectors X^k, X^{k-1}, ... does not run back to X^1, but terminates at the first blank encountered. The text materials and their various statistics appear in Appendix C.

Let the i th component error probability $P_e(i)$, of a decision scheme represent the ratio of the number of character samples misclassified to the total number of characters in the i th, i = 1, 2, ..., 7, text. The expected error probability P_E of the decision scheme, averaged over the seven texts and the corresponding standard deviation σ are then given by

$$P_E = \frac{1}{7} \cdot \sum_{i=1}^{7} P_e(i), \quad \sigma = \{ \frac{1}{7} \cdot \sum_{i=1}^{7} [P_e(i)]^2 - [P_E]^2 \} \qquad (4.1)$$

The ith component average number of features per pattern (ANFP) $N_f(i)$, associated with a decision scheme denotes the ratio of the total number of features required in classifying the character samples of the i th text to the total number of characters in that text. The expected ANFP, N_F of the decision scheme, averaged over the seven texts and the corresponding standard deviation are given by a similar set of equations as in (4.1).

The sequential decision schemes simulated throughout the rest of the thesis were assumed to be finite in nature, that is, a maximum of 25 features available from a character sample. Thus, at the end of the 25th decision state, irrespective of whether enough information was available or not, a final decision was always forced on the character sample*. The number of such forced decisions should be considered as a significant factor in evaluating the performance of finite sequential decision schemes. In the absence of any upper limit on the number of features available, the patterns on which terminal decisions were forced would require observation of more features before the normal termination of the observation process. The result is that the average number of features required per pattern would actually be higher than what is obtained with a finite number of features. The error probability would also decrease due to the observation of additional features. As an example, let us consider two equally reliable recognition schemes R_1 and R_2. Out of 10 patterns to be classified, let R_1 require 10 features for each one of the 5 patterns and all the 25 features for each

---

* or a feature from the neighbouring patterns is observed in case of CSPR schemes (Chapters VI and VII).

one of the remaining 5 patterns. Thus, the ANFP and the amount of forced decision associated with R_1 are 17.5 percent and 50.0 percent respectively. However, R_2 requires 20 features for each one of the 5 patterns and 15 features for each one of the remaining 5 patterns. The ANFP and amount of forced decision associated with R_2 are, therefore, 17.5 percent and 0.0 percent respectively. Based on only the error probability and the ANFP one would tend to conclude that both R_1 and R_2 are of equal effectiveness. In fact R_2 should be considered to be more effective than R_1, since R_1 would have observed additional features on 50 percent of the total patterns if larger dimensional feature vectors had been available from the patterns. A more accurate approach in evaluating the performance of the finite sequential schemes is, therefore, to take into account the number of forced decisions in-conjunction with the error probability and ANFP. The recognition results of sequential decision schemes throughout the rest of the thesis are always accompanied by the component forced decision $P_f(i)$ and (or) the expected forced decision P_F, averaged over all the seven texts.

In order to reduce the cost and time of performing a large number of recognition experiments, the results are presented in the thesis in a special way. To observe the trend of performance of the various decision schemes, only the first text with total of 538 character samples was used for the initial evaluation. Thus, $P_e(1)$ (and $P_f(1)$ for sequential schemes) for various values of $N_f(1)$ are presented in most cases. To add significance to the results, tests were then conducted with all seven texts using a specific number of features per pattern (or cost per feature in case of sequential scheme). The expected values P_E, P_F and N_F and the corresponding standard deviations are therefore available for meaningful comparisons of the schemes.

CHAPTER V

BAYES SEQUENTIAL DECISION SCHEMES FOR MULTICLASS PATTERN RECOGNITION PROBLEMS

5.1 Introduction

A few previous attempts [16]-[19] have been made to formulate sequential decision for multiclass recognition problems. However, the schemes are either suboptimal in nature, especially when $M > 2$ or of limited use due to the computational difficulty.

The Bayes sequential decision scheme which minimizes the expected cost of terminal decision, including the cost of feature observation, is essentially a backward procedure. It is quite attractive for multiclass problems. The application of such optimal sequential decision schemes based on the technique of dynamic programming [20], in the area of pattern recognition is due to Fu, Cardillo and Chien [23]-[26]. In order to maintain the true optimality of the decision scheme it is necessary to consider the entire future at every decision state. In other words, it is needed to work backward from optimal future to optimal present behaviour and so on back into the past. Difficulty in storage requirement and computation are therefore of major concern in designing an optimal sequential decision scheme. A recognition system, for example, with D statistically indpendent features per pattern, each one taking on q discrete values, would require a storage of $\sum_{i=1}^{D} (q)^D$ numbers for the risk functions alone. The optimal decision scheme is therefore, impracticable in most situations.

One way to overcome this difficulty is to devise a suboptimal

decision process. A suboptimal procedure, called the <u>one-state ahead truncation</u> approximation [65] or <u>limited-depth lookahead</u> mode [66], brings about a considerable reduction in the storage requirement and still maintains the important attributes of sequential schemes. At every decision state the suboptimal procedure assumes the next state to be the final state. Thus the risk both of continuing and of terminating the feature observation process are computed on the basis of the behaviour of the next state alone. The fact that the states beyond the immediate next one are dropped from consideration allows the risks to be computed in a forward manner.

It is however, necessary to know the "effectiveness" of the sequential decision schemes based on the one-state ahead trunction approximation. Unfortunately a general analytical assessment is not available and no explicit experimental assessment exits. Cardillo and Fu's [65] computer simulated experiments to illustrate the effectiveness of the suboptimal solution involve the on-line ordering of features. Also the experiments were designed to compare the suboptimal solution with the optimal one, which is of limited interest. A comparison between the suboptimal sequential and the optimal nonsequential schemes would be more meaningful, because then one would know whether the saving in the cost of feature observation provided by the sequential scheme is sufficient to warrant its use in place of the nonsequential scheme. Moreover, due to the oversimplicity of the recognition problem simulated and experimental methodology used [65], the results are of limited value. In fact, the conclusion based on such experimental procedures can be misleading [44] (also Chapter IV).

In this chapter the simple sequential decision scheme based on

one-state ahead approximation and the optimal simple nonsequential decision scheme are carefully compared from the points of view of both expected cost of a terminal decision and computational difficulty. The comparison is based on the computer simulated recognition experiments using the data set described in Chapter IV. A general compound sequential decision scheme for recognition problems with various forms of dependency is formulated. The compound sequential decision scheme with first order Markov dependent hypotheses is considered later and is simulated on the computer. Its performance is again compared with that of optimal compound nonsequential and suboptimal simple sequential decision schemes.

5.2 Formulation of the Decision Rule

We have at any instant k, k = 1, 2, .., of the sequential decision process a set $\bar{X}^k = X^1, X^2, \ldots, X^{k-1}, X^k$ of feature vectors, on the basis of which the k th pattern sample has to be classified. A very general multicategory recognition system is again considered and is assumed to satisfy the dependence relations (2.1-a), (2.1-b) and (2.2-c). Let:

$f[X^k, \ldots, X^1; n_k]$ be the minimum expected risk at state n_k of the entire sequential decision process, having observed the sequence of features $(x_1^k, \ldots, x_{n_k}^k)$ on the k th pattern and having available the set $\{X^{k-1}, \ldots, X^1\}$ of feature vectors from previous (k-1) patterns.

$R_s[X^k, \ldots, X^1; n_k; \delta^k]$ be the risk involved in making an optimal decision δ^k at state n_k of the decision process having available the set $\{x_1^k, \ldots, x_{n_k}^k\}$ of features from the k th

pattern and having the set $\{X^{k-1}, \ldots, X^1\}$ of feature vectors from previous (k-1) patterns.

C_{n_k} be the total cost of feature observations in continuing the sequential decision process to the n_k th state.

$c(n_k)$ be the cost associated with the n_k th feature component.

The risk involved in making a final decision δ^k at any state n_k of the sequential decision process amounts to,

$$R_s[X^k, \ldots, X^1; n_k; \delta^k], \quad n_k = 1, 2, \ldots, D.$$

If however, it is decided to continue the decision process through the observation of the $x^k_{n_k+1}$ th feature component, then the risk involved is given by

$$\{C_{n_k} + c(n_k + 1)\} + \int_{\Omega_{x^k_{n_k+1}}} f[(X^k, x^k_{n_k+1}), \ldots, X^1; n_k+1].$$

$$P(x^k_{n_k+1}|X^k)\,dx^k_{n_k+1}, \quad x^k_{n_k+1} \varepsilon \tilde{X}^k$$

where $\Omega_{x^k_{n_k+1}}$ denotes the feature subspace corresponding to the (n_k+1)st feature component and $\tilde{X}^k$ denotes the set $(x^k_{n_k+1}, \ldots, x^k_D)$ of features yet to be observed at state n_k.

Hence by the principle of optimality, the basic functional equation, governing the infinite sequence of expected risk $f[X^k, \ldots, X^1; n_k]$; $n_k = 1, 2, \ldots,$ is [20]

$$f[X^k, \ldots, X^1; n_k] = \text{Min}\begin{cases} \text{STOP:} \quad R_s[X^k, \ldots, X^1; n_k; \delta^k] \\ \text{CONTINUE:} \ \{C_{n_k} + c(n_k+1)\} + \int_{\Omega_{x^k_{n_k+1}}} f[(X^k, x^k_{n_k+1}), \ldots, X^1; \end{cases} \tag{5.1}$$

$$n_k+1] \cdot P(x^k_{n_k+1}|X^k)\, dx^k_{n_k+1}, \quad x^k_{n_k+1} \in \tilde{X}^k .$$

In case of a finite (fixed number of features) sequential decision process the optimum stopping rule can be computed backwards, starting from the known risk function of the final state. At the final state D, we have

$$f[X^k, \ldots, X^1; D] = R_s[X^k, \ldots, X^1; D; \delta^k] .$$

It is now possible to compute the risk for each and every state $n_k < D$, through the functional equation 5.1. More specificly, starting with the known value of $f[X^k, \ldots, X^1; D]$, we can write for state (D-1)

$$f[X^k, \ldots, X^1; (D-1)] = \text{Min}\begin{cases} \text{STOP:} \quad R_s[X^k, \ldots, X^1; (D-1); \delta^k] \\ \text{CONTINUE:} \ \{C_{D-1} + c(D)\} + \int_{\Omega_{x^k_d}} f[(X^k, x^k_D), \ldots, X^1; D] \end{cases}$$

$$\cdot\ P(x^k_D|X^k)\, dx^k_D, \quad x^k_D \in \tilde{X}^k.$$

Continuing in the similar fashion we can write for the first state of the decision process

$$f(X^k, \ldots, X^1; 1] = \text{Min}\begin{cases} \text{STOP:} \quad R_s[X^k, \ldots, X^1; 1; \delta^k] \\ \text{CONTINUE:} \quad C_1 + c(2) + \int_{\Omega_{x^k_2}} f[(X^k, x^k_2), \ldots, X^1; 2] \end{cases}$$

$$\cdot\ P(x^k_2|X^k)\, dx^k_2, \quad x^k_2 \in \tilde{X}^k$$

where the risk $f[(X^k, x_2^k), \ldots, X^1; 2]$ is available from the second state of the decision process.

It is necessary at every state to compute the terminal decision risk $R_s[X^k, \ldots, X^1; n_k; \delta^k]$, $n_k = 1, 2, \ldots, D$, employing an optimal decision rule to form δ^k. Using compound decision rule, the risk of deciding $X^k \sim \omega_i^k$, at any state n_k is given by

$$\rho_k(i) = \sum_{j=1}^{M} \ell(\omega_j^k; d_i^k)\, P(\bar{X}^k; \omega_j^k), \quad i = 1, 2, \ldots, M.$$

The Bayes decision rule is the one for which

$$\delta^k = d_q^k \quad \text{if} \quad \rho_k(q) = \underset{i}{\text{Min}}\{\rho_k(i)\}, \; i = 1, 2, \ldots, M.$$

The corresponding risk involved in making the decision is

$$R_s[X^k, \ldots, X^1; n_k; \delta^k] = \underset{i}{\text{Min}}\{\rho_k(i)\}, \; i = 1, 2, \ldots, M; \; n_k = 1, 2, \ldots, D.$$

Formulation of the optimal sequential decision process being complete, the storage and the computation difficulty associated with the procedure should now be apparent. In the next section a suboptimal sequential decision scheme based on one-state ahead truncation scheme is presented. The suboptimal procedure is easily implementable and requires far less storage than the optimal procedure.

5.3 A Suboptimal Decision Rule Based on One-State Ahead Truncation

In a suboptimal sequential decision scheme based on one-state ahead truncation approximation, the next possible state is always assumed to be the final state of the observation process. Thus, the risk of either continuing or terminating the feature observation process is computable

using a forward procedure and requires very little or no additional memory beyond that is required for the nonsequential decision scheme. Let

$R_c[X^k, \ldots, X^1; n_k]$ be the minimum expected risk at state n_k involved in continuing the feature observation process; assuming the next state to be a terminal state, having available the sequence $(x_1^k, \ldots, x_{n_k}^k)$ of features from k th pattern and having the set $\bar{X}^{k-1}$ of (k-1) feature vectors from the past.

Based on the one-state ahead truncation approximation it is possible to express the risk of continuing the observation process to (n_k+1) st state as follows:

$$R_c[X^k,\ldots,X^1;n_k] = \{C_{n_k} + c(n_k+1)\} + \int_{\Omega_{x_{n_k+1}^k}} \min_i \{\sum_{j=1}^{M} \ell(\omega_j^k; d_i^k) P(X^k, x_{n_k+1}^k, \ldots X^1; \omega_j^k)\} \cdot$$

$$P(x_{n_k+1}^k | X^k) dx_{n_k+1}^k; \; i = 1, 2, \ldots, M; \; x_{n_k+1}^k \varepsilon \tilde{X}^k. \tag{5.2}$$

Using the same mathematical notations as in Chapter II, the joint probability $P(X^k, x_{n_k+1}^k, X^{k-1}, \ldots, X^1; \omega_j^k)$ can be rewritten as

$$P(X^k, x_{n_k+1}^k, X^{k-1}, \ldots, X^1; \omega_j^k) = \sum_{q\varepsilon\{I_j\}} P(X^k, x_{n_k+1}^k, X^{k-1}, \ldots, X^1; W_q^k). \tag{5.3}$$

Since the dependence relations (2.1-2), (2.1-b) and (2.1-c) are assumed to be satisfied and (2.9-a) and (2.9-b) apply, one can express (5.3) as follows:

$$P(X^k, x^k_{n_k+1}, X^{k-1}, \ldots, X^1; \omega^k_j) = \begin{cases} \sum\limits_{q\varepsilon\{I_j\}} P(X^k, x^k_{n_k+1} | \bar{Y}^{k-1}; W^k_q) \sum\limits_{\ell=1}^{\xi^{k-1}} P(\bar{X}^{k-1}; W^{k-1}_\ell) \cdot \\ \qquad P(W^k_q | W^{k-1}_\ell)\}, \quad k > 1 \\ P(X^k, x^k_{n_k+1} | \omega^k_j) P(\omega^k_j), \quad k = 1. \end{cases} \tag{5.4}$$

The probability density $P(X^k, x^k_{n_k+1} | \bar{Y}^{k-1}, W^k_q)$ and the transition probability $P(W^k_q | W^{k-1}_\ell)$ are all assumed available such that (5.4) is recursively computable at every instant of the decision process. Thus, the minimum expected risk (5.2) of continuing the feature observation process in a sequential scheme can be computed at every instant of the decision process from a continuously growing set of information without* however, any increase in the memory requirement above what is needed for making decisions in nonsequential schemes.

The risk of terminating the sequential process can similarly be expressed in terms of the known quantities. Thus,

$$R_s[X^k, \ldots, X^1; n_k; \delta^k] = \underset{i}{\mathrm{Min}}[\sum_{q\varepsilon\{I_j\}} \sum_{j=1}^{M} \ell(\omega^k_j; d^k_i) P(X^k | \bar{Y}^{k-1}; W^k_q) \cdot \{ \sum_{s=1}^{\xi^{k-1}}$$

$$P(\bar{X}^{k-1}; W^{k-1}_s) P(W^k_q | W^{k-1}_s)\}], \quad i = 1, 2, \ldots, M.$$

A sequential decision rule based on the limited depth look-ahead mode may now be formulated as follows; at state n_k

$$\text{if} \begin{cases} R_c[X^k, \ldots, X^1; n_k] < R_s[X^k, \ldots, X^1; n_k; \delta^k]: & \text{CONTINUE: the observation process,} \\ R_c[X^k, \ldots, X^1; n_k] > R_s[X^k, \ldots, X^1; n_k; \delta^k]: & \text{STOP: the observation process.} \end{cases}$$

$$n_k = 1, 2, \ldots, D-1. \tag{5.5}$$

---

* The additional memory needed to store the costs of the feature components in suboptimal sequential decision schemes is usually very small.

It is therefore possible at every state to compute the risk, both of continuing and of terminating the observation process in a forward manner, using the one-state ahead truncation approximation.

5.4 Compound Sequential Decision Scheme for First Order Markov Dependent Hypotheses

A simpler recognition system is considered in this section, with the following dependence relations,

1. The pattern classes are only first order (m=1) Markov such that (3.14) applies for all i, j = 1, 2, ..., M.
2. The feature vector at any instant k is conditionally independent of the neighbouring feature vectors and the pattern identities, such that (3.15) applies for all i, j = 1, 2, ..., M.
3. The feature components within each class are statistically independent such that

$$P(x_1^k, x_2^k, \ldots, x_{n_k}^k \mid \omega_i^k) = \prod_{j=1}^{n_k} P(x_j^k \mid \omega_i^k), \tag{5.6}$$

$$k = 1, 2, \ldots; \; i, j = 1, 2, \ldots, M.$$

Equation (5.4) now reduces to

$$P(X^k, x_{n_k+1}^k, \ldots, X^1; \omega_j^k) = \begin{cases} P(X^k, x_{n_k+1}^k \mid \omega_j^k)\{\sum_{\ell=1}^{M} P(\bar{X}^{k-1}; \omega_\ell^{k-1}) P(\omega_j^k \mid \omega_\ell^{k-1})\}, & k > 1 \\ P(X^k, x_{n_k+1}^k \mid \omega_j^k) P(\omega_j^k), & k = 1, \end{cases} \tag{5.7}$$

where $P(\omega_j^k \mid \omega_\ell^{k-1})$, j, ℓ = 1, 2, ..., M, called the bigram transition probability, is assumed to be known. The risk of continuing the sequential decision process, at any state n_k, is

$$R_c[X^k,\ldots,X^1;n_k] = \{C_{n_k} + c(n_k+1)\} + \int_{\Omega_{x^k_{n+1}}} \underset{i}{\mathrm{Min}}\{ \sum_{j=1}^{M} \ell(\omega_j^k;d_i^k)P(X^k;x^k_{n_k+1}|\omega_j^k)\cdot$$

$$\sum_{q=1}^{M} P(\omega_j^k|\omega_q^{k-1})P(\bar{X}^{k-1};\omega_q^{k-1})\}\ P(x^k_{n_k+1}|X^k)dx^k_{n_k+1};\ i=1,2,\ldots,M;\ x^k_{n_k+1}\epsilon\tilde{X}^k. \tag{5.8}$$

For discrete valued feature components the integration sign in (5.8) is replaceable by a summation sign.

The conditional probability $P(x_{n_k+1}|X^k)$ is obtainable at every state from the known probability functions of the observed features. In pattern recognition problems the statistical independence among feature components either implies the class conditional independence of (5.6) or independence over all pattern classes such that

$$P(x_1^k,\ldots,x^k_{n_k}) = \prod_{i=1}^{n_k} P(x_i^k). \tag{5.9}$$

Both (5.6) and (5.9) can not in general be satisfied simultaneously [67], [68]. In computing the risk in (5.8), the class conditional independence (5.6) is assumed. Therefore, probability $P(x^k_{n_k+1}|X^k)$ has to be expressed based on the assumption that only (5.6) applies [68]. Thus

$$P(x^k_{n_k+1}|X^k) = \sum_{j=1}^{M} P(x_{n_k+1}|\omega_j^k)P(\omega_j^k;X^k), \qquad n_k=1,2,\ldots,D$$

where the a posteriori probability $P(\omega_j|X^k)$ is given by

$$P(\omega_j^k|X^k) = \frac{P(x^k_{n_k}|\omega_j^k)P(\omega_j^k|x^k_{n_k-1},\ldots,x_1^k)}{\sum_{i=1}^{M} P(x^k_{n_k}|\omega_i^k)\cdot P(\omega_i^k|x^k_{n_k-1},\ldots,x_1^k)}, \quad j=1,2,\ldots,M$$

which can be computed iteratively from the known likelihoods of the feature components.

The risk of terminating the sequential decision process becomes

$$R_s[X^k,\ldots,X^1;n_k;\delta^k] = \underset{i}{\mathrm{Min}}\{\sum_{j=1}^{M} \ell(\omega_j^k;d_i^k)P(X^k|\omega_j^k) \cdot \sum_{q=1}^{M} P(\bar{X}^{k-1};\omega_q^{k-1})$$

$$\cdot\ P(\omega_j^k|\omega_q^{k-1})\},\ i = 1,\ 2,\ \ldots,\ M.$$

For the loss matrix defined in (2.5), the optimum terminal decision at state n_k is to accept

$$\delta^k = d_i^k \quad \text{if} \quad P(\bar{X}^k|\omega_i^k)P(\omega_i^k) = \underset{j}{\mathrm{Min}}\{P(\bar{X}^k|\omega_j^k)P(\omega_j^k)\},\ j=1,\ 2,\ \ldots,\ M. \tag{5.10}$$

A suboptimal procedure in compound decision schemes is to use the previous decision instead of the information from the past patterns [34]. At instant k, we have

$$P(\bar{X}^k;\omega_i^k) = P(X^k|\omega_i^k) \cdot P(\omega_i^k;\ \bar{X}^{k-1}). \tag{5.11}$$

If $\delta^{k-1} = d_q^{k-1}$, corresponding to the set $\bar{X}^{k-1}$ of feature vectors, then using the past decision (5.11) reduces to

$$\begin{aligned} P(\bar{X}^k;\ \omega_i^k) &= P(X^k|\omega_i^k)P(\omega_i^k;\delta^{k-1}) \\ &= P(X^k|\omega_i^k)P(\omega_i^k;\omega_q^{k-1}),\ i,q=1,2,\ldots M. \end{aligned} \tag{5.12}$$

The risk both of continuing and of terminating the observation process using the past decision alone can now be obtained using (5.12) instead of (5.7).

5.5 Simple Sequential Decision Scheme

In simple (as opposed to compound) sequential schemes, the pattern classes are assumed to be statistically independent such that

$$P(\omega_i^k | \omega_j^{k-1}, \ldots, \omega_q^1) = P(\omega_i^k), \; k = 1, 2, \ldots; \; i,j,q = 1,2,\ldots,M.$$

Therefore, the information from the neighbouring pattern is of no use as far as the classification of the present pattern is concerned. Thus, deleting the set of feature vectors from the past patterns in our notations, one can now write for any state n_k

$$R_s[X^k; n_k; \delta^k] = \underset{i}{\mathrm{Min}}\{ \sum_{j=1}^{M} \ell(\omega_j^k; d_i^k) P(X^k | \omega_j^k) P(\omega_j^k)$$

and

$$R_c[X^k; n_k] = \{C_{n_k} + c(n_{k+1})\} + \int_{\Omega_{x^k_{n_k+1}}} \underset{i}{\mathrm{Min}} \{ \sum_{j=1}^{M} \ell(\omega_j^k, d_i^k) P(X^k, x^k_{n_k+1} | \omega_j^k) P(\omega_j^k) \}$$

$$\cdot \; P(x^k_{n_k+1} | X^k) \, dx^k_{n_k+1}, \; k = 1,2,\ldots \qquad (5.13)$$

The stopping rule based on the one-state ahead truncation approximation is the same as in (5.5). The optimum terminal decision at any state n_k is to accept

$$\delta^k = d_i^k, \text{ if } \; P(X^k | \omega_i^k) P(\omega_i^k) = \underset{j}{\mathrm{Min}}\{P(X^k | \omega_j^k) P(\omega_j^k)\},$$

$$j = 1, 2, \ldots, M. \qquad (5.14)$$

The set of risks derived in (5.13) is the same as defined in [11], and [25], where only the simple sequential decision schemes have been considered.

5.6 Experiments and Results

Several sets of experiments were conducted to compare the performance of optimal nonsequential and suboptimal (one-state ahead truncation) sequential decision schemes. The experiments described below were performed on Munson's data set, as reported in Chapter IV.

Experiment 5.1: The suboptimal simple sequential decision scheme of Section 5.5 and the optimal simple nonsequential schemes were considered in this set of experiments. The features were available in their natural order and were assigned equal amount of cost in case of the sequential scheme. Each character's a priori probability equalled that of English text. Quantities $P_e(1)$ and $P_f(1)$ for various values of $N_f(1)$ are shown in Figure 5.1. The results of recognition of all seven texts appear in Table 5.1, with the expected values P_E, P_F and N_F and the corresponding standard deviations being presented in Figure 5.1.

Experiment 5.2: The same set of decision schemes as in Experiment 5.1 were considered, except that the features were presented in three different orders. The purpose of this set of experiments is to compare the performance of the decision schemes when the features are preordered according to their "goodness". The features, in case of sequential schemes were again assumed equally costly. Three feature evaluation criteria were used to obtain the ordering of the feature components. These are listed below.

1. Bayes expression for error probability with English a priori probabilities (Chapter VIII).

2. On-line ordering technique using suboptimal sequential decision scheme (Chapter VIII).
3. Expected divergence between the class conditional densities. The feature components are assumed statistically independent within each class (Chapter VIII).

The results are presented in Figure 5.2.

Experiment 5.3: The optimal compound nonsequential and suboptimal compound sequential (Section 5.4) decision schemes were considered. The characters of the texts were assumed to be first order Markov, such that only bigram statistics were used and the dependency in pattern classes extended only to the finite past. In sequential schemes all the features were assigned equal amount of cost. Figure 5.3 shows the quantities $P_e(1)$, $P_f(1)$ for various values of $N_f(1)$. The results of recognition of all seven texts appear in Table 5.2 with the expected values P_E, P_F and N_F including the corresponding standard deviations, being presented in Figure 5.3.

5.7 Discussion of the Experimental Results

The results compiled in Figure 5.1, Figure 5.2 and Table 5.1 clearly indicate that the simple sequential decision scheme based on the one state-ahead truncation approximation performs significantly better than the optimal nonsequential decision scheme. We notice from Figure 5.1 that at 20 percent error rate, the suboptimal sequential decision scheme requires approximately 15.4 percent fewer features (on the average) per pattern sample than the optimal nonsequential decision scheme. This may mean a considerable saving in cost and time of feature observation

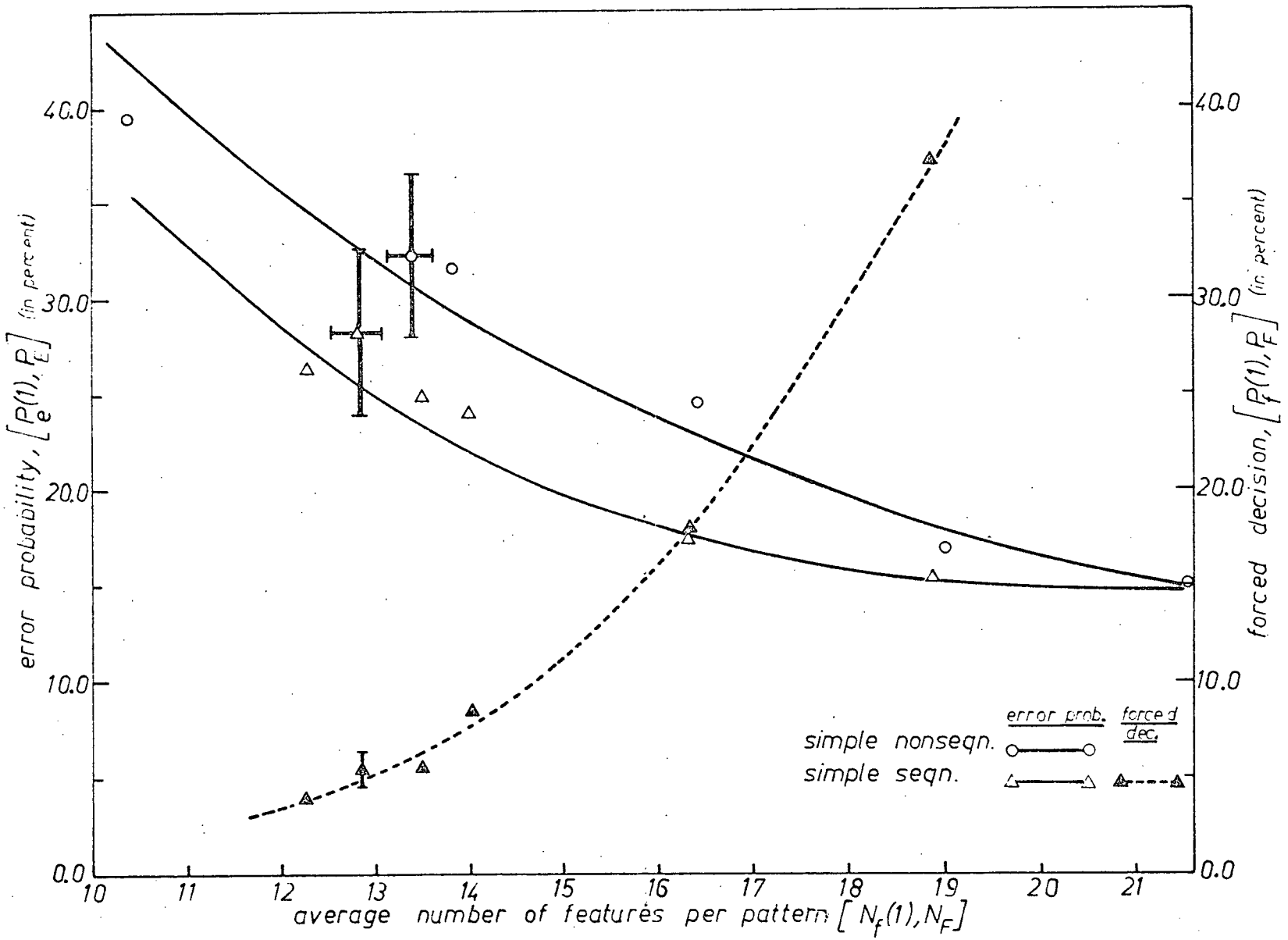

Figure 5.1 Comparison of simple suboptimal sequential and optimal non-sequential decision schemes.

Text		Set of Alphabets Used for		TYPE OF DECISION SCHEME				
				Nonsequential		Sequential		
No.	Size	Training	Testing	Avg. No. of Features Pattern	Error Prob. (in percent)	Avg. No. of Features Pattern	Error Prob. (in percent)	Forced Dec. (in percent)
1	538	22-147	1-21	13.8	31.8	13.5	25.0	5.7
2	420	1-21 43-147	22-42	13.6	38.3	12.8	36.1	6.9
3	579	1-42 64-147	43-63	13.3	33.2	12.6	28.6	5.9
4	516	1-63 85-147	64-84	13.1	28.1	12.6	24.7	6.0
5	467	1-84 106-147	85-105	13.5	30.6	12.9	26.1	5.5
6	557	1-105 127-147	106-126	13.1	37.9	12.5	33.9	5.7
7	486	1-126	127-143	13.1	26.1	12.7	24.3	3.9
Expected Values				13.4	32.3	12.8	28.4	5.7
Standard Deviations				0.274	4.25	0.31	4.41	0.83

Table 5.1 Recognition results of all seven texts using suboptimal simple sequential and optimal simple nonsequential decision schemes.

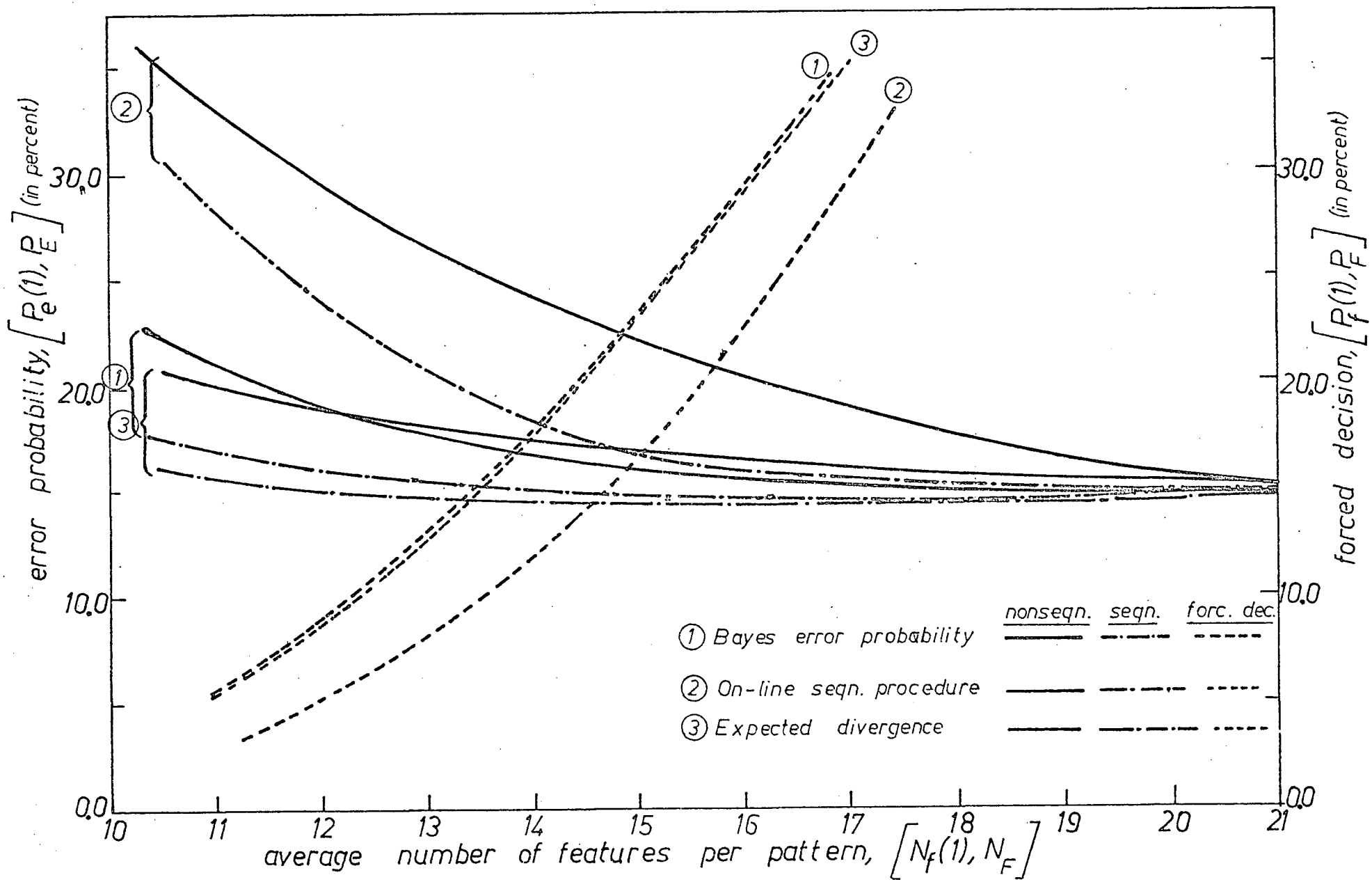

Figure 5.2 Comparison of simple suboptimal sequential and optimal nonsequential decision schemes with preordered sets of features.

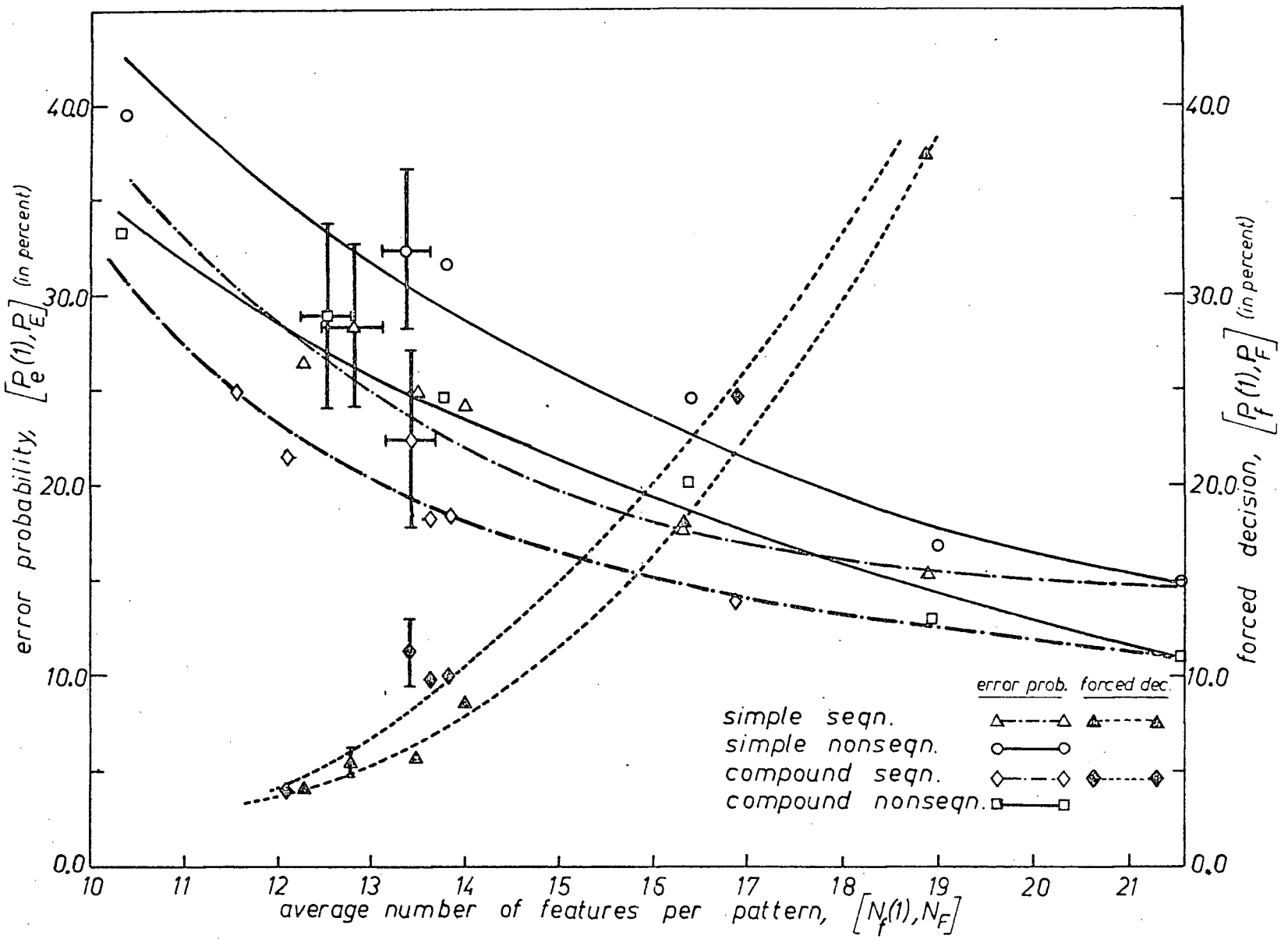

Figure 5.3 Comparison of simple sequential, simple nonsequential, compound sequential, and compound nonsequential decision schemes.

Text		Set of Alphabets Used for		TYPE OF DECISION SCHEME				
				Nonsequential		Sequential		
No.	Size	Training	Testing	Avg. No. of Features Pattern	Error Prob. (in percent)	Avg. No. of Features Pattern	Error Prob. (in percent)	Forced Dec. (in percent)
1	538	22-147	1-21	12.9	24.9	13.8	18.0	10.0
2	420	1-21 43-147	22-42	12.7	36.6	13.7	28.3	12.8
3	579	1-42 64-147	43-63	12.5	30.1	13.2	21.2	10.9
4	516	1-63 85-147	64-84	12.2	23.6	13.2	19.4	10.1
5	467	1-84 106-147	85-105	12.6	27.4	13.4	20.3	11.3
6	557	1-105 127-147	106-126	12.3	35.5	13.6	31.1	13.3
7	486	1-126	127-143	12.3	24.7	13.2	18.8	11.4
Expected Values				12.5	30.0	13.4	22.46	11.4
Standing Deviations				0.25	4.9	0.27	4.73	1.84

Table 5.2 Recognition results of all seven texts using suboptimal compound sequential and optimal compound nonsequential decision schemes.

during the classification of pattern samples. However, the sequential decision scheme, due to the nature of the stopping rule, requires more computation than the nonsequential scheme. In order to have an estimate of the computational complexity of the two schemes, Table 5.3 was prepared, which indicates the number of mathematical operations involved in implementing the nonsequential and the sequential decision rules. The total time required by the central processing unit (CPU) of the IBM system/360 Model 67 digital computer in recognizing the seven texts were also noted. Two different programs, one for the sequential and the other for the nonsequential decision schemes, written in Fortran IV, were used. No attempt was made to optimize the programs. Using a FORTRAN H compiler, the sequential decision scheme, observing 12.8 features on the average per pattern was found to require approximately 650 seconds of CPU time, while the nonsequential decision scheme using 13.4 features on the average per pattern needed 165 seconds of CPU time to classify all seven texts. In this particular case, the amount of computation for the suboptimal sequential scheme is higher by a factor of four than the optimal nonsequential scheme. Considering the fact that the sequential scheme requires significantly fewer number of features (on the average) for classification, the scheme may be quite attractive for pattern recognition and signal detection in spite of its extra computational complexity. It is possible that through some computational technique or approximation [25] the amount of computation required for the suboptimal sequential decision scheme can further be reduced without much sacrifice in the performance, making it even more attractive.

The advantage of using contextual information in sequential decision schemes is evident from Figure 5.3. The compound sequential

NONSEQUENTIAL SCHEME

Total Complexity for D States

MATHEMATICAL OPERATIONS					
multiplication	division	addition	substraction	maximization	minimization
D.M	-	-	-	1	-

SEQUENTIAL SCHEME

Complexity Per State

Process involved	MATHEMATICAL OPERATIONS					
	multiplication	division	addition	subtraction	maximization	minimization
risk of continuing	M + 1	-	3M + 1	1	1	-
risk of terminating	-	-	M	1	1	-
decision	-	-	-	-	-	1
updating	M	-	1	-	-	-
TOTAL/STATE	2M + 1	-	4M + 1	2	2	1

Table 5.3 Comparison of simple sequential and simple nonsequential.

decision scheme with bigram information performs significantly better than the simple sequential and the compound nonsequential schemes. At 17.5% error rate for example, the compound sequential decision scheme based on the one-state ahead truncation approximation, requires approximately 15 percent fewer features (on the average) per pattern than the suboptimal simple sequential scheme and approximately 19 percent fewer than the optimal compound nonsequential scheme. The amount of computation for the compound sequential decision scheme, based on the CPU time needed for the classification of seven texts, was found to be higher by a factor of three than the compound nonsequential decision scheme. Thus, the compound sequential decision scheme, which is easy to implement and requires no more storage than the compound nonsequential decision scheme, may be quite attractive for recognition problems with memory in source and observation medium.

CHAPTER VI

SUBOPTIMAL COMPOUND SEQUENTIAL PATTERN RECOGNITION SCHEMES FOR DEPENDENT HYPOTHESIS PROBLEMS

6.1 Introduction

In many recognition problems, a subset of available features may contain very little discriminatory information or some pattern samples may be so "noisy" that the available features may not sufficiently characterize the patterns. In such circumstances the sequential decision schemes discussed so far in the thesis always observe some features which provide very little or no additional information for discrimination but increase the expected cost of a terminal decision. Because the process of feature observation in a sequential decision scheme is continued until sufficient confidence for a final decision is achieved or the cost of feature observation becomes excessive. In recognition problems with dependent hypotheses, this form of redundancy in the feature observation can be avoided if the required additional feature is allowed to be observed on any one of the pattern samples of the sequence. The feature may be observed either on the pattern under consideration (one to be decided) as in the classical case (compound sequential decision scheme of Chapter V), or any one of the neighbouring patterns. The pattern selected and the feature actually observed on it is the one which provides the maximum amount of additional information.

In this chapter a compound sequential pattern recognition (CSPR) scheme [69] for dependent hypothesis problems is proposed. The scheme, during the process of classification of a pattern sample of a sequence

may observe the required additional feature on any one of the patterns of the sequence. A final decision on the pattern under consideration is drawn on the basis of the total available information. The scheme therefore, observes the most useful features from each pattern sample, resulting in a termination of the feature observation process earlier than the equally reliable classical sequential and nonsequential decision schemes.

Another obvious merit of the CSPR scheme is that it permits the look-ahead mode [33] of decision in the sequential recognition scheme. That is, additional contextual information from the subsequent patterns may also be available for the classification of the pattern sample under consideration. Such a sequential decision scheme using the knowledge of contextual dependencies both from the previous and the succeeding patterns of the sequence would be quite effective for dependent hypothesis problems from the point of view of minimizing the expected cost of a terminal decision.

The CSPR scheme may be particularly attractive for recognition problems with variable dimensional feature vectors [39], [52], [54]. It is possible that even after observing the entire set of available features from a pattern sample, especially if the set is of relatively small dimensionality, the information may not be adequate to make a final decision with sufficient confidence. If the pattern classes are dependent, the CSPR scheme may gain the additional confidence for a final decision through the observation of features from neighbouring patterns. For the same reason, the CSPR scheme will be useful for finite sequential recognition problems where a terminal decision is forced to be made at the end of the final state irrespective of whether sufficient information is available or not.

The suboptimal sequential decision rule for the CSPR scheme in its general form is formulated in Section 6.3. Three special cases based on simple but realistic assumptions are then formulated in Section 6.4.

6.2 Statement of the Problem

The decision function for the CSPRS is derived on the basis of the following assumptions:

1. Memory both in the source and observation medium may exist, such that the dependence relation (2.1-a), (2.1-b) and (2.1-c) are all satisfied. However, a decision can be delayed until additional information from the subsequent patterns is available. Thus, the dependencies extend also to the finite future.
2. Only one feature may be observed at any decision state.
3. One state-ahead truncation procedure is used to approximate the optimum backward procedure for implementing the decision rule.

At any instant k, let a string of t patterns be <u>under observation</u> and let

$$\begin{cases} (X^{k+t_1}, \ldots, X^k, \ldots, X^{k-t_2}), & \text{if } k \geq t_2 \\ (X^{k+t_1}, \ldots, X^k, \ldots, X^1), & \text{if } k < t_2 \end{cases}$$

be the corresponding set of feature vectors, where

t_1 denotes the extent of dependency in the future. That is, the number of patterns looked ahead at every instant of the decision process,

thus, $t_2 = (t - t_1 + 1)$

we shall denote $t_o = (t_2 + 1)$.

The required additional feature can be observed on any one of these t patterns. While trying to decide on the k th pattern of the sequence, the decisions of the previous t_2 patterns are assumed to be tentative. The decisions, therefore, may be altered if necessary due to the observation of additional features. Let

$$N^t = \begin{cases} (n_{k+t_1}, n_{k+t_1-1}, \ldots, n_k, \ldots, n_{k-t_2}), & k \geq t_2 \\ (n_{k+t_1}, n_{k+t_1-1}, \ldots, n_k, \ldots, n_1), & k < t_2 \end{cases}$$

denote a decision state of the sequential process and let N_i^t, $i = 1, 2, \ldots, z(t)$, where

$$z(t) = \begin{cases} (t)^D, & k \geq t \\ (k)^D, & k < t \end{cases}$$

be any convenient way of subscripting the decision states. Let

$$\eta(N_i^t;\ \zeta^k) \triangleq \{N_i^t(j);\ j = k-t_2, \ldots, k, \ldots, k+t_1\}$$

be a set of ζ^k decision states where each element $N_i^t(j)$ denotes a state to which transition from state N_i^t is possible through the observation of the n_j th feature on the j th pattern of the sequence. Since only one feature is allowed to be observed at any decision state, the number of elements in the set $\eta(N_i^t;\ \zeta^k)$ at any instant k is given by

$$\zeta^k \leq \begin{cases} t\ , & k \geq t_2 \\ k\ , & k < t_2\ . \end{cases}$$

The set $\eta(N_i^t;\ \zeta^k)$ is the null set when no further transition is possible from state N_i^t.

Associated with each decision state is a set

$$\Delta^{t_o} \triangleq \begin{cases} (\delta^k, \delta^{k-1}, \ldots, \delta^{k-t_2}), & k \geq t_2 \\ (\delta^k, \delta^{k-1}, \ldots, \delta^1), & k < t_2 \end{cases}$$

of t_o optimum (Bayes) terminal decisions, where δ^j, $j = (k-t_2), \ldots, k$, represents the optimum decision associated with the j th pattern.

Finally, let:

$r_j[\delta]$ be the component risk involved in forming an optimal terminal decision on the j th, $j = (k-t_2), \ldots, k$, pattern of the sequence based on the set $\bar{X}^{j+t_1}$ of feature vectors.

$R_s[\Delta^{t_o}:N_i^t]$ be the minimum expected compound risk of forming the set Δ^{t_o} of optimum decisions at state N_i^t of the decision process based on the available set of pattern vectors.

$\Phi_j[N_i^t(n_q)]$ be the expected component risk involved with the j th, $j = (k-t_2), \ldots, k$, pattern of the sequence in observing an additional feature x_{n_q} on the q th, $q = (k-t_2), \ldots, k, \ldots, (k+t_1)$, pattern

$R_c[\eta(N_i^t;\zeta^k):\hat{x}]$ be the minimum expected compound risk of observing the best <u>additional feature $\hat{x}$</u> from one of the ζ^k patterns at state N_i^t of the decision process.

$C(N_i^t)$ be the total cost of feature observation in continuing the decision process up to the state N_i^t.

6.3 Formulation of the Suboptimal Sequential Decision Rule

A suboptimal sequential decision rule based on one state-ahead truncation approximation, where the next decision state is considered to be the final state may be expressed at any instant k, as follows:

$$\text{if:}\quad \begin{cases} R_s[\Delta^{t_o}:N_i^t] > R_c[\eta(N_i^t:\zeta^k);\hat{x}]: & \text{CONTINUE the decision process} \\ R_s[\Delta^{t_o}:N_i^t] < R_c[\eta(N_i^t:\zeta^k);\hat{x}]: & \text{STOP the decision process.} \end{cases} \tag{6.1}$$

The problem now is to express the risks in terms of known quantities requiring a minimum amount of computation and storage. At any instant k, the component risk involved in the decision d_i^j i.e. the decision that $X^j \sim \omega_i^j$ is

$$\rho_j(i) = \sum_{u=1}^{M} \ell(\omega_u^j;d_i^j)\ P(\bar{X}^{j+t_1};\ \omega_u^j),\quad i = 1, 2, \ldots, M .$$

If dependence relations (2.1-a), (2.1-b) and (2.1-c) apply and the dependencies may extend to the finite future, then

$$P(\bar{X}^{j+t_1};\ \omega_u^j) = \sum_{i\varepsilon\{I_u\}} P(\bar{X}^{j+t_1};\ W_i^j)$$

$$= \sum_{i\varepsilon\{I_u\}} \sum_{q=1}^{\xi} P(X^{j+t_1}|\bar{Y}^{j+t_1-1};\ W_q^{j+t_1}) \cdot P(\bar{X}^{j+t_1-1};W_q^{j+t_1},W_i^j)$$

$$= \sum_{i\varepsilon\{I_u\}} \sum_{q=1}^{\xi} \sum_{s=1}^{\xi} \ldots \sum_{v=1}^{\xi} \{P(X^{j+t_1}|\bar{Y}^{j+t_1-1};W_q^{j+t_1})$$

$$P(X^{j+t_1-1}|\bar{Y}^{j+t_1-2};W_s^{j+t_1-1})\ldots P(X^{j+1}|\bar{Y}^j;W_v^j) \cdot P(\bar{X}^j;W_i^j)\} \cdot \{P(W_q^{j+t_1}|W_s^{j+t_1-1})$$

$$\ldots P(W_v^{j+1}|W_i^j)\} \tag{6.2}$$

where

$$P(\bar{X}^j;W_i^j) = P(X^j|\bar{Y}^{j-1};W_i^j) \sum_{\ell=1}^{\xi} P(W_i^j|W_\ell^{j-1}) \cdot P(\bar{X}^{j-1};W_\ell^{j-1}), \tag{6.3}$$

(the superscript of ξ is neglected for the sake of simplicity). Thus $P(\bar{X}^{j+t_1}; \omega_u^j)$ is computable recursively starting with the joint probability $P(\bar{X}^{k-t_2-1}; W^{k-t_2-1})$ available at instant $(k-t_2-1)$ and the known state transition probabilities.

The decision d_ℓ^j on the j th pattern of the sequence is optimum, that is,

$$\delta^j = d_\ell^j, \quad \text{if } \rho_j(\ell) = \underset{q}{\text{Min}}\{\rho_j(q);\ q = 1, 2, \ldots M\}$$

and the corresponding risk involved is given by

$$r_j[\delta] = \underset{v}{\text{Min}}\{ \sum_{i\varepsilon\{I_u\}} \sum_{u=1}^{M} \ell(\omega_u^j; d_v^j) P(\bar{X}^{j+t_1}; W_i^j)\},\ v = 1, 2, \ldots, M. \tag{6.4}$$

Thus, the expected compound risk of forming the set Δ^{t_o} of optimum decisions at state N_i^t is

$$R_s[\Delta^{t_o}:N_i^t] = \frac{1}{t_o} \sum_{j=(k-t_2)}^{k} r_j[\delta]. \tag{6.5}$$

However, at state N_i^t the risk involved with j th pattern of the sequence, if an additional feature is desired from the q th pattern, may be expressed as follows:

$$\Phi_j[N_i^t(n_q+1] = C(N_i^t) + c(n_q+1) + \int_{\Omega_{x_{n_q+1}^q}} \underset{v}{\text{Min}}\{ \sum_{u=1}^{M} \ell(\omega_u^j; d_v^j)$$

$$\cdot\ P(\bar{X}^{j+t_1}, x_{n_q+1}^q; \omega_u^j)\}\ P(x_{n_q+1}^q | X^q)\, dx_{n_q+1}^q;$$

$$q = k - t_2, \ldots, k, \ldots, k+t_1; \; j = k - t_2, \ldots, k;$$

$$x^q_{n_q+1} \varepsilon \tilde{X}^q. \tag{6.6}$$

Following our assumptions, (6.6) can be rewritten as follows:

$$\Phi_j[N^t_i(n_q+1)] = C(N^t_i) + c(n_q+1) + \int_{\Omega_{x^q_{n_q+1}}} \underset{v}{\mathrm{Min}}\{ \sum_{i\varepsilon\{I_u\}} \sum_{u=1}^{M} \ell(\omega^j_u; d^j_v)$$

$$\cdot \; P(\bar{X}^{j+t_1}, x^q_{n_q+1}; W^j_i)\} \; P(x^q_{n_q+1} | X^q) dx^q_{n_q+1} \tag{6.7}$$

where

$$P(\bar{X}^{j+t_1}, x^q_{n_q+1}; W^j_i) = P(X^{j+t_1}, \ldots, X^j, \ldots [X^q, x^q_{n_q+1}], \ldots, X^1; W^j_i)$$

can be expressed in terms of known quantities as in (6.2). Since

$$\int_{\Omega_{x^q_{n_q+1}}} P(x^q_{n_q+1} | X^q) dx^q_{n_q+1} = 1,$$

(6.7) for any $k \geqslant j \geqslant (q + t_1 + 1)$ reduces to the following expression:

$$\Phi_j[N^t_i(n_q+1)] = C(N^t_i) + c(n_q+1) + \sum_{i\varepsilon\{I_u\}} \sum_{u=1}^{M} \ell(\omega^j_u; d^j_\ell) P(\bar{X}^{j+t_1}; W^j_i).$$

Thus, the expected compound risk associated with the set of t_o patterns of observing the additional feature $x^q_{n_q+1}$ at state N^t_i is given by

$$\frac{1}{t_o} \sum_{j=k-t_2}^{k} \Phi_j[N^t_i(n_q+1)], \; q = k-t_2, \ldots, k, \ldots, k+t_1; \; N^t_i(n_q+1) \varepsilon \eta(N^t_i; \zeta^k).$$

The minimum expected risk of continuing the feature observation process at state N_i^t, through the observation of on additional feature on any one of the ζ^k patterns under observation is then,

$$R_c[\eta(N_i^t;\zeta^k):\hat{x}] = \underset{q}{\mathrm{Min}}\{\frac{1}{t_o}\sum_{j=k-t_2}^{k} \Phi_j[N_i^t(n_q+1)]\}, \tag{6.8}$$

$$q = k-t_2,\ldots,k, \ldots, k+t_1.$$

If the feature observation process is to be continued, then the best feature to be observed at state N_i^t is any solution to (6.8). That is, at state N_i^t, $\hat{x} = x^q_{n_q+1}$, if

$$\frac{1}{t_o}\sum_{j=k-t_2}^{k} \Phi_j[N_i^t(n_q+1)] = \underset{\ell}{\mathrm{Min}}\ \frac{1}{t_o}\{\sum_{j=k-t_2}^{k} \Phi_j[N_i^t(n_\ell+1)]\} \tag{6.9}$$

$$q,\ \ell = k-t_2,\ldots,k,\ldots,k+t_1.$$

6.4 Special Cases -- CSPRS Type 1, Type 2 and Type 3

The compound sequential pattern recognition scheme proposed in Section 6.3 is very general in form. Therefore, a large variety of sequential feature "sampling" and classification schemes may be developed using different types of dependence relations, cost allocations and decision functions. In this section, three special cases based on fairly simple but realistic assumptions are formulated. The cases are as follows:

CSPRS Type 1:

In the CSPRS Type 1 system it is assumed that the dependencies

extend only to the finite past ($t_1 = 0$). The pattern classes are assumed to be first order Markov with known state transition probability. The feature vectors are considered conditionally independent as in (3.15).

Under these circumstances, $t = t_2 + 1$, and (6.4) reduces to

$$r_j(\delta) = \underset{q}{\mathrm{Min}}\{\sum_{u=1}^{M} \ell(\omega_u^j : d_q^j)\ P(\bar{X}^j;\omega_u^j)\},\quad q = 1, 2, \ldots, M,$$

where $P(\bar{X}^j;\omega_u^j)$ can be expressed in the form shown in (6.3) and for first order Markov dependence among the pattern classes reduces to the following form:

$$P(\bar{X}^j;\omega_u^j) = P(X^j|\omega_u^j) \sum_{v=1}^{M} P(\bar{X}^{j-1};\omega_v^{j-1})\ P(\omega_u^j|\omega_v^{j-1}).$$

The expected compound risk $R_s[\Delta^t;N_i^t]$ is now given by

$$R_s[\Delta^t;N_i^t] = \frac{1}{t} \sum_{j=k-t+1}^{k} r_j(\delta)\,. \tag{6.10}$$

Similarly, the minimum expected risk of continuing the decision process may be expressed as follows:

$$R_c[\eta(N_i^t:\zeta^k);\hat{x}] = \underset{q}{\mathrm{Min}}\{\frac{1}{t} \cdot \sum_{j=k-t+1}^{k} \Phi_j[N_i^t(n_q+1)]\},\ q = k-t+1, \ldots, k$$

$$= \underset{q}{\mathrm{Min}}\{C(N_i^t) + c(n_q+1) + \frac{1}{t} \sum_{j=k-t+1}^{k} [\int_{\Omega_{x_{n_q+1}^q}} \underset{v}{\mathrm{Min}}\{\sum_{u=1}^{M} \ell(\omega_u^j;\ d_v^i)$$

$$\cdot\ P(\bar{X}^j;x_{n_q+1}^q;\omega_u^j)\}\ P(x_{n_q+1}^q|X^q)dx_{n_q+1}^q]\},\ x_{n_q+1}^q \varepsilon \tilde{X}^q;\ q = k-t+1, \ldots, k.$$

Besides the memory required for storing the probability distribution of the feature components and the bigram statistics, some additional memory is needed to store the set of likelihoods, $\{P(X^j|\omega_i^j);\ i = 1, 2, \ldots, M\}$, $i = (k-t+1), \ldots, (k-1)$. The set $\{P(\bar{X}^{k-t};\ \omega_q^{k-t});\ q = 1, 2, \ldots, M\}$ of joint probabilities are assumed available at every instant k and can be obtained recursively from one instant to the other. Thus, the total additional memory requirement at any instant k is (t.M) numbers, which grow linearly with t, the number of patterns under observation.

CSPRS Type 2:

The CSPRS Type 2 is equivalent to the CSPRS Type 1, with the exception that the expected compound risk either of continuing or of terminating the feature observation process is equal to the risk involved only with the pattern under consideration. Thus, at any instant k an additional feature can be observed if necessary, on any one of the t patterns; however, the risk of the k th pattern alone is assumed to be effected. Therefore, the expected compound risk of forming the set Δ^{t_o} of terminal decisions is given by

$$R_s[\Delta^{t_o}:N_i^t] = r_k(\delta)$$

$$= \underset{j}{\mathrm{Min}}\{\sum_{u=1}^{M} \ell(\omega_u^k;d_j^k)\ P(\bar{X}^k;\omega_u^k)\}.$$

Similarly, the minimum expected risk of continuing the decision process at state N_i^t can be expressed as follows:

$$R_c[\eta(N_i^t:\zeta^k):\hat{x}] = \underset{q}{\mathrm{Min}}\{\Phi_k[N_i^t(n_q+1)]\},\ q = k-t+1,\ \ldots,\ k$$

$$= \underset{q}{\mathrm{Min}}\{C(N_i^t) + c(n_q+1) + \int_{\Omega_{x^q_{n_q+1}}} \underset{j}{\mathrm{Min}}\{\sum_{u=1}^{M} \ell(\omega_u^k;d_i^k)$$

$$P(\bar{X}^k, x^q_{n_q+1};\omega_u^k)\}\ P(x^q_{n_q+1}|X^q)dx^q_{n_q+1}\}\ ,$$

$$x^q_{n_q+1}\varepsilon\tilde{X}^q;\ j = 1,\ 2,\ \ldots,\ M;\ q = k - t + 1,\ \ldots,\ k$$

The scheme is, therefore, simpler in structure and requires less computation than the CSPRS Type 1.

CSPRS Type 3:

In CSPRS Type 3, the decision on a pattern may be delayed to have additional information from the subsequent patterns. Thus, the dependencies extend to the finite future as well. It is assumed that the pattern classes are first order Markov and the feature vectors are conditionally independent.

Let $t_1 = 1$, therefore $t = t_2$ and (6.4) reduces to

$$r_j(\delta) = \underset{v}{\mathrm{Min}}\{\sum_{u=1}^{M} \ell(\omega_u^j;d_v^j)\ P(\bar{X}^{j+1};\omega_u^j)\},\ v = 1,\ 2,\ \ldots,\ M.$$

The joint probability $P(X^{j+1};\omega_u^j)$ is expressible in the form shown in (6.2) and for first order Markov dependence among the pattern classes reduces to

$$P(\bar{X}^{j+1};\omega_u^j) = P(\bar{X}^j;\omega_u^j) \sum_{q=1}^{M} P(X^{j+1}|\omega_q^{j+1}) \cdot P(\omega_q^{j+1},\omega_u^j)$$

where the probability $P(\bar{X}^j;\omega_u^j)$ is given by (6.3). The expected compound risk $R_s[\Delta^{t_o}:N_i^t]$ is the same as in (6.10). However, at any state n_k the risk of continuing the decision process is

$$R_c[\eta(N_i^t:\zeta^k):\hat{x}] = \underset{q}{\mathrm{Min}}\{\frac{1}{t_o}\sum_{j=k-t_2}^{k} \Phi_j[N_i^t(n_q+1)]\},$$

$$q = k-t_2, \ldots, k, k+1$$

$$= \underset{q}{\mathrm{Min}}\{C(N_i^t) + c(n_q+1) + \frac{1}{t_o}\sum_{j=k-t+1}^{k}[\int_{\Omega_{x^q_{n_q+1}}} \cdot \underset{v}{\mathrm{Min}}\{\sum_{u=1}^{M} \ell(\omega_u^j;d_v^j)P(\bar{X}^{j+1},x^q_{n_q+1};\omega_u^j)\}\cdot$$

$$P(x^q_{n_q+1}|X^q)dx^q_{n_q+1}]\}.$$

The advantage with the CSPRS Type 3 is that it permits look-ahead mode [33] of decision in sequential classifiers. Thus, the additional information from the subsequent patterns may result in the earlier termination of the feature observation process in sequential decision schemes.

CHAPTER VII

ANALYTICAL AND EXPERIMENTAL EVALUATION OF THE CSPR SCHEMES

7.1 Introduction

In this chapter, the suitability and the performance capability of the CSPR schemes proposed in Chapter VI are assessed both analytically and experimentally. For the analytical part, a recognition problem with binary hypotheses and first order Markov dependence is considered. The effect of observing the additional feature either on the pattern under consideration or the neighbouring one, on the system error probability is then analyzed. For the experimental part, multiclass problems are considered. The results of simulating the three CSPR schemes discussed in Chapter VI are presented. Finally, the advantage of using the CSPR scheme in recognition problems with variable dimensional feature vectors is pointed out.

7.2 Analytical Evaluation Using Two-Class, First Order Markov Problem

The two pattern classes ω_1 and ω_2 of the recognition problem are assumed to be first order Markov such that at any instant k

$$P(\omega_1^{k+1}|\omega_2^k) = P(\omega_2^{k+1}|\omega_1^k) = p.$$

Therefore, $P(\omega_1) = P(\omega_2) = 0.5$.

Let the features be statistically independent with $P(x_i^k|\omega_j^k)$, $i = 1, 2, \ldots, D$; $j = 1, 2$, a univariate Gaussian distribution function with mean

μ_j and variance σ_i^2. Thus

$$P(x_i^k|\omega_j^k) = [(2\pi)^{-1/2} \cdot \sigma_i^{-1}] \quad \exp[-(x^k-\mu_j)^2/2\sigma_i^2].$$

Without any loss of generality it may be assumed that $\mu_2 = \mu$ and $\mu_1 = 0$. Finally, let the feature vectors be conditionally independent.

At any instant k, the k th component conditional risk for the decision d_j^k

$$R_k = \int \sum_{i=1}^{M} \ell(\omega_i^k; d_j^k) \; P(\bar{X}^k|\omega_i^k) \; P(\omega_i^k) d\bar{X}^k.$$

For the type of loss matrix defined in (2.5) the risk becomes the probability of error and, for k = 2, may be written as follows:

$$e_k(n_k;n_{k-1}) = 1 - \sum_{j=1}^{2} \int_{\Omega_k(j)} P(X^k, X^{k-1}|\omega_j^k) \; P(\omega_j^k) dX^k \; dX^{k-1}, \qquad (7.1)$$

where $\Omega_k(j)$ denotes the feature subspace corresponding to the decision d_j^k at instant k, and n_k and n_{k-1} are the dimensionality of feature vectors X^k and X^{k-1} respectively.

The error probability is minimized if the decision rule is such that

$$\delta^k = d_j^k, \text{ if } P(\bar{X}^k;k_j) = \underset{i}{\text{Max}}\{P(\bar{X}^k;\omega_i^k); \; i, j = 1, 2\}$$

in which case

$$\Omega_k(j) = \{\bar{X}^k : P(\bar{X}^k;\omega_j^k) \geq P(\bar{X}^k;\omega_i^k), \; i = 1, 2\}. \qquad (7.2)$$

Since,

$$P(\bar{X}^k;\omega_j^k) = P(X^k|\omega_j^k) \sum_{i=1}^{2} P(X^{k-1};\omega_i^{k-1}) \; P(\omega_j^k|\omega_i^{k-1}),$$

(7.2) is given by

$$\Omega_k(j) = \{X^k, X^{k-1}: \frac{P(X^k|\omega_j^k)}{P(X^k|\omega_i^k)} \geq \frac{\sum_{q=1}^{2} P(X^{k-1};\omega_q^{k-1})P(\omega_i^k|\omega_q^{k-1})}{\sum_{s=1}^{2} P(X^{k-1};\omega_s^{k-1})P(\omega_j^k|\omega_s^{k-1})}, \; i, j = 1, 2\}, \quad i \neq j. \tag{7.3}$$

Following the assumptions of the Markov dependence of the pattern classes, we now can express (7.1) as follows:

$$e_k(n_k;n_{k-1}) = \int_{\Omega_k(2)} P(X^k|\omega_1^k) \cdot \sum_{q=1}^{2} P(X^{k-1}|\omega_q^{k-1}) \cdot P(\omega_1^k|\omega_q^{k-1})\, dX^k \cdot dX^{k-1}. \tag{7.4}$$

For statistically independent feature components within each class we can write

$$P(X^k|\omega_i^k) = \prod_{j=1}^{n_k} P(x_j^k|\omega_i^k)$$

$$= [(2\pi)^{-n_k/2} \cdot \prod_{j=1}^{n_k} \sigma_j^{-1}] \cdot \exp[-0.5 \sum_{j=1}^{n_k} (x_j^k - \mu_i)\sigma_j^{-2}].$$

Let $\theta_i = (\sigma/\sigma_i)^2$, where σ is a constant. Therefore,

$$\frac{P(X^k|\omega_2^k)}{P(X^k|\omega_1^k)} = \exp[-\{\mu^2 \sum_{j=1}^{n_k} \theta_j - 2\mu \sum_{j=1}^{n_k} x_j^k \cdot \theta_j\}/2\sigma^2].$$

Taking logarithm on both sides yields

$$\ln\{\frac{P(X^k|\omega_2^k)}{P(X^k|\omega_1^k)}\} = (\frac{\mu}{\sigma^2}) \cdot \sum_{j=1}^{n_k} \theta_j (x_j^k - 0.5\mu).$$

Define,

$$f(X^{k-1}) = \frac{\sum_{q=1}^{2} P(X^{k-1};\omega_q^{k-1})P(\omega_1^k|\omega_q^{k-1})}{\sum_{s=1}^{2} P(X^{k-1};\omega_s^{k-1})P(\omega_2^k|\omega_s^{k-1})}$$

$$= \frac{(1-p)\cdot P(X^{k-1}|\omega_1^{k-1}) + p\cdot P(X^{k-1}|\omega_2^{k-1})}{p\cdot P(X^{k-1}|\omega_1^{k-1}) + (1-p)\cdot P(X^{k-1}|\omega_2^{k-1})} .$$

Equation (7.3) therefore, reduces to

$$\Omega_k(2) = \{X^k, X^{k-1}: \sum_{\ell=1}^{n_k} \theta_\ell x_\ell^k \geqslant 0.5 \sum_{q=1}^{n_k} \theta_q + \frac{\sigma^2}{\mu} \ln f(X^{k-1})\}$$

Let

$$v = \sum_{\ell=1}^{n_k} \theta_\ell x_\ell^k - 0.5\mu \sum_{q=1}^{n_k} \theta_q - \frac{\sigma^2}{\mu} \ln f(X^{k-1}).$$

Since x_ℓ^k, $\ell = 1, 2, \ldots, n_k$ are Gaussian random variables, v is a Gaussian random variable with mean and variance obtained as follows:

$$m_v = E[v]$$

$$= E[\sum_{\ell=1}^{n_k} \theta_\ell x_\ell^k] - 0.5\mu \sum_{q=1}^{n_k} \theta_q - \frac{\sigma^2}{\mu} \ln f(X^{k-1}) .$$

If $x_\ell^k \sim \omega_1^k$, then the random variable x^k is a zero mean process and therefore,

$$m_v = -0.5\mu \sum_{q=1}^{n_k} \theta_q - \frac{\sigma^2}{\mu} \ln f(X^{k-1}), \quad x_\ell^k \sim \omega_1^k .$$

Similarly, it can be shown that

$$\sigma_v^2 = E[v^2] - E[v]^2$$

$$= \sigma^2 \sum_{\ell=1}^{n_k} \theta_\ell, \quad x_\ell^k \sim \omega_1^k \quad .$$

Equation (7.4) now becomes

$$e_k(n_k;n_{k-1}) = \int_{X^{k-1}} [(1-p)\cdot P(X^{k-1}|\omega_1^{k-1}) + p \cdot P(X^{k-1}|\omega_2^{k-1})]dX^{k-1}$$

$$\cdot \int_{v>0} [(2\pi)^{-1/2} \cdot \sigma_v^{-1}]\exp[-0.5(v-m_v)^2/\sigma_v^2]dv$$

$$= \frac{(1-p)}{\sqrt{2\pi}\,\sigma_v} \int_{X^{k-1}} \int_{v>0} P(X^{k-1}|\omega_1^{k-1}) \cdot \exp[-0.5(v-m_v)^2/\sigma_v^2]dvdX^{k-1}$$

$$+ \frac{p}{\sqrt{2\pi}\,\sigma_v} \int_{X^{k-1}} \int_{v>0} P(X^{k-1}|\omega_2^{k-1})\exp[-0.5(v-m_v)^2/\sigma_v^2]dvdX^{k-1}.$$

At this stage let us assume that $p \lessapprox 1$ and since $1 \geq P(X^{k-1}|\omega_i) \geq 0$, $i = 1, 2$, we can write

$$f(X^{k-1}) = \frac{P(X^{k-1}|\omega_1^{k-1})}{P(X^{k-1}|\omega_2^{k-1})}$$

and, therefore,

$$\ln f(X^{k-1}) = -\frac{\mu}{\sigma^2} \sum_{i=1}^{n_{k-1}} \theta_i(x_i^{k-1} - 0.5\mu).$$

Let $y = (v-m_v)/\sigma_v$. Then

$$e_k(n_k;n_{k-1}) = \frac{(1-p)}{\sqrt{2\pi}} \int_{X^{k-1}} \int_b P(X^{k-1}|\omega_1^{k-1}) \cdot \exp[-0.5y^2]dydX^{k-1}$$

$$+ \frac{p}{\sqrt{2\pi}} \int_{X^{k-1}} \int_b P(X^{k-1}|\omega_2^{k-1}) \cdot \exp[-0.5y^2]dydX^{k-1} \tag{7.5}$$

where

$$b = -m_v/\sigma_v$$

$$= \{0.5\mu \sum_{q=1}^{n_k} \theta_q - \sum_{i=1}^{n_{k-1}} \theta_i(x_i^{k-1} - 0.5\mu)\}/\sigma[\sum_{\ell=1}^{n_k} \theta_\ell]^{1/2}.$$

Consider separately the parts of (7.5)

$$(a) = \frac{(1-p)}{\sqrt{2\pi}} \int_{X^{k-1}} \int_b P(X^{k-1}|\omega_1^{k-1})\exp[-0.5y^2]dydX^{k-1}$$

$$=(1-p) \frac{\prod_{j=1}^{n_{k-1}} \theta_j^{1/2}}{(2\pi)^{n_{k-1}/2}(\sigma)^{n_{k-1}}} \int_{-\infty}^{\infty} \exp[-0.5 \sum_{u=1}^{n_{k-1}} (x_u^{k-1})^2\theta_u/\sigma^2]dX^{k-1}\{\frac{1}{\sqrt{2\pi}} \int_b \exp[$$

$$-0.5y^2]dy\} \tag{7.6}$$

Let

$$z_i = x_i^{k-1} \cdot \theta_i^{1/2} .$$

Therefore,

$$b = \sum_{i=1}^{n_{k-1}} z_i b_i' + b_o$$

where

$$b_i' = -\frac{1}{\sigma}\left[\frac{\theta_i}{\sum_{j=1}^{n_k} \theta_j}\right]^{1/2}$$

and

$$b_o = 0.5\mu[\sum_{i=1}^{n_k} \theta_i + \sum_{j=1}^{n_{k-1}} \theta_j]/\sigma[\sum_{i=1}^{n_k} \theta_i]^{1/2}.$$

Equation (7.6) then reduces to the following expression

$$(a) = \frac{(1-p) \cdot \prod_{j=1}^{n_{k-1}} \theta_j^{1/2}}{(2\pi)^{n_{k-1}/2} \cdot (\sigma)^{n_{k-1}}} \int_{-\infty}^{\infty} \int_{b}^{\infty} \exp[-0.5(\sum_{q=1}^{n_{k-1}} z_q^2 + y^2)/\sigma^2]\, dydz^{k-1}. \quad (7.7)$$

After proper coordinate transformation (7.7) can be expressed as follows

$$(a) = \frac{(1-p)}{(2\pi)^{n_{k-1}/2} \cdot \sigma^{n_k}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp[-0.5(\sum_{q=1}^{n_{k-1}} z_q'^2)/\sigma^2]dz. \frac{1}{\sqrt{2\pi}} \int_{b} \exp[0.5y'^2]dy \quad (7.8)$$

where z' and y' are now independent random variables.

Equation (7.8) finally reduces to

$$(a) = 0.5(1-p)[1 - \mathrm{erf}\{(\mu/2\sigma)(A_{n_k,n_{k-1}}/2)^{1/2}\}]$$

where

$$A_{n_k,n_{k-1}} = \sum_{i=1}^{n_k} \theta_i + \sum_{j=1}^{n_{k-1}} \theta_j$$

and $\mathrm{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{o}^{x} \exp[-y^2]dy.$

It can similarly be shown that the second part of (7.5), denoted (b), reduces to

$$(b) = 0.5p[1 - \mathrm{erf}\{(\mu/2\sigma)B_{n_k,n_{k-1}} \cdot (2A_{n_k,n_{k-1}})^{-1/2}\}]$$

where

$$B_{n_k,n_{k-1}} = \sum_{i=1}^{n_k} \theta_i - \sum_{j=1}^{n_{k-1}} \theta_j .$$

Thus at any decision state with n_k and n_{k-1} number of features on the k th and the (k-1) st pattern respectively, the k th component error probability is given by

$$e_k(n_k;n_{k-1}) = 0.5 - 0.5[(1-p)\mathrm{erf}\{(\mu/2\sigma)(A_{n_k,n_{k-1}}/2)^{1/2}\}$$

$$+ p\ \mathrm{erf}\{(\mu/2\sigma)B_{n_k,n_{k-1}}\cdot(2A_{n_k,n_{k-1}})^{-1/2}\}] \qquad (7.9)$$

7.3 Results and Discussion for the Two-Class, First Order Markov Problem

In Figures 7.1, 7.2, and 7.3 the error probability $e_k(n_k;n_{k-1})$ for various combinations of n_k and n_{k-1} are presented, the distribution of the variance of the feature components being different in each case. The following three cases were considered.

Case 7.1: The feature components were all of uniform variance, thus

$$\sigma_i = \sigma, \quad i = 1, 2, \ldots, D.$$

Case 7.2: The variance of the feature components increased exponentially. Thus, the variance σ_i of the i th feature component was given by

$$\sigma_i = \sigma \cdot a_1 \exp[i \cdot a_2], \ i = 1, 2, \ldots, D$$

where a_1 and a_2 were prespecified constants.

Case 7.3: The variance of the feature components increased linearly. Thus, the variance σ_i of the i th feature component is given by

$$\sigma_i = \sigma\cdot(a_1' + a_2' \cdot i), \ i = 1, 2, \ldots, D$$

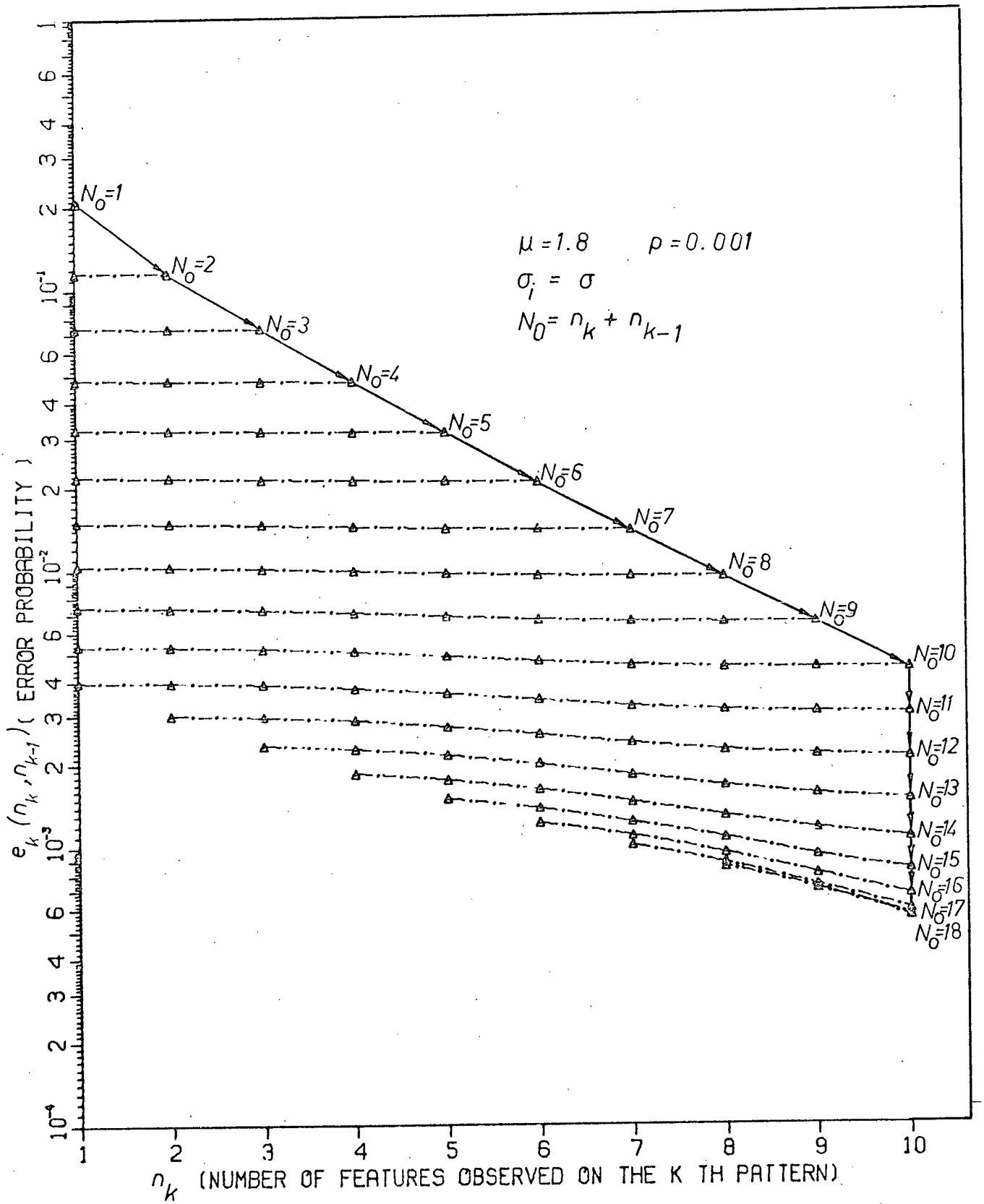

Figure 7.1 System error probability for various combinations of features from the present and the previous patterns. The arrows show the path of minimum error probability. The features' variance is uniformly distributed.

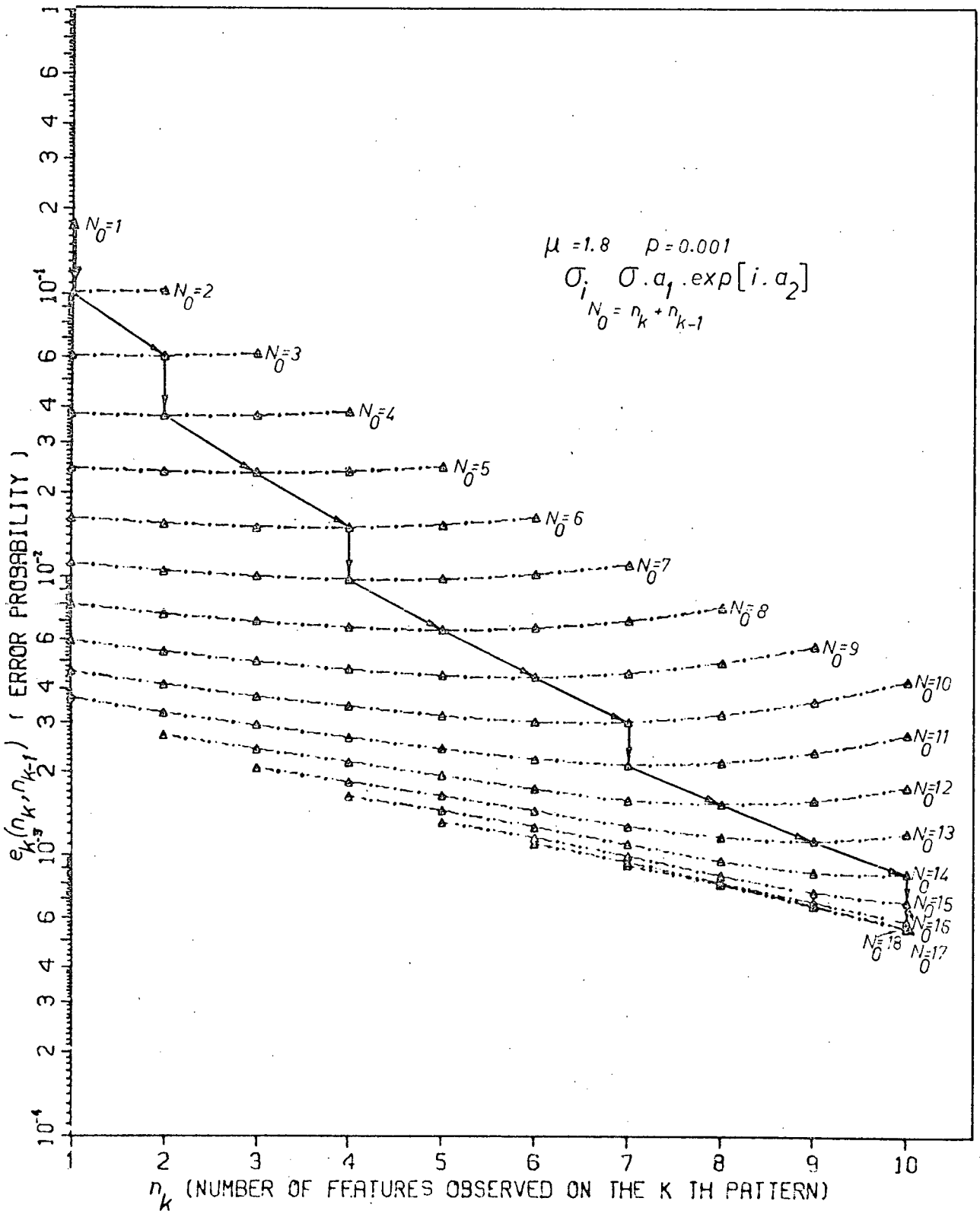

Figure 7.2 System error probability for various combinations of features from the present and the previous patterns. The arrows show the path of minimum error probability. The features' variance is exponentially distributed.

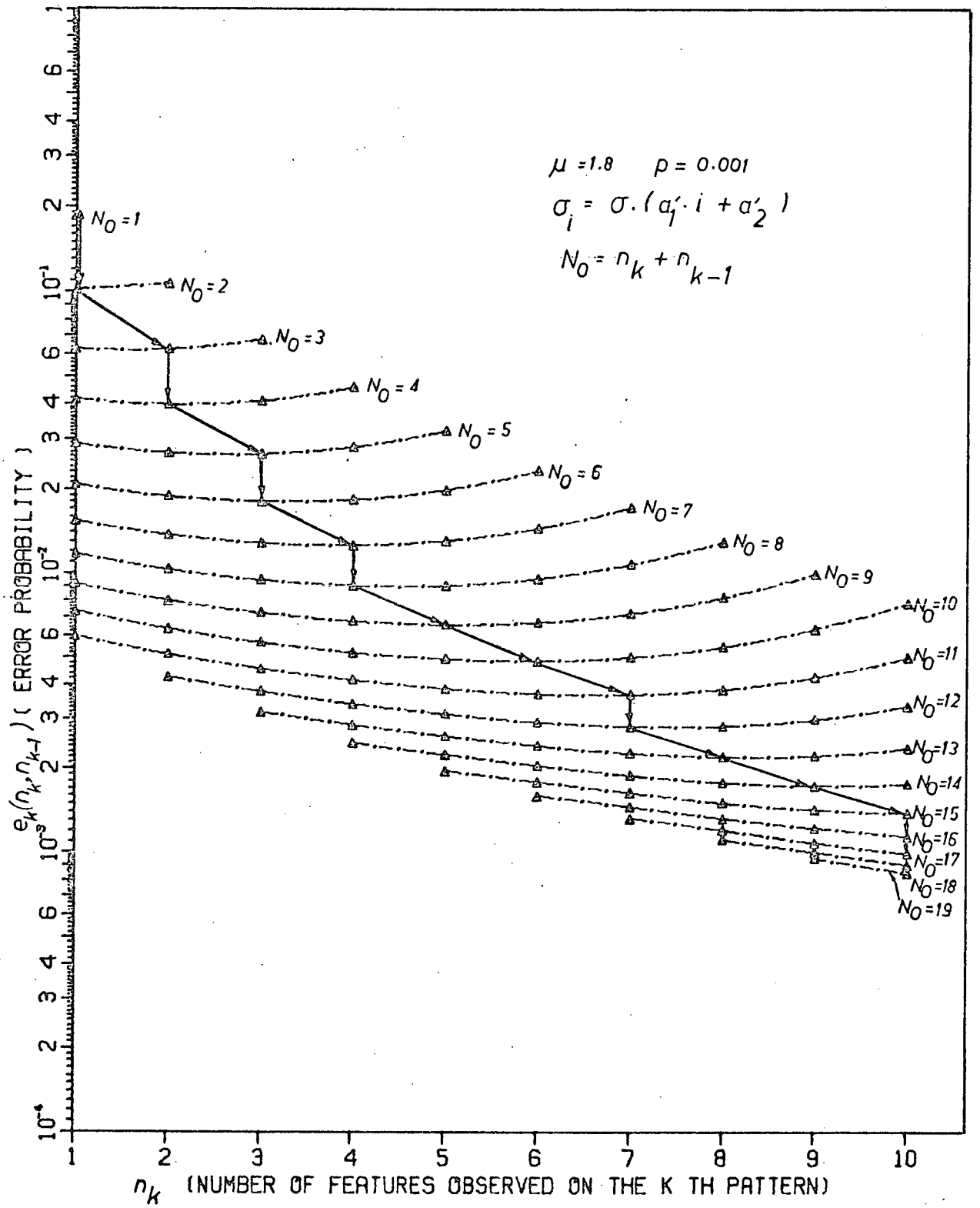

Figure 7.3 System error probability for various combinations of features from present and previous patterns. Features' variance is linearly distributed.

where a_1' and a_2' are two prespecified constants.

A maximum of ten features were assumed available per pattern sample (D = 10). Also, for each one of the three cases $\mu = 1.8$, $p = 0.008$ and $\sigma = 1$ were assumed. Note the way in which the error probabilities are plotted in Figure 7.1, 7.2, and 7.3. A curve on each figure corresponds to the variation of the error probability for various combinations of n_k and n_{k-1}, the features from the k th and k-1 st patterns, with the constraint that

$$n_k + n_{k-1} = \text{a constant} = N_o .$$

One can view the curve corresponding to $N_o = i$ to be the locus of the error probability associated with the i th state. The error probability associated with a state is dependent on how the state is arrived at. Transition from one state (except the final) to the next is possible through the observation of an additional feature either on the k th or (k-1) st pattern. The arrows in the figures show the locus of minimum error probability and therefore, the optimum way of sampling the pattern features.

When the features are equally effective as in Figure 7.1 we notice that at every decision state (until features from the k the pattern are exhausted) a feature from the pattern under consideration (k th one) is preferable to one from the neighbouring (k-1 st) pattern. However, when the features are not all equally effective, as in Figures 7.2 and 7.3, we observe that at some states the error probability is minimized by selecting the additional feature from the neighbouring pattern rather than the pattern to be decided. Thus, when the features are of unequal quality, having the choice to select an additional feature from either

the pattern under consideration or one of the neighbouring patterns may mean an earlier termination of the feature observation process. This in turn results in a reduction in the required number of features per pattern for terminal decisions.

The total number of states in which a feature from the neighbouring pattern is preferable over a feature from the pattern under consideration, denoted by $s(n_{k-1})$, depends on the distribution of the effectiveness of the various feature components. The standard deviation of the features' variance σ_f was found to be higher in Case 7.3 ($\sigma_f = 0.287$) than in Case 7.2 ($\sigma_f = 0.188$) and as a result we notice from Figures 7.2 and 7.4 that the value of $s(n_{k-1})$ is higher in Case 7.3 than in Case 7.2. Plotted in Figure 7.4 is the $s(n_{k-1})$ for several values of standard deviation, σ_f of the features' variance. The means of the features' variance were the same in each case. As shown in Figure 7.4, the higher is σ_f, that is, the more the feature components differ from each other in their effectiveness, the larger is the value of $s(n_{k-1})$. This indicates that in situations where the features are quite dissimilar in their effectiveness, the CSPR schemes would be more efficient than the classical sequential and nonsequential schemes.

7.4 Experimental Evaluation of CSPR Schemes Using Multiclass Problems

In order to assess the performance improvements obtainable using the three CSPR schemes discussed in Section 6.4, a set of recognition experiments were simulated on the digital computer. The experiments and their implications are discussed below.

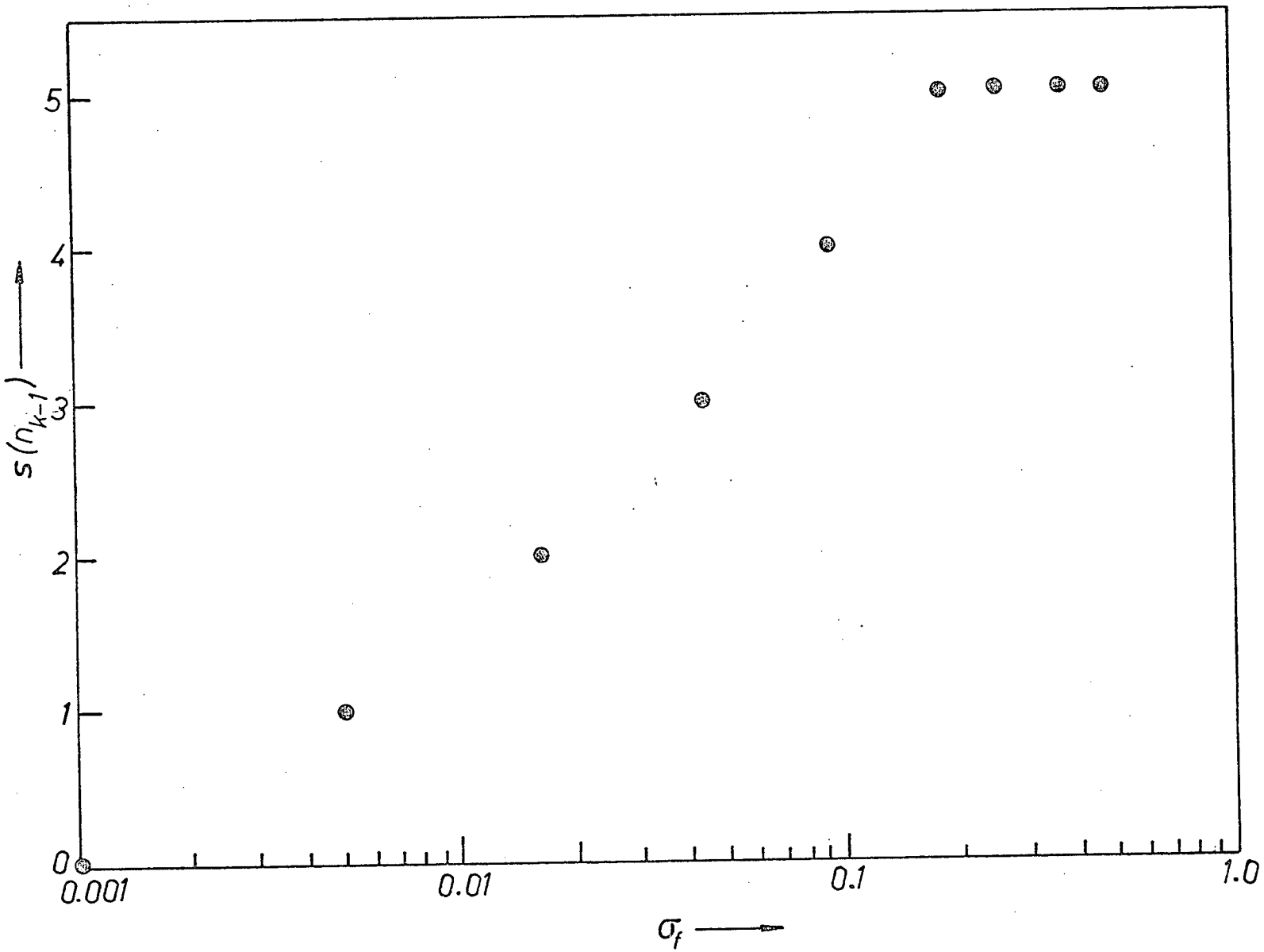

Figure 7.4 Effect of standard deviation of the features' variance on the total number of states in which a feature from the previous pattern is preferable to one from the present pattern.

Experiment 7.1:

The CSPRS Type 1 with $t = 2$ was considered. The cost of observing any one of the feature components from a pattern sample was assumed to be the same but dependent on the position of the pattern on which the feature was observed. A feature, if observed on the pattern under consideration was assumed to be less costly than if it was observed on one of the neighbouring patterns. At instant k the total cost of a feature component from the j th pattern was

$$c(n_j + 1) + \kappa(k - j)\cdot c(n_j + 1), \; j = (k - t + 1), \ldots, k \qquad (7.10)$$

where κ was a prespecified constant and $c(n_j + 1)$ was the cost of a feature component from the pattern under consideration.

The performance of CSPRS Type 1 is illustrated in Figure 7.5. The results of all seven texts are shown in Table 7.1. For comparison purposes the performance of simple sequential, simple nonsequential, compound sequential and compound nonsequential decision schemes are also illustrated in Figure 7.5. The figure indicates that the CSPRS Type 1 requires significantly fewer features (on the average) per pattern than equally reliable simple sequential and simple nonsequential decision schemes. Using the same dependence relations, the CSPRS Type 1 also performs significantly better than the compound nonsequential scheme. At 20 percent error rate, for example, the CSPRS Type 1 requires about 21.0 percent fewer features on the average per pattern than the compound nonsequential scheme. Compared to a compound sequential decision scheme, the CSPRS Type 1 using the same bigram dependence relation among the pattern classes performs better, although the percentage improvement in performance is not large. At 20.0 percent error rate, a 4.6 percent

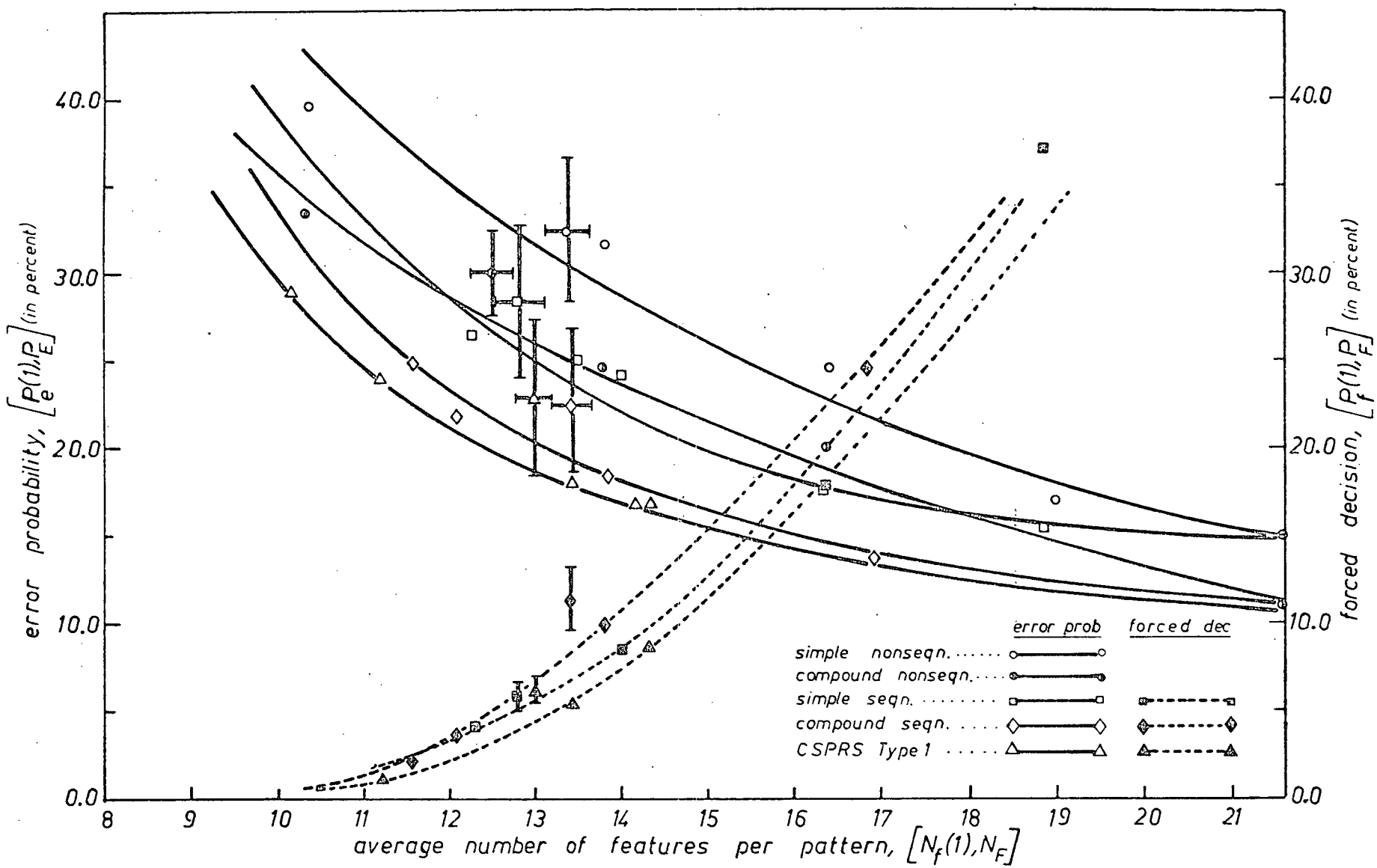

Figure 7.5 Comparison of CSPRS Type 1, compound sequential, compound nonsequential, simple sequential, and simple nonsequential decision schemes.

Text		Set of Alphabets Used for		Avg. No. of Features Pattern	Error Prob. (in percent)	Avg. No. of Features Observed (in percent)		Forced Decision (in percent)
No.	Size	Training	Testing			On k th Pattern	On (k-1) st Pattern	
1	538	22-147	1-21	13.5	18.2	71.9	28.1	5.6
2	420	1-41 43-147	22-42	13.2	30.1	73.6	26.4	6.2
3	579	1-42 64-147	43-63	12.9	20.3	73.4	26.6	5.7
4	516	1-63 85-147	64-84	12.9	18.5	75.8	24.2	3.9
5	467	1-84 106-147	85-105	13.3	22.0	72.6	27.4	7.3
6	557	1-105 127-147	106-126	13.0	30.1	72.5	27.5	7.7
7	489	1-126	127-147	12.8	20.8	73.8	26.2	6.7
Expected Values				13.1	22.8	73.4	26.6	6.2
Standard Deviations				0.23	4.73	1.18	1.18	1.17

Table 7.1 Recognition results of all seven texts using CSPRS Type 1

saving in the average number of features per pattern occurs for the Type 1 relative to the compound sequential scheme. Table 7.1 shows that the CSPRS Type 1 requiring 13.08 features on the average per pattern, observed 26.6 percent of the total used features on the previous pattern and the remaining 73.4 percent on the pattern under consideration. The number of features observed from the previous patterns is, in general, dependent on the parameter κ of (7.10). The more expensive it was to observe a feature on the previous patterns the less frequently were such features observed.

Experiment 7.2:

The CSPRS Type 2 with $t = 2$ was implemented. The feature components were assumed equally costly irrespective of the position of the pattern on which they were observed.

The recognition results based on the first text appear in Figure 7.6. Also included in Figure 7.6 are the results of CSPRS Type 1 and the sequential compound decision scheme. Figure 7.6 shows that the CSPRS Type 2 requires slightly more features on the average than the equally reliable CSPRS Type 1. Compared to the compound sequential decision scheme, the CSPRS Type 2 performed better as long as the average number of features required per pattern was below 75 percent of the total available features. About 18 to 20 percent of the total features used were found to be observed from the previous pattern sample and the rest from the pattern under consideration.

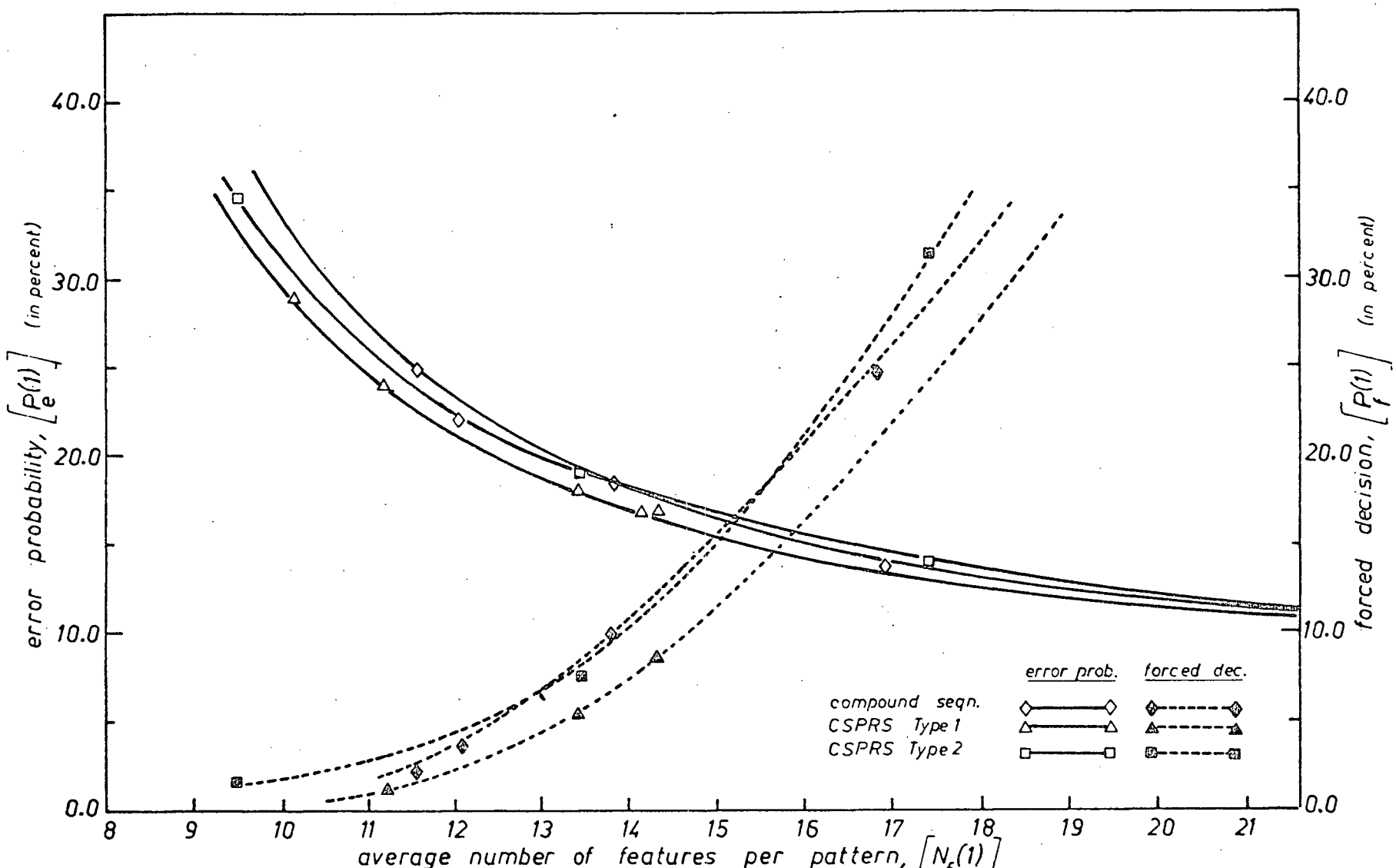

Figure 7.6 Comparison of CSPRS Type 1, CSPRS Type 2, compound sequential, and compound nonsequential decision schemes.

Experiment 7.3:

The CSPRS Type 3 with t = 3 was considered. Thus, additional information from the following as well as the previous patterns was available for the classification of the present one. The features were all equally costly as in Experiment 7.2.

The recognition results based on the first text are presented in Figure 7.7. Also included in Figure 7.7 are the results of CSPRS Type 1, Type 2 and the ordinary compound sequential decision scheme. It is evident from Figure 7.7 that the CSPRS Type 3 performs better than the CSPRS Type 1, Type 2 and the compound sequential decision scheme. However, the percentage improvement in performance relative to CSPRS Type 1 is not large. As it is pointed out in Section 6.4, the CSPRS Type 3 permits decisions to be delayed until some information from the subsequent patterns are available. The CSPR scheme requiring on the average 7.5 features per pattern was found to observe 60.5 percent and 16.7 percent of the total used features on the subsequent and the previous patterns respectively. The feature requirement from the previous and the subsequent patterns can be controlled by making the cost of features dependent on the pattern's position as in Experiment 7.1.

In Figure 7.8 the recognition capability of each feature component is illustrated. Two sets of results are presented, one analytical and the other experimental. The analytical results were obtained using the Bayes expression for the probability of error as given by (8.8). The seven likelihood matricies obtained from the training samples provided the knowledge of the probability distributions of the feature components. The experimental one is based on the actual recognition of the seven English texts, each feature component being used separately.

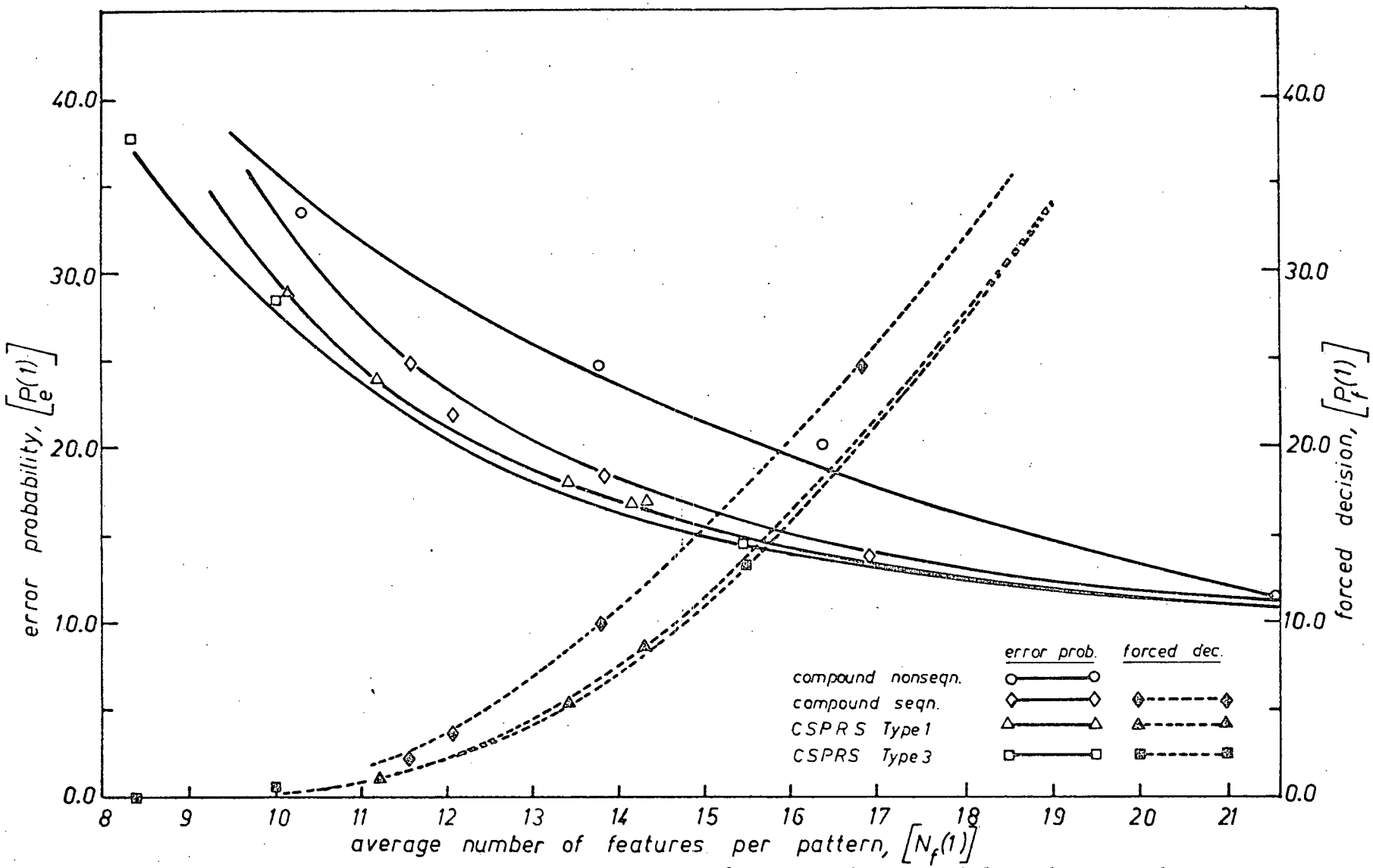

Figure 7.7 Comparison of CSPRS Type 1, CSPRS Type 3, compound sequential, and compound nonsequential decision schemes.

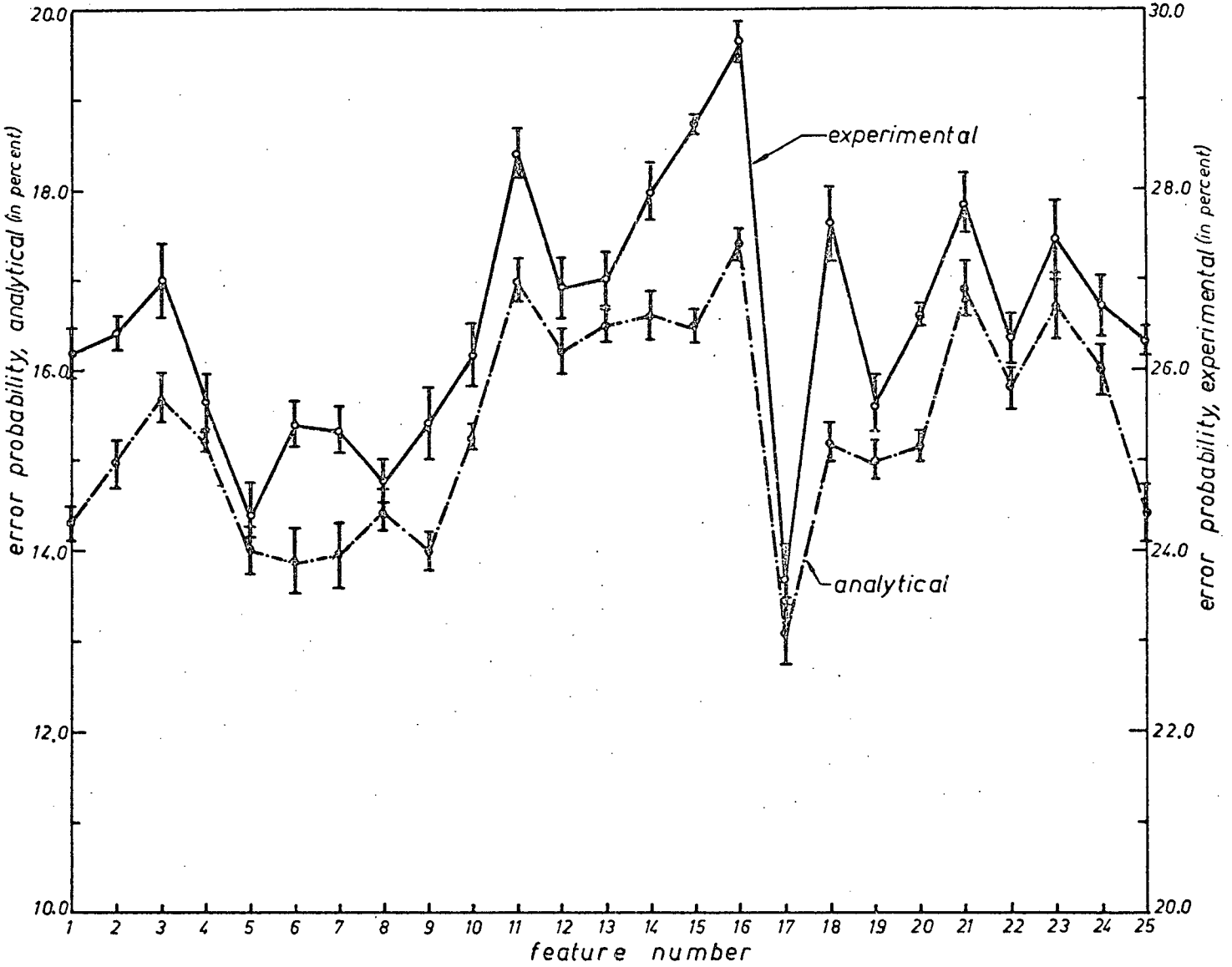

Figure 7.8 Analytical and experimental evaluation of the discrimination capability of the 25 feature components.

The a priori probabilities of the pattern classes in both cases equalled those of the characters of English text. The discrepancy between the analytical and the experimental results as observed in Figure 7.8, is due to the fact that in the experimental procedure the blanks and the periods were included and assumed to be perfectly recognizable. We notice from the figure that the majority of the 25 feature components have a recognition capability within a narrow range of 14 to 17 percent when evaluated analytically and 25 to 28 percent when evaluated experimentally. Therefore, the 25 features used in simulating the CSPR schemes are more or less of equal effectiveness as far as the recognition capability is concerned.

If the variance of the features' effectiveness is low, very little or nothing is gained by observing the additional feature on the neighbouring patterns rather than the pattern under consideration. It therefore, intuitively appears that in a situation like this where the features are equally effective the CSPR schemes may not perform any better than the ordinary compound sequential decision scheme. We arrived at a similar conclusion by analyzing the results of the two-class problem in Section 7.3. That the three CSPR schemes did not perform significantly better than the ordinary sequential decision scheme is most likely due to the similar effectiveness of the features. It is anticipated that the more dissimilar the feature components are in their effectiveness the more efficient and useful the CSPR schemes would be as seuqential decision schemes for pattern recognition problems with dependent hypotheses.

7.5 The CSPR Schemes and Recognition Systems with Variable Dimensional Feature Vectors

For recognition systems with variable or relatively low dimensional feature vectors, the CSPR schemes may be quite attractive as sequential classifiers. It is possible that a pattern sample having a low dimensional feature vector can not be classified with sufficient confidence even after observing the entire set of features from the pattern. One way to increase the confidence in such a situation is to observe additional features from the neighbouring patterns, as in the CSPR schemes. Table 7.2 and Figure 7.9 illustrates the relation between the system error probability, and the number of features observed on the pattern under consideration and the neighbouring patterns. We note from the table and the figure that it is possible to maintain the feature requirement from the pattern under consideration more or less constant and still reduce the error probability of the system by observing additional features from the neighbouring patterns. With $t = 2$, for example, the error probability was reduced from 24.1 percent to 17.1 percent by observing on the average only 3.4 more features from the previous pattern, while the number of features from the pattern under consideration was maintained approximately constant. Thus, by selecting a suitable value for the average number of features from the pattern under consideration (depending on how low the dimensionality of the feature vectors goes), the feature requirement from the neighbouring patterns can be adjusted until desired accuracy is achieved.

CSPRS TYPE - 1

t = 2, Bigram Statistics, Unordered Features

Expt. No.	Text		Av. No. of Feats. Observed on the		Error Probability (in percent)	Forced Decision (in percent)
	No.	Size (chars.)	k th Pat.	(k-1)st Pat.		
1	1	538	9.9	1.3	24.1	1.3
2	1	538	9.6	3.8	18.2	5.6
3	1	538	9.7	4.5	16.9	8.9
4	1	538	9.6	4.7	17.1	8.7

t = 3, Bigram Statistics, Unordered Features

Expt. No.	Text		Av. No. of Feats. Observed on the			Error Probability (in percent)	Forced Decision (in percent)
	No.	Size (chars.)	k th Pat.	(k-1)st Pat.	(k-2)nd Pat.		
1	1	538	9.2	2.2	1.2	19.5	2.2

Table 7.2 Performance of CSPRS Type 1 for various amounts of features from the neighbouring patterns.

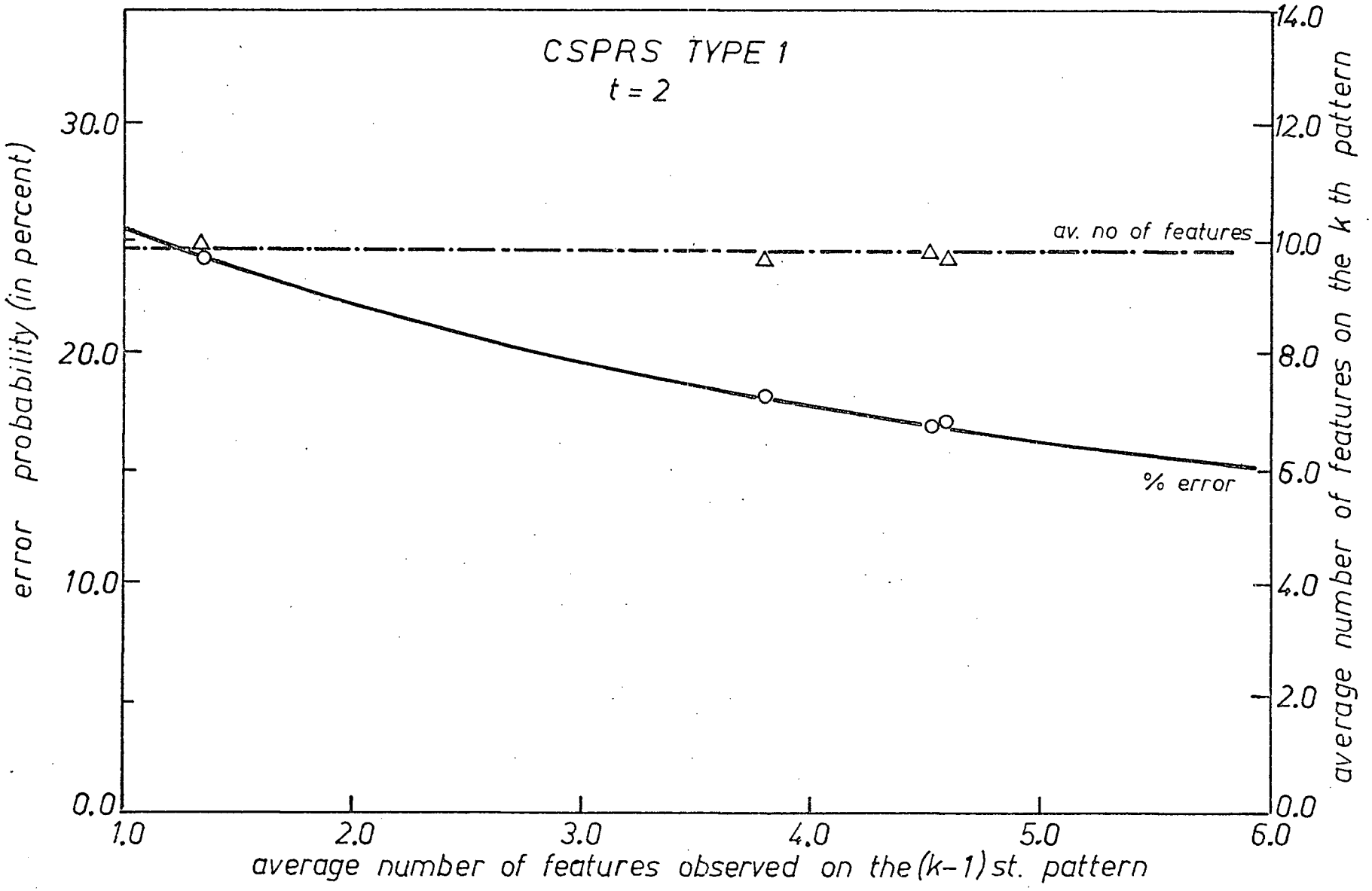

Figure 7.9 Effect of observing features from the previous pattern, on error probability and the feature requirement from the present pattern in CSPRS Type 1.

CHAPTER VIII

FEATURE ORDERING AND ON-LINE SEQUENTIAL DECISION SCHEMES

8.1 Introduction

The expected cost of a terminal decision in sequential decision schemes is strongly dependent on the order in which the features are observed. The sequential schemes discussed so far in the thesis are always provided with a preordered set of features. Thus, at every state the feature to be observed is known in advance. It is, however, possible to formulate a sequential decision scheme which at every decision state would select the feature which is best suited for that state [25], [26]. Therefore, a feature which together with the already observed features on the pattern sample provides the maximum amount of additional information is observed at every state. Such an on-line ordering of features results in an earlier termination of the feature observation process.

Unfortunately, the implementation of the optimal on-line sequential decision scheme* (based on dynamic programming) is impracticable in most situations due to the amount of computation and storage involved. One amenable solution is to use the suboptimal decision rule based on one-state ahead truncation approximation. However, the question is how effective is the suboptimal on-line sequential decision scheme both from the points of view of computational difficulty and recognition performance. Cardillo and Fu's [65] results of simulated recognition experiments even

*From now on the sequential decision scheme with on-line feature ordering is called the on-line sequential decision and the sequential decision scheme without any feature ordering is called usual sequential decision scheme.

though they are meant to determine the effect of one-state ahead truncation approximation, do not enable one to significantly evaluate the suitability of the suboptimal on-line sequential decision scheme. More over, as it is pointed out in Chapter V, the oversimplicity of the recognition problem simulated and the experimental methodology used, the results are of very limited value.

In this chapter the performance of the suboptimal on-line sequential decision scheme is evaluated on the basis of a set of significant experiments simulated on a digital computer. Its computational complexity is analyzed and the situations in which the on-line ordering of features may be useful are discussed. The capability of the on-line sequential decision scheme as a feature ordering criterion is assessed and compared with those of three other feature evaluation criteria. A modified on-line sequential (MOLS) decision scheme based on limited length (LS) of search over available features is proposed. The MOLS decision scheme requires less computation than the on-line sequential decision scheme and performs better than the usual sequential decision scheme. The advantage of incorporating limited length of search into the sequential decision schemes using a set of preordered features is also pointed out.

8.2 Formulation of the Sequential Decision Scheme with On-Line Ordering of Features

At any decision state i, let $f_j(i)$ denote the j th component feature of a set of D features available in their natural order. Let

$$\Pi_b(q) = \{i, j, \ldots, \ell\}, \; i, j, \ldots, \ell = 1, 2, \ldots, D;$$
$$i \neq j \neq \ldots, \neq n; \; b = 1, 2, \ldots, (q)^M$$

be a set of q indicies of the features $f_i(1)$, $f_j(2)$, ..., $f_\ell(q)$, observed at 1, 2, ..., q th states respectively. Therefore, $\overline{\Pi}_b(q)$, the complement of the set $\Pi_b(q)$ is the set of indicies of the features yet to be observed at state q. Let:

$R_s[X^k;\ n_k;\ \delta^k|\Pi_b(n_k)]$ be the risk involved in forming an optimal terminal decision δ^k at state n_k of the decision process having available the sequence of features $(x_1^k, \ldots, x_{n_k}^k)$ selected in such a way that the feature indicies form the set $\Pi_b(n_k)$.

$R_c[X^k;\ n_k|\Pi_b(n_k)]$ be the minimum expected risk involved at state n_k of continuing the feature observation process having available the sequence of features $(x_1^k, \ldots, x_{n_k}^k)$ from the k th pattern and selected in such a way that the feature indicies form the set $\Pi_b(n_k)$.

If at state n_k it is decided to terminate the feature observation process, then the risk involved in optimum (Bayes) decision δ^k is given by

$$R_s[X^k;n_k;\delta^k|\Pi_b(n_k)] = \underset{i}{\mathrm{Min}}\{\sum_{j=1}^{M} \ell(\omega_j^k;d_i^k)P(X^k|\omega_j^k;\Pi_b(n_k))P(\omega_j^k)\},$$
$$i = 1, 2, \ldots, M. \qquad (8.1)$$

If the elements of the loss-matrix is given by (2.5), then the optimum decision $\delta^k = d_q^k$, if

$$P(X^k|\omega_q^k,\Pi_b(n_k))P(\omega_q^k) = \underset{i}{\mathrm{Max}}\{P(X^k|\omega_i^k;\Pi_b(n_k))P(\omega_i^k)\}.$$

If however, an additional feature is desired at state n_k, then the risk involved in continuing the observation process can be expressed as follows:

$$R_c[X^k;n_k|\Pi_b(n_k)] = C_{n_k} + \underset{q\varepsilon\bar{\Pi}_b(n_k)}{\text{Min}} \{c(n_k+1) + \int_{\substack{\Omega \\ x^k_{n_k+1}}} \underset{i}{\text{Min}}[\sum_{j=1}^{M} \ell(\omega_j^k;d_i^k) \tag{8.2}$$

$$P(X^k,x^k_{n_k+1},f_q(n_k+1)|\omega_j^k)P(\omega_j^k)] \cdot P(x^k_{n_k+1};f_q(n_k+1)|X^k,\Pi_b(n_k))dx^k_{n_k+1}\},$$

$$i = 1, 2, \ldots, M; \; n_k = 1, 2, \ldots, D-1.$$

The decision rule for the on-line sequential decision scheme based on the one-state ahead truncation approximation is:

$$\left[\begin{array}{ll} R_c[X^k;n_k|\Pi_b(n_k)] < R_s[X^k;n_k;\delta^k|\Pi_b(n_k)]: & \text{CONTINUE} \\ R_c[X^k;n_k|\Pi_b(n_k)] > R_s[X^k;n_k;\delta^k|\Pi_b(n_k)]: & \text{STOP}. \end{array}\right. \tag{8.3}$$

If the observation process is decided to be continued to the (n_k+1) st state then the best feature to be observed is any solution to (8.2). In other words, for the (n_k+1) st state

$$\hat{x} = f_q(n_k+1); \text{ if } c(n_k+1) + \int_{\substack{\Omega \\ x^k_{n_k+1}}} \underset{i}{\text{Min}}[\sum_{j=1}^{M} \ell(\omega_i^k;d_j^k)P(X^k,x^k_{n_k+1},f_q(n_k+1)|\omega_j^k;\Pi_b(n_k))$$

$$P(\omega_j^k)] \cdot P(x^k_{n_k+1},f_q(n_k+1)|X^k;\Pi_b(n_k))\, dx^k_{n_k+1}$$

$$= \underset{s\varepsilon\bar{\Pi}_b(n_k)}{\text{Min}} \{c(n_k+1) + \int_{\substack{\Omega \\ x^k_{n_k+1}}} \underset{i}{\text{Min}}[\sum_{j=1}^{M} \ell(\omega_j^k;d_i^k)$$

$$P(X^k, x^k_{n_k+1}, f_s(n_k+1) | \omega^k_j; \Pi_b(n_k)) P(\omega^k_j)] P(x^k_{n_k+1}; f_s(n_q+1) | X^k, \Pi_b(n_k)) dx^k_{n_k+1}\},$$

$$q \varepsilon \bar{\Pi}_b(n_k); \; i = 1, 2, \ldots, M.$$

In case only a prespecified feature component is available for observation at state n_k, the risks defined in (8.1) and (8.2) reduce to those of the usual sequential decision scheme of Section 5.5.

8.3 Results and Discussion of Recognition Experiments Using the On-Line Sequential Decision Scheme

The simple suboptimal on-line sequential decision scheme of Section (8.2) was simulated on the digital computer using the data set described in Chapter IV. The feature components were assumed equally costly and also statistically independent within each class. The performance of the decision scheme for various values of average number of features per pattern is shown in Figure 8.1. The results of recognition of all seven texts are presented in Table 8.1 and the expected results with corresponding standard deviations are also shown in Figure 8.1. For the ease of comparison the performance of the simple usual sequential and the optimal nonsequential decision schemes also appear in Figure 8.1.

It is evident from Figure 8.1 that if all the features available from a pattern sample are not observed, the on-line sequential decision scheme is definitely more effective than the usual sequential decision scheme from the point of view of expected cost of a terminal decision. At 17.5 percent error rate, for example, the on-line sequential scheme requires 44.6 percent and 33.6 percent fewer features on the average per pattern than the optimal nonsequential and suboptimal usual sequential

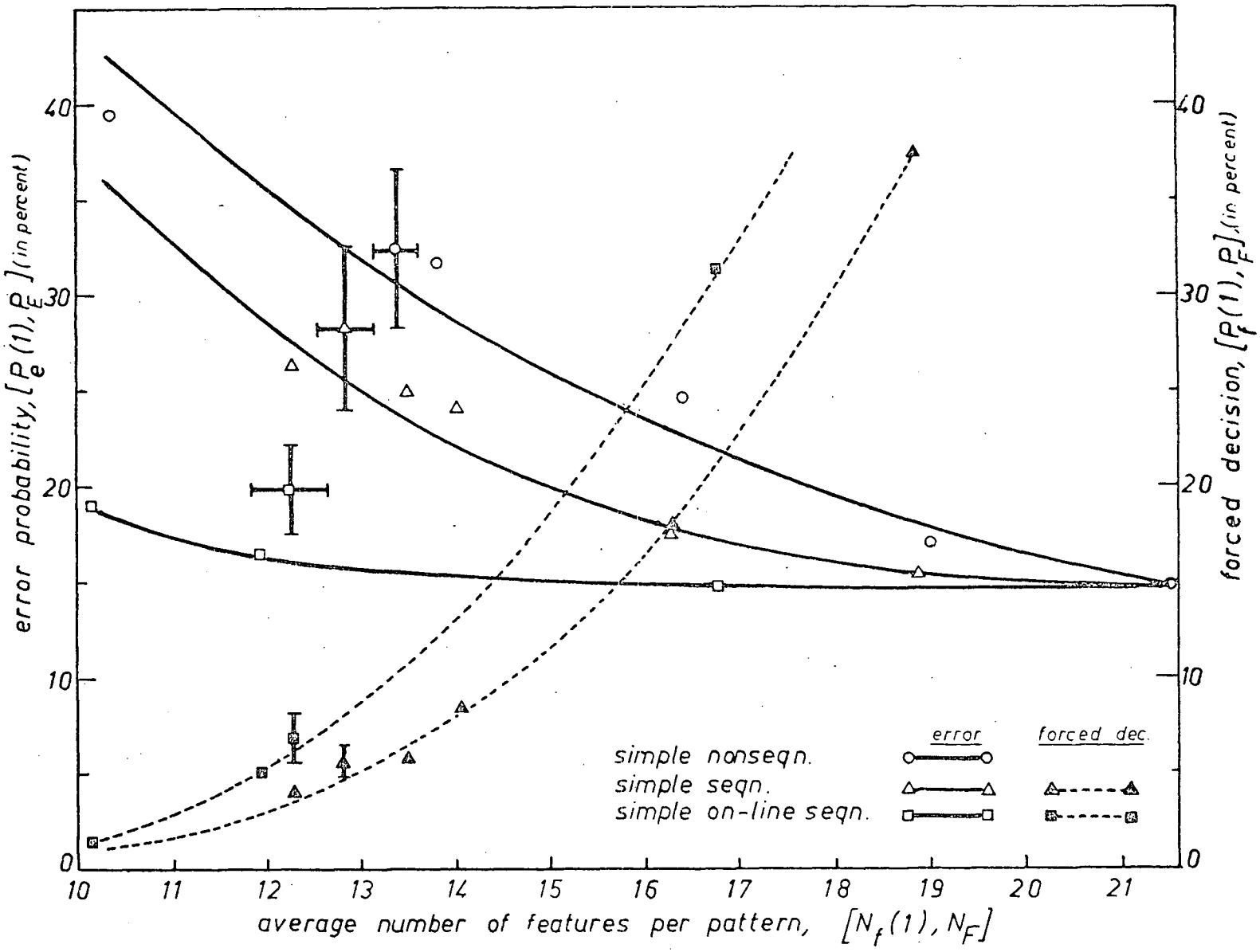

Figure 8.1 Comparison of suboptimal on-line sequential, usual sequential, and optimal nonsequential decision schemes.

Text		Alphabet Set Used For		Avg. No. of Features Pattern	Error Prob. (in percent)	Forced Decision (in percent)
NO.	Size	Training	Testing			
1	538	22-147	1-21	11.9	16.5	5.2
2	420	1-21 43-147	22-42	11.8	22.1	6.2
3	579	1-42 64-147	43-63	12.0	19.5	6.2
4	516	1-63 85-147	64-84	12.0	18.4	6.8
5	467	1-84 106-147	85-105	12.7	18.2	7.7
6	557	1-105 106-147	106-126	13.0	24.4	9.7
7	489	1-126	127-143	12.1	20.2	6.96
Expected Values				12.2	19.9	6.9
Standard Deviations				0.42	2.45	1.32

Table 8.1 Recognition results of all seven texts using suboptimal simple sequential decision scheme with on-line ordering of features.

decision schemes, respectively. At the same error rate, the usual sequential decision scheme requires only 14.7 percent fewer features than the optimal nonsequential decision scheme. Thus, the on-line sequential decision scheme based on the one-state ahead truncation approximation may be quite attractive for recognition problems where a trade-off between the cost of feature observation and reliability of the system is desired.

The only disadvantage with the on-line sequential decision scheme is that at every decision state it has to search for the best feature and therefore, requires additional computation. It is also necessary to remember the features observed at various states but the storage and computation involved for such process is negligible. The computational complexity due to the involved search is of primary concern. At any state n_k, the decision scheme needs to test $(D-n_k)$ feature components to determine the feature to be observed, if necessary, at (n_k+1)st state. If the on-line sequential decision scheme requires on the average ℓ features per pattern for terminal decisions, then the average amount of searches $N(\ell)$, involved per pattern is

$$N(\ell) = \ell(2D - \ell + 1)/2, \quad \ell = 1, 2, \ldots, D. \tag{8.4}$$

For $\ell = D$, that is, if all the D features are needed per pattern, (8.4) reduces to

$$N(D) = D(D + 1)/2. \tag{8.5}$$

Considering (8.5) as the reference, we can express the average computational complexity due to the on-line ordering of features as follows:

$$S_c(\alpha) = \alpha(2\alpha - D\alpha + 1)/(D + 1), \tag{8.6}$$

where $\alpha = \ell/D$.

In Figure 8.2 $S_c(\alpha)$ for $D = 25$ is plotted as a function of α.

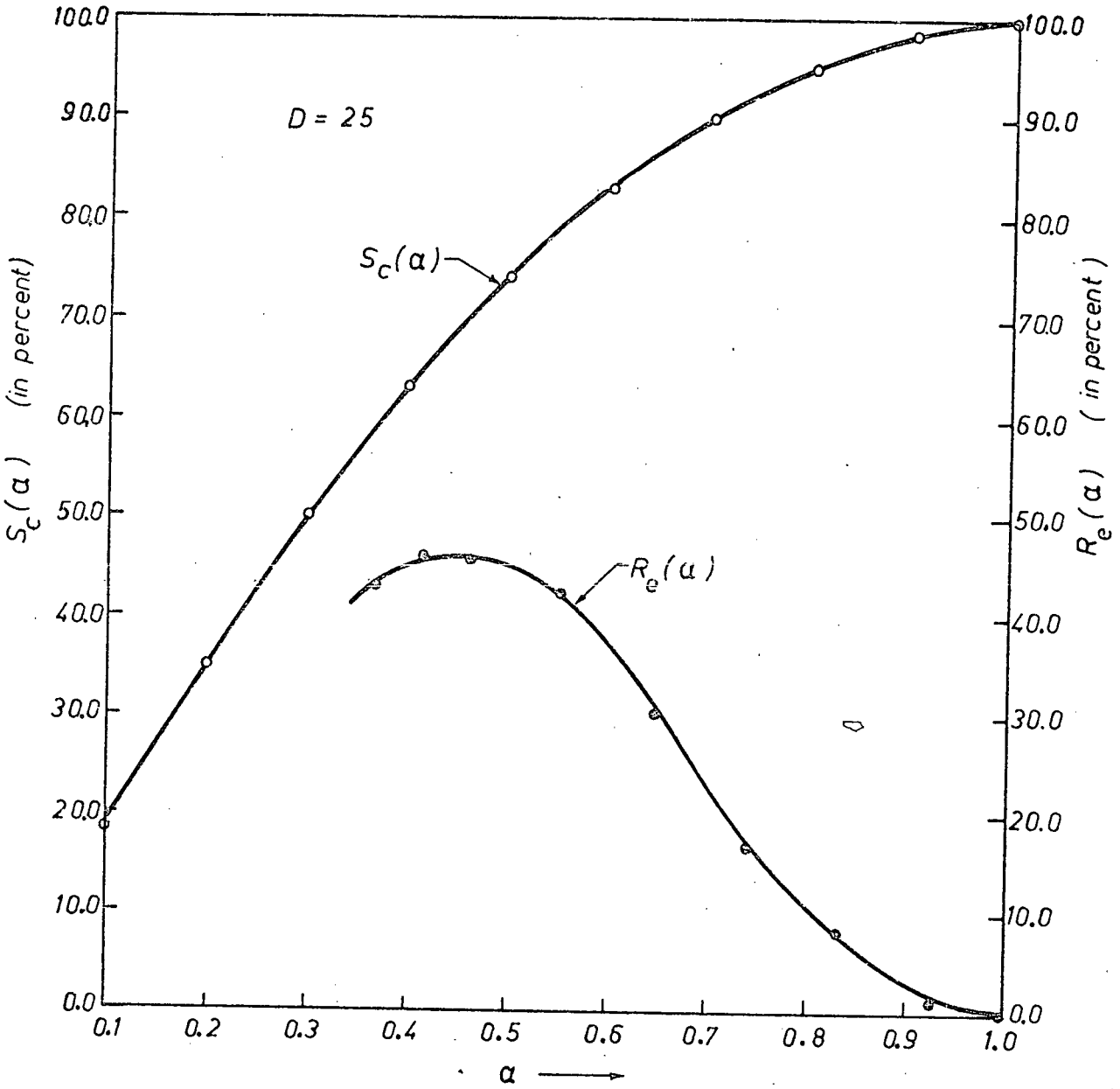

Figure 8.2 Comparison of computational complexity and the reduction in the error probability due to on-line ordering of features.

Also shown in the Figure 8.2 is $R_e(\alpha)$ which denotes the reduction in the error probability obtainable using the on-line sequential decision scheme relative to the usual sequential decision scheme. Formally, $R_e(\alpha)$ is defined as follows:

$$R_e(\alpha) = (P_e - P'_e)/P_e$$

where for some α

P_e denotes the error probability due to the simple suboptimal decision scheme without feature ordering.

and P'_e denotes the error probability due to the on-line sequential decision scheme.

Figure 8.2 shows that for any increase in value of α beyond 0.45 the computational complexity $S_c(\alpha)$ increases and the reduction in the error probability $R_e(\alpha)$ decreases. A range of values of α therefore, exists for which the on-line ordering of features is most effective from the points of view of computational difficulty and proficiency in reducing system error probability. Figure 8.2 indicates that the on-line ordering of features is suitable when on the average around 45 percent of the total available features per pattern is adequate for terminal decisions. The optimum range in practice however, would depend on the cost of additional computation and the expected cost of a terminal decision.

8.4 Feature Evaluation Using the On-Line Sequential Decision Scheme

Selection of a "good" set of features and classification of patterns based on the selected set of features are the two essential aspects of a pattern recognition problem. The size of such a feature set is dictated by the data storage and processing limitations and its

objective lies in the minimization of the system error probability. In general the minimization of error probability for the selection of a set of features for multiclass problems is often impossible. An expression for the system error probability is not always available, and even if it may be available, the expression may be too cumbersome for analytical minimization. For these reasons various suboptimal (optimal in some cases) approaches from the information theory point of view which are easier to implement and have bound on the error probability have been proposed for feature selection [71] - [79]. Nevertheless, the computational part can still be tedious, and impractical in some situations. To select the best subset of n features from a set of D, one has to search over $\binom{D}{n}$ combinations, except in some special cases where the criterion function for a subset of features is expressible in terms of the function of the individual components.

One simple technique is to use an <u>information measure</u> to individually rank each feature component and eliminate those with low rankings. This technique is by far the simplest and has been used by others for the selection of a subset of features [77], [80]-[82]. Implicit in this technique is the assumption that a subset of features containing the ones which are individually ranked to be good by some feature evaluation criterion, is a good choice. The validity of such assumption is however, highly questionable. In fact it has been shown theoretically that the best subset may even be composed of the features which are worst when ranked on an individual basis [83], [84].

Ordering of features, as we have mentioned earlier, is important for sequential decision schemes. The expected cost of terminal decisions is highly dependent on the order in which the features are observed. The

advantage of the sequential decision scheme with on-line feature ordering is that it is capable of selecting the best features by itself during the classification process. But on-line ordering scheme may not always be desirable, in which case we must decide on the ordering in which a usual sequential decision scheme should observe the features. Several approaches for this purpose are considered in this section.

The frequency of usage of a particular feature component in an on-line sequential decision scheme is dependent on the effectiveness of that feature. Therefore, the on-line sequential decision scheme can be considered as a scheme for the evaluation and the ordering of features. The ordered set of features may be useful for sequential decision schemes with a preordered set of features and the selection of subsets of features for nonsequential decision schemes. Notice that the evaluation of features using the on-line sequential decision scheme is not on the basis of individual performance, because the selection of a feature at every state is dependent on the features already observed in previous states.

Table 8.2 shows the usage of the 25 feature components at various decision states of the sequential classification process. The results are based on the recognition of all seven texts using a simple on-line sequential decision scheme. On the average 12.25 features per pattern sample were observed. The preference for one feature component over others at different decision states is evident from Table 8.2 which indicates that a ranking of the features is possible using the on-line sequential decision scheme.

Let U_{ij} denote the total usage of the j th feature component at i th state and let $\hat{f}_j(i)$ represent that the j th (in natural order)

decision state \ total usage of feature components (in percent)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	14.3	0.0	62.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	22.0	0.0	0.0	0.0	0.0	0.0	0.0	1.4	0.0
3	0.0	10.6	0.0	4.6	4.2	0.0	17.5	6.3	0.0	0.0	1.7	12.6	3.1	1.4	0.0	7.1	18.8	0.0	0.0	0.0	0.0	9.5	0.0	2.6	0.0
4	0.0	9.7	0.2	9.2	6.9	0.0	5.3	9.8	0.0	0.0	3.0	7.9	5.3	4.6	0.1	8.9	14.0	0.0	0.8	0.0	0.0	7.4	0.0	6.9	0.0
5	1.3	7.5	1.1	9.2	6.4	0.0	5.4	7.0	0.5	0.2	3.9	5.7	6.5	10.9	1.1	5.9	5.8	0.6	2.9	0.1	0.0	6.9	0.8	8.5	0.0
6	1.5	7.1	2.7	7.8	7.4	0.3	2.0	6.4	1.5	0.7	4.0	5.3	7.8	10.7	2.1	5.0	3.8	1.7	4.1	0.9	0.0	8.7	1.7	5.2	0.2
7	0.8	7.2	2.8	8.6	6.2	0.3	1.6	5.3	3.0	1.0	3.7	5.2	5.6	7.4	4.3	5.6	3.3	2.7	5.6	0.7	0.0	7.0	2.1	6.0	0.3
8	1.1	6.4	3.8	7.4	6.9	0.4	0.5	4.9	3.5	1.6	5.1	3.6	5.7	6.1	4.2	6.5	3.0	3.4	4.4	1.2	0.0	5.7	2.4	4.3	0.4
9	0.7	6.0	4.4	5.5	6.6	1.0	0.3	4.2	3.6	2.3	4.5	4.0	6.5	4.4	4.2	4.8	1.8	3.4	4.9	1.4	0.0	5.1	2.4	4.0	0.9
10	0.4	5.4	4.1	4.7	4.9	1.2	0.4	3.5	3.7	2.3	5.0	3.4	4.0	4.4	4.5	3.7	2.1	3.8	4.8	1.9	0.0	3.5	3.7	3.6	0.6
11	0.4	3.2	4.3	3.4	4.3	1.1	0.3	2.7	3.7	2.7	5.2	3.3	3.8	3.6	3.4	5.2	1.3	4.0	3.8	1.4	0.0	3.6	3.4	3.6	0.9
12	0.4	2.4	3.7	3.1	3.2	1.5	0.0	3.5	4.0	2.5	5.1	2.6	4.4	3.0	3.4	3.1	0.9	3.1	3.7	2.0	0.0	1.9	4.0	3.0	0.7
13	0.6	1.2	3.9	2.6	2.3	1.4	0.3	2.8	2.9	3.8	4.3	1.8	4.3	2.6	3.8	3.5	0.9	2.9	3.6	1.7	0.0	2.1	2.1	3.4	0.7
14	0.6	1.5	4.3	2.1	1.5	0.9	0.2	3.6	2.2	4.4	2.9	1.3	3.7	1.4	3.1	2.6	0.7	2.0	3.2	1.6	0.0	1.8	2.5	2.5	0.9
15	0.5	1.1	3.1	2.0	1.7	1.4	0.0	3.4	2.8	3.9	2.1	2.2	2.3	1.0	1.8	2.9	0.5	1.5	2.4	2.0	0.0	2.1	2.5	1.9	0.7
16	0.7	1.0	2.7	0.8	0.9	1.3	0.0	2.1	2.3	3.8	1.8	2.1	1.7	1.4	2.5	1.9	0.2	2.5	2.0	1.3	0.0	1.1	2.8	2.0	1.1
17	0.8	1.0	1.6	1.0	0.9	1.6	0.0	2.6	2.8	2.5	1.6	1.4	1.4	0.6	1.6	1.5	0.5	2.2	1.5	1.5	0.0	1.3	1.8	2.0	0.9
18	0.1	1.3	1.2	0.3	0.5	2.5	0.0	1.1	2.5	1.8	0.7	2.3	0.9	0.5	1.9	0.9	0.3	2.1	0.9	3.1	0.0	1.0	2.0	1.3	1.0
19	0.6	1.5	1.0	0.2	0.3	2.4	0.0	0.7	1.7	1.1	0.7	1.6	0.4	0.6	1.0	0.7	0.1	1.7	1.2	2.0	0.0	1.2	2.8	1.1	1.8
20	0.9	1.2	1.4	0.3	0.3	2.2	0.0	0.6	1.3	0.8	0.3	1.0	0.3	0.5	1.1	0.3	0.0	1.3	1.2	2.2	0.0	1.2	1.3	0.9	1.6
21	1.0	0.7	0.6	0.3	0.1	3.0	0.0	0.6	0.6	1.0	0.3	0.5	0.2	0.2	0.9	0.5	0.2	1.1	0.8	1.7	0.0	0.7	1.3	0.5	2.1
22	1.3	0.7	0.6	0.2	0.1	1.8	0.0	0.4	0.4	0.9	0.4	0.3	0.1	0.3	0.5	0.1	0.1	0.9	0.7	1.6	0.0	0.4	1.0	0.5	2.6
23	1.3	0.1	0.4	0.5	0.0	2.1	0.0	0.2	0.4	0.6	0.2	0.2	0.1	0.1	0.6	0.1	0.1	0.6	0.3	1.3	0.0	0.4	1.1	0.3	2.0
24	2.3	0.0	0.7	0.2	0.0	1.0	0.0	0.0	0.2	0.7	0.2	0.0	0.0	0.1	0.7	0.0	0.0	0.4	0.0	0.6	0.0	0.1	1.4	0.1	1.7
25	4.0	0.0	0.2	0.1	0.0	0.4	0.0	0.0	0.0	0.4	0.1	0.0	0.2	0.1	0.3	0.0	0.0	0.1	0.0	0.7	0.0	0.0	0.8	0.0	0.8

Table 8.2 The usage of the feature components at various decision states of an on-line sequential decision scheme. The row and column corresponding to a encircled number indicate the feature number and its rank respectively.

feature is the the i th best. A method of ranking the features based on Table 8.2 is to decide $f_j(i)$

$$\text{if:} \quad \left[\begin{array}{l} \sum_{\ell=1}^{q} U_{\ell j} \neq \underset{g}{\text{Max}}\{ \sum_{\ell=1}^{q} U_{\ell g} \}, \quad \text{for all } i > q \geqslant 1 \\ \\ \text{and } \sum_{\ell=1}^{i} U_{\ell j} = \underset{g}{\text{Max}}\{ \sum_{\ell=1}^{i} U_{\ell g} \}, \end{array} \right. \tag{8.7}$$

$$i, j, g, \ell = 1, 2, \ldots, D.$$

The ordering of the features based on (8.7) is shown in Table 8.2 where the column and the row corresponding to a circled number represent the feature number and its rank respectively. We shall refer to this procedure of ordering the features as the <u>on-line ordering procedure no. 1</u>.

Figure 8.3 shows the frequency of use of the 25 feature components employing an on-line sequential decision scheme, where the frequency refers to the ratio of number of pattern samples in which the feature component was used to the total number of pattern samples classified. The results are again based on the recognition of all seven texts and using 12.25 features per pattern on the average. Note that some features were used much more often than the others depending on their effectiveness. For example, feature number 21 was observed on each and every pattern sample classified, whereas feature number 25 was observed only on 20 percent of the total pattern samples. Thus, a ranking of the feature components on the basis of the frequency of usage of the features is possible in the on-line sequential decision scheme. We shall refer to this procedure of ordering the features as the <u>on-line ordering procedure no. 2.</u>

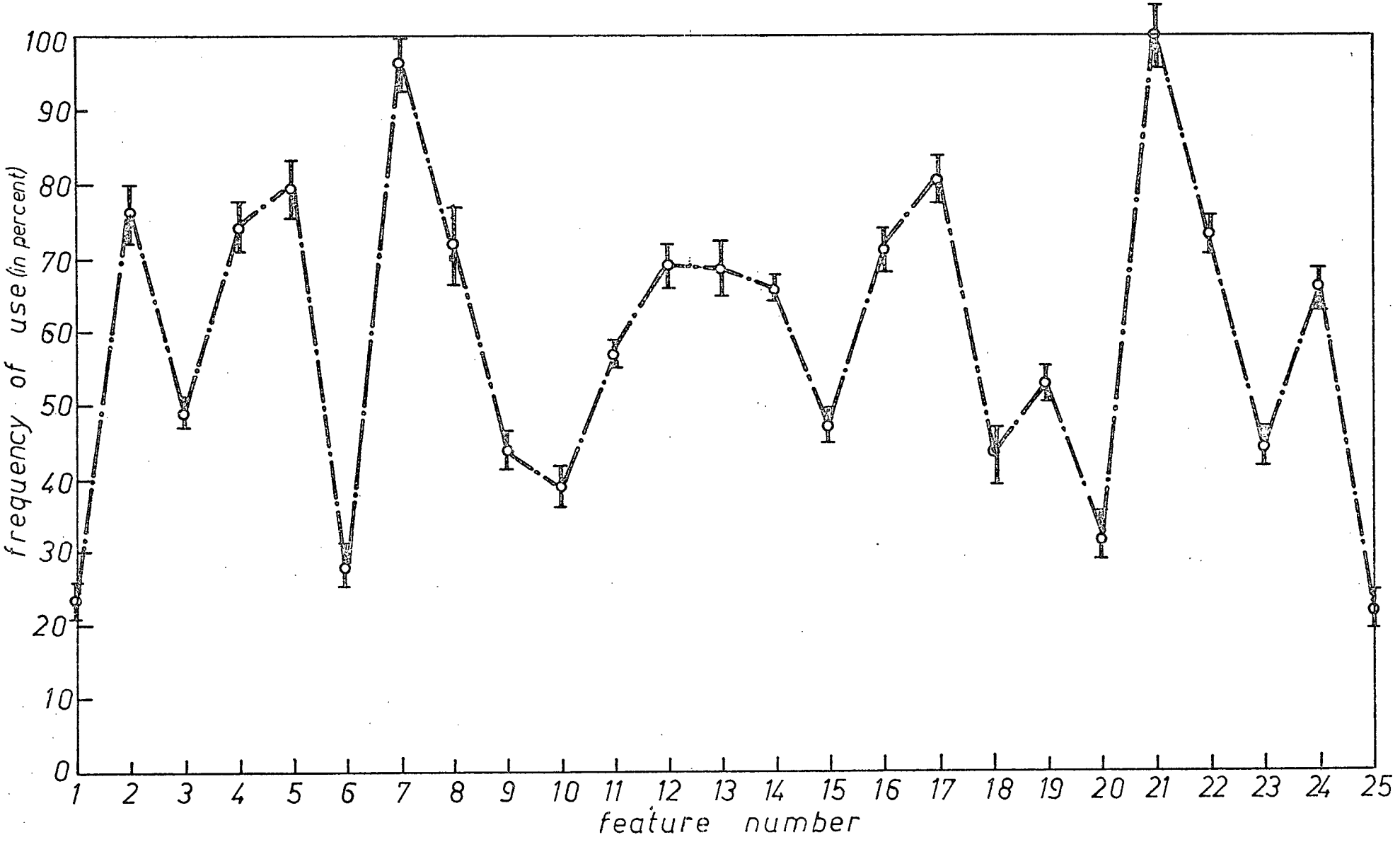

Figure 8.3 Frequency of usage of the feature components in on-line sequential decision scheme.

In order to assess the capability of the on-line sequential decision scheme as a feature ordering criterion, three other standard feature evaluation criteria were also considered for the purpose of comparisons. The features were evaluated individually. Thus, the subset to be selected was assumed to contain only one feature. The three criteria are as follows:

1. Bayes Error Probability: for statistically independent features within each class, the Bayes error probability for the i th feature component is given by

$$p_i(\varepsilon) = 1 - \int_{\Omega_{x_i}} \operatorname*{Max}_j \{P(x_i|\omega_j)P(\omega_j)\}\, dx_i, \tag{8.8}$$

$$i = 1, 2, \ldots, D;\ j = 1, 2, \ldots, M.$$

The the i th feature component is considered to be more effective than the ℓ th one, if

$$p_i(\varepsilon) < p_\ell(\varepsilon), \quad i, \ell = 1, 2, \ldots, D.$$

2. The Mutual Information Between the Features and the Pattern Classes: for statistically independent features within each class the mutual information between the i th features and the set of pattern classes Ω_ω is given by

$$I[x_i;\Omega_\omega] = H(\Omega_\omega) - H(\Omega_\omega|x_i) \tag{8.9}$$

where

$$H(\Omega_\omega) = -\sum_{j=1}^{M} P(\omega_j)\ \ell n\ P(\omega_j)$$

and the equivocation

$$H(\Omega_\omega|x_i) = -\sum_{j=1}^{M} P(\omega_j) \int_{\Omega_{x_i}} P(x_i|\omega_j) \cdot$$

$$\ln\left[\frac{P(x_i|\omega_j)P(\omega_j)}{P(x_i)}\right] dx_i;\ i = 1, 2, \ldots, D.$$

Again the i th component feature is considered to be more effective than the ℓ th one when

$$I[x_i;\ \Omega_\omega] > I[x_\ell;\ \Omega_\omega],\ i, \ell = 1, 2, \ldots, D.$$

3. <u>The Expected Divergence Between the Class-Conditional Densities</u>: the expression for the expected divergence for the ℓ th feature component is given by

$$J(x_\ell) = \sum_{i=1}^{M} \sum_{j=1}^{M} P(\omega_i)P(\omega_j) \int_{\Omega_x} \{P(x_\ell|\omega_i) - P(x_\ell|\omega_j)\}.$$

$$\ln\{P(x_\ell|\omega_i)/P(x_\ell|\omega_j)\},\ \ell = 1, 2, \ldots, D. \qquad (8.10)$$

The advantage of this criterion is that for statistically independent features within each class, the expected divergence for a subset of features is expressable as a sum of expected divergences of the individual feature components. Thus, if $X = (x_1, x_2 \ldots x_{n_k})$ is a set of n_k statistically independent features, then

$$J(X) = \sum_{i=1}^{n_k} J(x_i),\ n_k = 1, 2, \ldots, D.$$

A subset of features, containing the ones which are individually ranked to be good by expected divergence is, therefore, the best subset.

Spearman rank correlation coefficients, which provide a non-parametric quantitative measure of similarity between a pair of orderings [85], [86], were computed to compare the orderings of features obtainable using on-line sequential decision scheme and the three feature evaluation criteria. Table 8.3 shows the correlation coefficients between various pairs of orderings.

It is evident from Table 8.3 that the on-line ordering procedures no. 1 and no. 2 result in almost the same ordering of the feature components. The orderings are however quite different from the ones obtainable using the three other measures. The closest to the on-line ordering procedures is the expected divergence. The correlation coefficients are 0.231 and 0.201 between the expected divergence and the on-line ordering procedure no. 1, and between the expected divergence and the on-line ordering procedure no. 2 respectively.

In order to assess the suitability of the various feature ordering procedures for sequential decision schemes without on-line ordering, a group of recognition experiments were conducted. The simple sequential classifier was presented with features preordered with on-line ordering procedures as well as the three other feature evaluation criteria. The results of recognition based on the first text are presented in Figure 8.4. The results of all seven texts are included in Table 8.4 the expected values and the corresponding standard deviations are also shown in Figure 8.4. The results show that the sequential decision scheme performs better when the features are presented after being preordered using the on-line ordering procedures than when natural ordering is used. Procedure no. 1 seems to be more effective than procedure no. 2 as far as the performance of the sequential decision scheme is concerned.

Natural Ordering	On-Line Ordering No. 2	Ordering with Mutual Information	Ordering with expected divergence	On-Line Ordering No. 1	
-0.402	0.048	0.811	0.781	0.063	Ordered with Error Probability
	0.111	-0.245	-0.223	0.095	Natural Ordering
		-0.148	-0.201	0.953	On-line Ordering No. 2
			0.994	-0.179	Ordered with Mutual Information
				-0.232	Ordered with Expected Divergence

Table 8.9 Spearman rank correlation coefficients between various sets of ordered features.

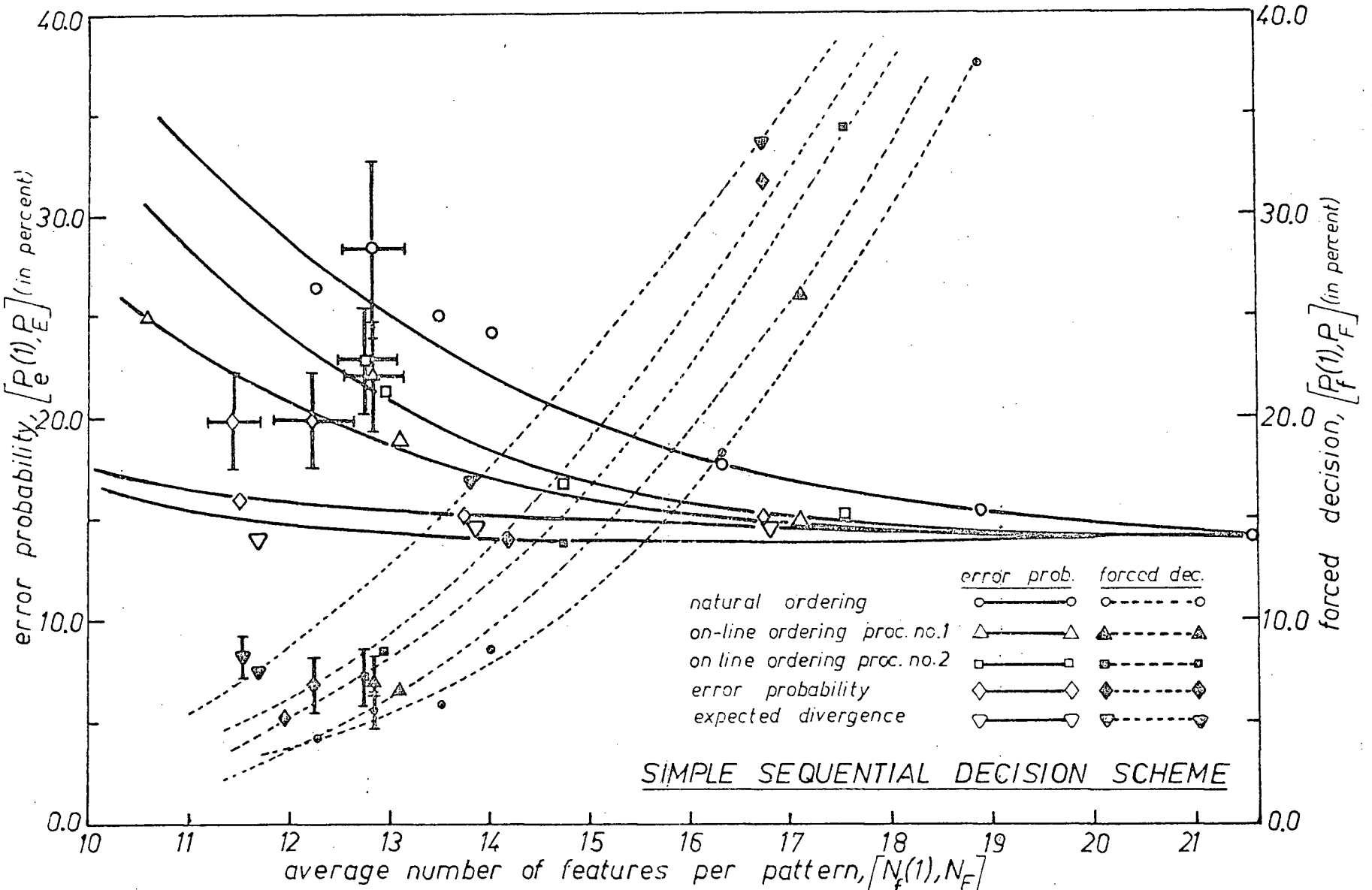

Figure 8.4 Comparison of feature ordering criteria on the basis of the performance of the simple suboptimal sequential decision scheme.

Text No.	Feature Ordering Criterion											
	On-line Proce. No. 1			On-line Proce. No. 2			Error Probability			Expected Divergence		
	Av. No. of feats/Pat.	Error Prob. %	Forced dec. %	Av. no. of feats/Pat.	Error Prob. %	Forced dec. %	Av. No. of feats/Pat.	Error Prob. %	Forced dec. %	Av. No. of feats/Pat.	Error Prob. %	Forced dec. %
1	12.9	19.1	6.7	12.9	21.3	8.5	11.5	15.8	7.2	11.6	14.3	7.0
2	13.1	23.7	7.8	2.8	23.3	8.5	11.8	21.4	9.5	11.9	20.2	9.2
3	12.4	22.0	5.3	12.3	23.1	4.3	11.3	18.8	6.9	11.2	19.1	7.2
4	12.4	20.5	6.4	12.5	20.3	7.3	11.1	19.1	6.0	10.9	16.4	6.6
5	12.7	20.0	5.8	13.0	20.5	6.4	11.5	20.5	9.0	11.8	18.6	8.7
6.	13.2	26.8	6.8	13.2	28.7	7.5	11.7	24.2	7.7	12.0	24.2	9.5
7	12.7	21.6	4.3	12.6	23.1	4.1	11.1	18.6	4.9	10.8	19.4	6.1
Mean	12.8	22.0	6.1	12.8	22.9	7.3	11.4	19.8	6.7	11.5	18.9	7.7
Std. Dev.	0.29	2.43	1.05	0.28	2.63	1.48	0.26	2.43	1.71	0.428	2.86	1.25

Table 8.4 Recognition results of all seven texts using suboptimal simple sequential decision scheme with sets of preordered features.

We note that the sequential decision scheme performs better when presented with features preordered using either error probability or expected divergence rather than the on-line procedures. The error probability and the expected divergence therefore, seem to be better criteria to obtain an ordering of features for sequential decision schemes without on-line ordering of features. From the computational point of view the error probability and expected divergence are also better methods than the on-line procedures.

8.5 Modified On-Line Sequential Decision Scheme with Limited Length of Search

The principal disadvantage of the on-line sequential decision scheme is its computational complexity. In order to be able to determine the best feature at any state ℓ, a search over $(D-\ell)$ feature components is required and this can impose a serious limitation on the desirability of the on-line sequential scheme, especially when D is large. One good compromise would be to permit the on-line selection of features but only on the basis of a limited search over the available features. Thus, at any decision state ℓ, instead of searching over $(D-\ell)$ features, the scheme may be allowed to search over q features only, where q is a prespecified constant and $D \geqslant q \geqslant 1$. The choice of q would depend on the costs of computation, feature components and terminal decision errors.

The risk of continuing the observation process, involved with modified on-line sequential (MOLS) decision scheme based on limited length of search (LS) q, can be expressed as follows:

$$R_c[X^k;n_k|\Pi_n(n_k)] = C_{n_k} + \underset{u\varepsilon\overline{\Pi}_b'(n_k)}{\text{Min}} \{c(n_k+1) + \int_{\Omega_{x_{n_k+1}}^k} \underset{i}{\text{Min}}$$

$$[\sum \ell(\omega_j^k; d_i^k) P(X^k x_{n_k+1}^k; f_u(n_k+1) \mid \omega_j^k, \Pi_b(n_k)\ ) P(\omega_i^k)]$$

$$P(x_{n_k+1}^k, f_u(n_k+1) \mid X^k; \Pi_b(n_k)\ ) dx_{n_k+1}^k\}, \qquad (8.11)$$

$$i = 1, 2, \ldots, M$$

where $\bar{\Pi}_b'(n_k)\ \varepsilon \bar{\Pi}_b(n_k)$, is the set of indicies of the first q feature components yet to be observed at decision state n_k.

For q = 1, (8.11) reduces to (5.13) and the MOLS decision scheme reduces to the usual sequential decision scheme with a set of preordered features. However, for $q = (D - n_k)$ for every state n_k, (8.11) is equivalent to (8.2) and the MOLS decision scheme reduces to the on-line sequential decision scheme of subsection 8.2. For any other value of q between these two extremes, the MOLS decision scheme requires less computation than the on-line sequential decision scheme and performs better than the usual sequential decision scheme.

In order to assess the suitability of the proposed MOLS decision scheme a set of recognition experiments were simulated. The results of recognition based on the first text are presented in Figure 8.5. Seven different lengths of search, including the cases of on-line sequential and usual sequential decision schemes, were considered. As shown in Figure 8.5 the higher the length of search, the lower is the expected cost of a terminal decision. However, considering a length of search of only 5, the recognition performance is improved by more than 50 percent of what is achieved by using the full length of search (LS = 25), as in the case of on-line sequential decision scheme. The computation involved with LS = 25 is, of course, far more than what is involved with LS = 5. By implementing the MOLS scheme, one can make limited use of the

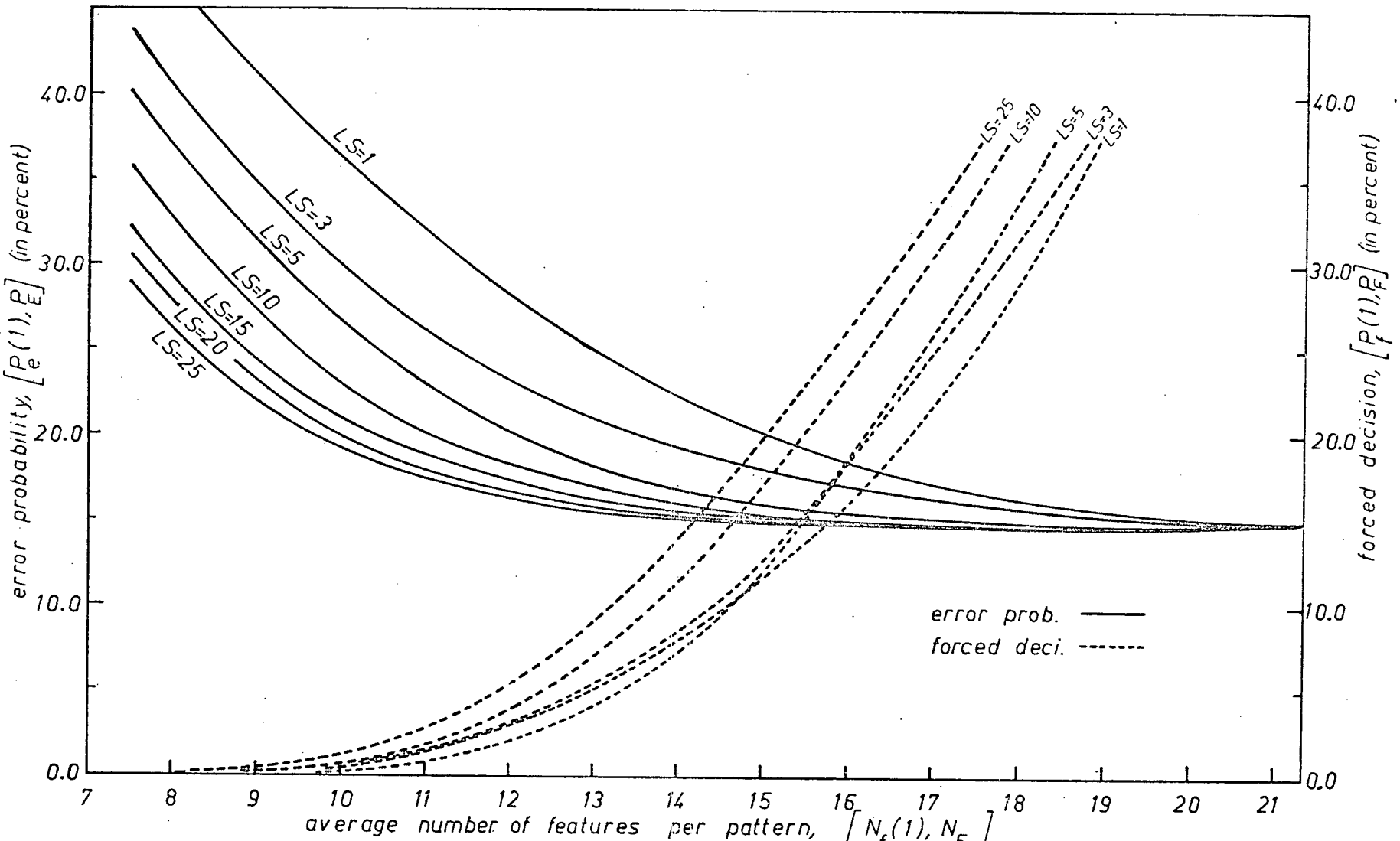

Figure 8.5 Evaluation of MOLS decision scheme with various lengths of search.

advantages of the on-line ordering of features without incuring excessive computation.

The trade-off between the computational complexity (or length of search) and the expected cost of a terminal decision of the MOLS decision scheme is analyzed as follows. Let q be the desired length of search and ℓ be the average number of features per pattern required for a terminal decision using the MOLS scheme. Therefore, the average amount of search involved with each pattern is

$$N(\ell, q) = \{q \cdot (2D - q + 1) - (D - \ell) \cdot (D - \ell + 1)\}/2. \qquad (8.12)$$

For full length of search, i.e. q = D, as in the on-line sequential decision scheme, (8.12) reduces to (8.5). Combining (8.5) and (8.13), the computational complexity for the MOLS scheme relative to the on-line sequential decision scheme may be expressed as follows:

$$S(\alpha, \beta) = \frac{(\beta^2 - 2\beta -) + (1-\alpha)(1-\alpha+\kappa)}{(\alpha^2 - 2\alpha - \kappa)} \qquad (8.12)$$

where

$$\alpha = \ell/D$$

$$\beta = q/D$$

$$\kappa = 1/D.$$

The computational complexity $S(\alpha, \beta)$ is plotted in Figure 8.6 for various values of α and β. Also shown in Figure 8.6 is $E(\alpha, \beta)$ which denotes the increase in the error probability relative to the on-line sequential decision scheme due to limited length of search. Formally, $E(\alpha, \beta)$ is defined as follows:

$$E(\alpha, \beta) = (P''_e - P'_e)/P'_e$$

where, for some values of α and β

P'_e denotes the error probability due to the on-line sequential decision scheme.

P''_e denotes the error probability due to the suboptimal MOLS decision scheme.

The trade-off between the computational complexity and expected cost of a terminal decision is clearly indicated in Figure 8.6. The question of a trade-off arises only when the average number of features per pattern for a terminal decision is less than approximately 70 percent of the number of total features available from each pattern sample. Figure 8.6 shows that for $\alpha = 0.56$, for example, the increase in the length of search from 25 ($\beta = 1$) to 10 ($\beta = 0.4$) is only 12.5 percent. However, the saving in the additional computation is about 50 percent. Thus, the MOLS decision scheme with suitable length of search allows an appropriate balance of computational complexity and expected cost of a terminal decision in sequential decision schemes with on-line ordering of features.

8.6 Sequential Recognition Scheme with Limited Length of Search Over Preordered Features

By restricting the sequential classifier to observe a suitable set of preordered features, the total expected cost of terminal decisions can be considerably reduced and at the same time the additional computation involved with on-line ordering of features can be avoided. It is however, difficult to obtain an ordering of features that would guarantee the minimum error probability in a sequential scheme. Moreover, in sequential schemes the best feature at every state is always dependent on the information already available from the pattern and can not be prespecified.

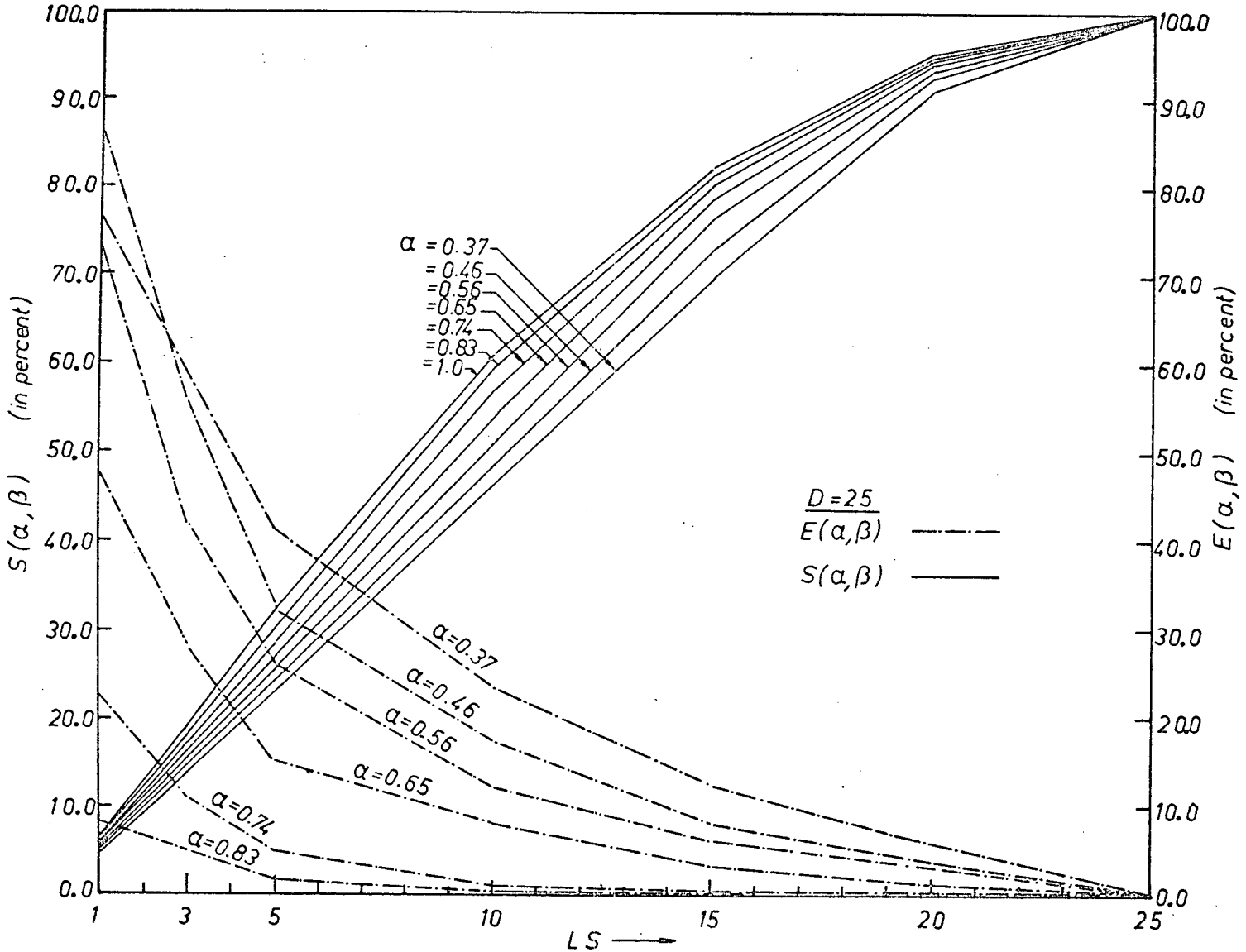

Figure 8.6 Comparison of computational complexity and the increase in the error probability due to the limited length of search in the MOLS decision scheme.

Thus, by restraining the order of observing the features one is liable to introduce redundancy in the observed features and therefore, increase the expected cost of a terminal decision. One way of avoiding such redundancy would be to provide the sequential classifier with a suitable set of preordered features and at the same time permit the on-line ordering of features on the basis of a limited length of search. Thus, the additional computation involved is fairly small but the decision scheme has the option to choose from a few feature components. The limited search facility allows the sequential decision scheme to make the best use of a set of preordered features. It is of course, obvious that if full length of search (LS = 25) is employed, no matter how the features are ordered, the computational complexity and the performance of the decision scheme would be equal to those of the on-line sequential decision scheme discussed before.

Figure 8.7 illustrates the performance of the simple suboptimal sequential decision scheme using a set of preordered features with and without the facility of the limited length of search. Three different orderings of features (shown in the figure) were considered and LS = 5 was used in each case. It is clear from Figure 8.7 that the sequential decision scheme with LS = 5 performs better in each case than the usual sequential decision scheme, observing the prespecified feature at every state. Limited amount of search incorporated into sequential decision schemes requires only a little additional computation but may mean a considerable saving in expected cost of terminal decisions. It is therefore, recommended that sequential decision schemes using a set of preordered features should always have the facility of limited length of search.

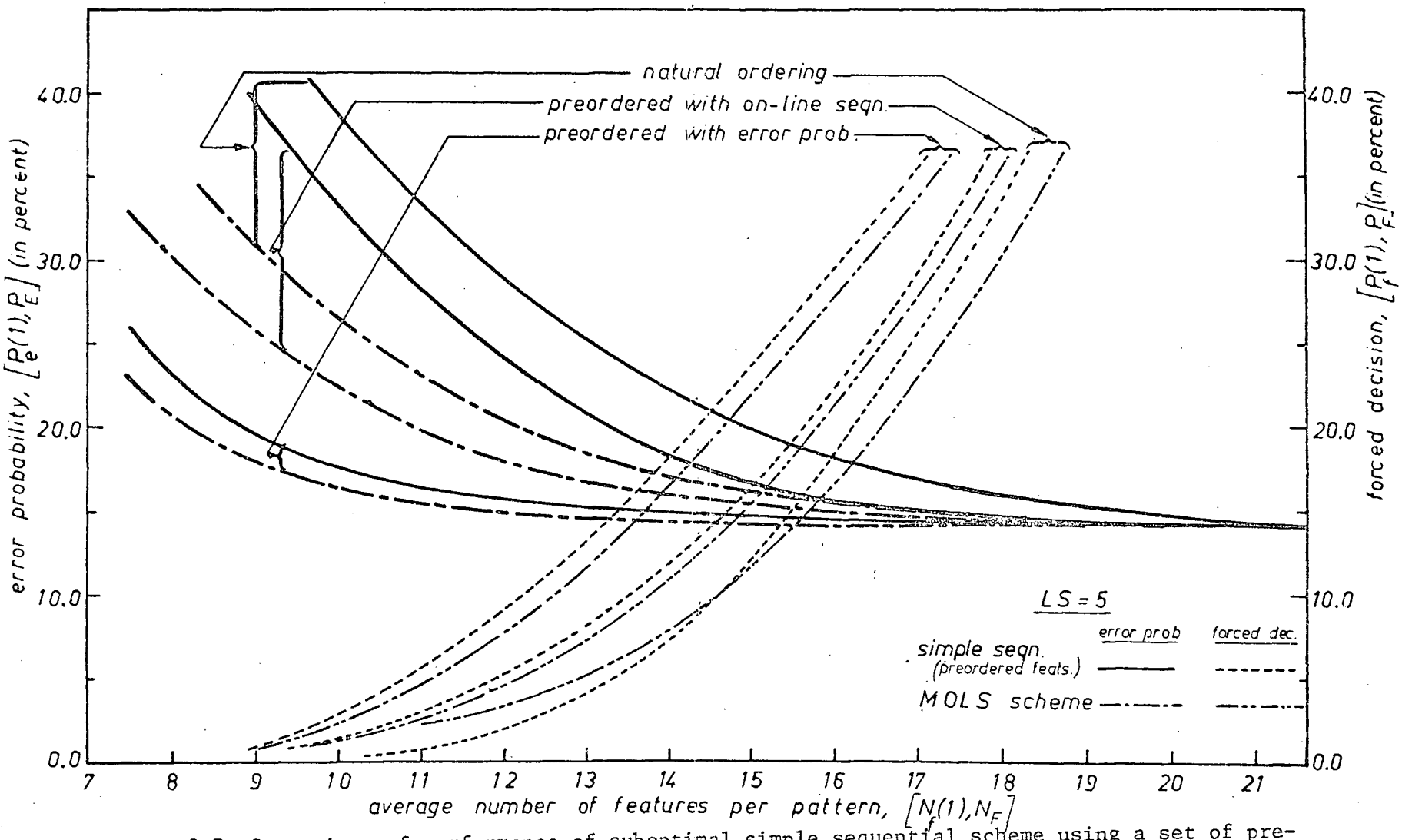

Figure 8.7 Comparison of performance of suboptimal simple sequential scheme using a set of preordered features with and without limited search facility.

CHAPTER IX

CONCLUSIONS AND DISCUSSION

9.1 General Discussion

Various sequential schemes for both pattern classification and feature ordering have been considered in this thesis. The sequential decision schemes require a fewer features on the average than the equally reliable nonsequential decision schemes. When the features are costly, either because of elaborate computations or because of danger and risk associated (biomedical and industrial applications) with the measurement operation, the sequential decision schemes become more attractive than the nonsequential ones. The sequential decision schemes may also be suitable in situations where it is necessary to minimize the time needed for a terminal decision (real time applications). The computation and storage required for the implementation of the sequential decision schemes are, of course, more than those needed for nonsequential schemes. From the computational point of view sequential decision schemes may some times be difficult or even impossible to implement. However, the sub-optimal sequential decision schemes, based on the one-state ahead truncation approximation considered in this thesis, retain the important attributes of the sequential process and are quite practicable. Utilization of contextual information in nonsequential decision schemes is quite common and it reduces the system error probability. Use of such contextual information, in sequential decision schemes is possible through the OCSPRT and the compound sequential decision scheme formulated in Chapters II and V respectively. For dependent hypotheses recognition

problems the OCSPRT and the compound sequential decision scheme result in an expected cost of terminal decisions that is lower than that of simple sequential (including SPRT) and the compound nonsequential decision schemes.

The CSPR schemes permit the observation of the required additional features either on the pattern under consideration or any one of the neighbouring patterns. In dependent hypothesis recognition problems the schemes, therefore observe only the best features from each pattern sample. As a result the expected cost of terminal decisions is lower than that of classical compound sequential and nonsequential decision schemes. The more dissimilar the feature components are in their effectiveness the more efficient the CSPR scheme tend to be in comparison with the compound sequential and compound nonsequential decision schemes. The CSPR schemes also allow the delaying of a decision in sequential schemes for additional information from the subsequent patterns to be available.

Because the CSPR schemes need to search, at every decision state, for the pattern on which the desired additional feature is to be observed, these schemes are computationally more involved than the simple and the compound sequential decision schemes of Chapter V. It may be desirable to employ some efficient computational approximation in order to reduce the complexity and the time required for a terminal decision.

The on-line ordering of features is peculiar to sequential decision schemes and results in a considerable reduction in the expected cost of terminal decision compared to usual sequential decision schemes. Although additional computation is required for on-line ordering, a compromise between such additional computation and expected cost of terminal decision is possible using the MOLS decision scheme. In MOLS

decision scheme on-line ordering of features is attainable but only on the basis of a limited length of search over the available features. Finally, it is recommended that the sequential decision scheme using a set of preordered features should always have the facility of limited length of search over the available features.

9.2 Suggestions for Future Research

Some suggestions for further work follow naturally from the present investigation. These are summarized below:

1. The derivation of the expression for the average number of features per pattern required for a terminal decision (Chapter III) is based on the assumption that at any instant k the information from the entire past is available ($t = k$). It will be of interest to obtain such an expression for any value of $k > t > 1$.
2. A wide variety of CSPR schemes is possible to be studied but only three of them have been considered in Chapter VI. Detail study of the effects of allocating costs to the feature components, and using various forms of dependence relations among the pattern classes and feature vectors is needed for further understanding of the CSPR schemes. The CSPRS Type 1 and Type 2 with larger value of t should also be considered.
3. It will be interesting to know how the CSPR scheme performs in situations where the feature components are quite dissimilar in their effectiveness. Usefulness of the CSPR schemes in recognition problems with variable dimensional feature vectors should also be determined.

4. The CSPR schemes considered in Chapters VI and VII are always provided with a set of preordered features. A possible generalization is to allow the on-line ordering of features such that the decision schemes have the choice of observing any one of the remaining features from any one of the patterns of the sequence.

5. Simulation of the sequential decision schemes formulated for dependent hypotheses recognition problems (Chapters III, V and VII) is based on the assumption that the pattern classes are first order Markov such that only bigram probabilities are used in the thesis. Higher order dependence, such as second order Markov dependence (trigram) should also be considered to determine the change in the expected cost of terminal decisions.

6. During the training phase, the likelihoods $P(x_i = j|\omega_\ell)$ $j = 0, 1, \ldots, 16$; $\ell = 1, 2, \ldots, 26$, for each feature i was estimated from a fixed number of pattern samples. In sequential schemes especially with on-line ordering, some features are observed more often than the others. Thus, instead of using a fixed number of training samples for each feature, one should use more training samples for the features which are more frequently used. An adaptive unsupervised sequential decision scheme is therefore needed where as soon as a pattern sample is classified the statistics of only the used feature components are updated. Such a sequential scheme would not only reduce the cost of terminal decisions during the testing phase but also reduce the cost of estimating the re-

quired parameters during the training phase.

7. The feature extraction scheme proposed in Chapter IV promises to be a simple but effective feature extraction method for pattern recognition and should be investigated. The effect of considering masks of a size different than 4×4 and possibly nonsquare should be determined. Use of fewer regions, and/or quantization of the number of black points in a region to a few values, would reduce required computation time and storage capacity. Effect of different weighting systems of the points of the mask should also be determined.

8. Finally, a logical extension would be to implement in practical problems the various sequential decision schemes formulated in this thesis. Possible areas of applications are speech recognition, biomedical engineering [87], [88], phased array radar, and weather prediction.

APPENDIX-A

Detailed Derivation of Equation 2.9

In the following derivations the subscripts are omitted for the sake of convenience.

For $k \geq t$, we can write

$$P(X^k|X^{k-1},\ldots,X^{k-t+1};W^k) = \frac{P(X^k,\ldots,X^{k-t+1};\omega^k,\ldots,\omega^{k-m})}{P(X^{k-1},\ldots,X^{k-t+1};\omega^k,\ldots,\omega^{k-m})}$$

$$= \frac{P(X^k,\ldots,X^{k-r};\omega^k,\ldots,\omega^{k-m}|X^{k-r},\ldots,X^{k-t+1})}{P(X^{k-1},\ldots,X^{k-r};\omega^k,\ldots,\omega^{k-m}|X^{k-r},\ldots,X^{k-t+1})} \quad \text{(A-1)}$$

For any $r \geq m$ and no delay in the system (A-1) reduces to

$$P(X^k|X^{k-1},\ldots,X^{k-t+1};W^k) = \frac{P(X^k,\ldots,X^{k-r};\omega^k,\ldots,\omega^{k-m})}{P(X^{k-1},\ldots,X^{k-r};\omega^k,\ldots,\omega^{k-m})}$$

$$= P(X^k|\bar{Y}^{k-1};\ W^k) \ , \quad \text{(A-2)}$$

where $\bar{Y}^{k-1} = (X^{k-1},\ \ldots,\ X^{k-r})$, $k \geq t$. It can similarly be shown that (A-2) holds for $k < t$. Thus (2.9-a) follows immediately.

To obtain (2.9-b), note that for any k we can write

$$P(W^k|W^{k-1};\ \bar{Y}^{k-1}) = \frac{P(\bar{Y}^{k-1}|W^k,\ W^{k-1})P(W^k;\ W^{k-1})}{P(W^{k-1};\ \bar{Y}^{k-1})} \quad \text{(A-3)}$$

for system without any delay (A-3) reduces to

$$P(W^k|W^{k-1};\ \bar{Y}^{k-1}) = \frac{P(\bar{Y}^{k-1};W^{k-1})P(W^k,\ W^{k-1})}{P(W^{k-1};\bar{Y}^{k-1})}$$

$$P(W^k|W^{k-1};\ \bar{Y}^{k-1}) = \frac{P(\bar{Y}^{k-1}|W^{k-1})P(W^k,\ W^{k-1})}{P(W^{k-1};\ \bar{Y}^{k-1})}$$

$$= \frac{P(\bar{Y}^{k-1}|W^{k-1})P(W^k|W^{k-1})P(W^{k-1})}{P(\bar{Y}^{k-1}|W^{k-1})P(W^{k-1})}$$

$$= P(W^k|W^{k-1})\ .$$

APPENDIX B

Derivation of $E[\ell n\ \gamma_{n_k}]$ and $E[\ell n\ \lambda^k]$ -- for Use in Chapter III

The expected values $E[\ell n\ \gamma_{n_k}]$ and $E[\ell n\ \lambda^k]$ are derived here. The results are analogus to Wald's results [2], [46].

At any instant k the sequential observation process is terminated when the probability ratio γ_{n_k} exceeds one of the two stopping boundaries. Neglecting any excess over the stopping boundaries, we can write

for $X^k \sim \omega_2$,

$$\gamma_{n_k} = \begin{cases} T_2 & \text{with probability } (1-\epsilon) \\ T_1 & \text{with probability } \epsilon \ . \end{cases}$$

Then

$$E[\ell n\ \gamma_{n_k}] = (1-\epsilon)\ell n T_2 + \epsilon \ell n\ T_1,\ X^k \sim \omega_2,$$

however, for $\epsilon_{12} = \epsilon_{21} = \epsilon$

$$T_2 = 1/T_1 = T = (1-\epsilon)/\epsilon$$

Thus

$$E[\ell n\ \gamma_{n_k}] = e_2(\gamma) = (1-2\epsilon)\ n(\frac{1-\epsilon}{\epsilon}),\quad X^k \sim \omega_2 .$$

One can similarly show that for $X^k \sim \omega_i$

$$E[\ell n\ \gamma_{n_k}] = e_1(\gamma) = (2\epsilon-1)\ \ell n(\frac{1-\epsilon}{\epsilon}).$$

Let

$$f_i = \ell n\ \frac{P(x_j^k|\omega_2^k)}{P(x_j^k|\omega_1)},\ j = 1, 2, \ldots, D;\ k = 1, 2, \ldots .$$

The feature vectors are assumed to be of sufficiently large dimensionality such that the observation process terminates before the final state with proability one*. For class-conditionally independent feature components, we have

$$P(X^k|\omega_j^k) = \prod_{i=1}^{D} P(x_i^k|\omega_j^k), \; j = 1, 2$$

therefore, the f_i's form statistically independent random variables.

If the process of feature observation is terminated at state n_k we obtain

$$\sum_{i=1}^{D} f_i = \sum_{j=1}^{n_k} f_j + \sum_{\ell=n_k+1}^{D} f_\ell, \; n_k = 1, 2, \ldots, D ,$$

$$= \ell n \; \lambda^k + \sum_{\ell=n_k+1}^{D} f_\ell \quad . \tag{B.1}$$

The random variable n_k is dependent only on the first n_k feature components and is independent of all the f_ℓ, $\ell > n_k$. The expected value of (B.1) therefore, yields

$$D \cdot E[f] = E[\lambda^k] + \{D - E[n_k]\} \cdot E[f]$$

or

$$E[\lambda^k] = E[n_k] \cdot E[f] \, .$$

$$= \begin{cases} e_2(n) \cdot e_2(f), & \text{if } X^k \sim \omega_2 \\ e_1(n) \cdot e_1(f), & \text{if } X^k \sim \omega_1 \, . \end{cases}$$

---

*A more rigorous streatment requires taking the limit as D approaches infinity [2].

APPENDIX-C

THE TEXTS FOR RECOGNITION EXPERIMENTS

AND THEIR ASSOCIATED STATISTICS

** TEXT NUMBER 1 **

THE PHILOSOPHICAL DIFFERENCE BETWEEN HUMAN PERCEPT
ION AND COMPUTER RECOGNITION OF PRINTED CHARACTERS
SEEMS TO DEPEND ON THE RESEARCHERS BIAS. THE COGN
ITIVE PSYCHOLOGIST MAINTAINS THAT THE DISTINGUISHI
NG CHARACTERISTICS OF THE ENTIRE STIMULUS GIVE RIS
E TO THE PERCEPT WITH NO SINGLE ATTRIBUTE BEING EI
THER NECESSARY OR SUFFICIENT IN CONTRAST THE ENGIN
EER VIEWS PATTERN RECOGNITION AS EXTRACTING VARIOU
S MATHEMATICAL MEASURES FROM THE STIMULUS BY USING
VARIOUS TRANSFORMATIONS. ALTHOUGH THESE TWO APPRO
ACHES ARE NOT NECESSARILY ANTITHETICAL

a. SOURCE..... "Character Recognition: Performance of Human Computers", by B. Blesser and E. Peper <u>M.I.T. Quarterly Progress Report</u>, No. 98, 1970, pp. 185

b. TOTAL NUMBER OF CHARACTERS..... 538

c. NUMBER OF CHARACTERS WITHOUT BLANKS AND PERIODS..... 465

d. SET OF ALPHABETS USED IN FORMING THE TEXT..... 1 - 21

** TEXT NUMBER 2 **

FROM A MATHEMATICAL STATISTICAL POINT OF VIEW THE ASSUMPTION OF RANDOM SAMPLING MAKES IT POSSIBLE TO DETERMINE THE SAMPLING DISTRIBUTION OF A PARTICUL AR STATISTIC GIVEN SOME PARTICULAR POPULATION DIST RIBUTION. IF THE VARIOUS VALUES OF THE STATISTIC C AN ARISE FROM SAMPLES HAVING UNDETERMINED OR UNKNO WN PROBABILITIES OF OCCURRENCE THEN THE STATISTICI AN HAS NO WAY TO DETERMINE THE SAMPLING DISTRIBUTI ON OF THAT STATISTIC

a. SOURCE. ..... Statistics

by W. L. Hays

Holt, Rinehart and Winston, Inc.

1963, pp. 216

b. TOTAL NUMBER OF..... 420

CHARACTERS

c. NUMBER OF CHARACTERS

WITHOUT BLANKS AND PERIODS...421

d. SET OF ALPHABETS USED IN

FORMING THE TEXT.....22 - 42

** TEXT NUMBER 3 **

IN SEARCHING FOR REGIONS OF THE CORTEX THAT MIGHT CONTRIBUTE TO THE PERFORMANCE OF THE HIGHER INTELL ECTUAL PROCESSES WE WOULD NOT BE INCLINED TO PAY T OO MUCH ATTENTION TO THOSE CORTICAL AREAS THAT HAV E ALREADY BEEN ESTABLISHED TO BE TERMINAL POINTS F OR THE PERIPHERAL NERVES AND THAT THEREFORE ARE KN OWN TO BE PRIMARILY ENGAGED IN THE RECEIPT OF SENS ORY INFORMATION OR THE TRANSMISSION OF MOTOR COMMA NDS TO THE OUTLYING REGIONS OF THE BODY. OF THE SE VERAL HUNDRED SQUARE INCHES OF SURFACE AREA IN THE CEREBRAL CORTEX ONLY ABOUT ONE FOURTH IS USED FOR THESE SENSORIMOTOR PROCESSES

a. SOURCE..... The Machinery of the Brain

by E. Wooldridge

McGraw-Hill, 1963

pp. 145

b. TOTAL NUMBER OF CHARACTERS..... 580

c. NUMBER OF CHARACTERS

WITHOUT BLANKS AND PERIODS..... 483

d. SET OF ALPHABETS USED IN

FORMING THE TEXT..... 43 - 63

** TEXT NUMBER 4 **

DEAR SIR. I HAVE REVIEWED YOUR IDEA FOR DESIGN ENT
RY SWEEP GENERATOR AND AM HAPPY TO ACCEPT IT FOR P
UBLICATION IN ELECTRONIC DESIGN. UNFORTUNATELY PUB
LICATION WILL BE DELAYED FOR A FEW MONTHS DUE TO T
HE LARGE BACKLOG OF IDEAS FOR DESIGN WE NOW HAVE O
N HAND. PRIOR TO PUBLICATION THOUGH YOU WILL RECEI
VE A COPY OF THE FINAL EDITED VERSION OF YOUR MATE
RIAL FOR APPROVAL. IN ADDITION YOU WILL RECEIVE PA
YMENT FOR YOUR ENTRY SHORTLY. YOU ARE NOW ELIGIBLE
FOR THE MOST VALUABLE OF ISSUE AWARD AS DETERMINE
D BY OUR READERS

a. SOURCE..... A Letter from the Editor of a Publishing Company

b. TOTAL NUMBER OF CHARACTERS..... 516

c. NUMBER OF CHARACTERS WITHOUT BLANKS AND PERIODS..... 421

d. SET OF ALPHABETS USED IN FORMING THE TEXT..... 64 - 84

** TEXT NUMBER 5 **

THE SHARE OF TOTAL MANUFACTURING ASSETS HELD BY TH
E TOP COMPANIES MAY BE INCREASING BUT PROBABLY NOT
NEARLY AS MUCH AS THE GOVERNMENT SUGGESTS. AND WH
AT SEEMS LIKE A DRAMATIC RISE IN SALES CONCENTRATI
ON WITHIN A MAJOR CATEGORY OF CONSUMER GOODS INDUS
TRIES VANISHES ALMOST ENTIRELY WHEN THE RATHER PRI
MITIVE STATISTICAL TECHNIQUES USED IN THE GOVERNME
NT STUDIES ARE SET ASIDE IN FAVOR OF METHODS THAT
ARE MORE SOPHISTICATED THOUGH QUITE COMMONLY EMPLO
YED BY ECONOMISTS

a. SOURCE......"Bigness is a Number Game"

by S. Rose

Fortune

vol. LXXX, No. 6, November, 1969

pp. 112-115

b. TOTAL NUMBER OF CHARACTERS..... 467

c. NUMBER OF CHARACTERS WITHOUT

BLANKS AND PERIODS..... 394

d. SET OF ALPHABETS USED IN

FORMING THE TEXT..... 85 - 105

** TEXT NUMBER 6 **

GEORGE C. SCOTT COMES AS CLOSE TO FITTING HIS DEFI
NITION OF THE IDEAL ACTOR AS ONE MAN CAN WITHOUT B
REAKING APART INTO THREE DISPARATE INDIVIDUALS. IN
HIS LIFE OFFSTAGE HE HAS BEEN STUBBORNLY EVEN VIO
LENTLY INDIVIDUAL WHEN HE IS ACTING HE CREATES A C
HARACTER AND HIDES HIS INDIVIDUALITY WITH SINGULAR
SUCCESS AS THE MAN IN ROW TEN HE IS A PERFECTIONI
ST CRITIC MORE DEMANDING OF HIMSELF THAN OF THOSE
AROUND HIM. IN MORE THAN A DOZEN STAGE AND SCREEN
ROLES IN A STEADILY GROWING CAREER SCOTT HAS DEMON
STRATED THAT HE IS ONE OF THE BEST OF CONTEMPORARY
ACTORS

a. SOURCE..... Time

March 22, 1971

pp. 51

b. TOTAL NUMBER OF CHARACTERS..... 557

c. NUMBER OF CHARACTERS WITHOUT

BLANKS AND PERIODS..... 457

d. SET OF ALPHABETS USED IN FORMING

THE TEXT..... 106 - 126

** TEXT NUMBER 7 **

THE RANGELANDS OF THE WESTERN UNITED STATES. THESE
VAST TRACTS ARE MAINLY UTILIZED FOR THE GRAZING O
F CATTLE. THEY COULD BE MUCH MORE PRODUCTIVE NOT O
NLY FROM THE ECONOMIC STANDPOINT BUT ALSO IN THE R
ECREATIONAL AND THE AESTHETIC SENSE. THE CENTURY O
LD WAR OVER THE USE OF THE WIDE OPEN SPACES OF THE
UNITED STATES IS NOW ESCALATING YEAR BY YEAR IN S
COPE AND COMPLEXITY. IN THE NINETEENTH CENTURY IT
WAS A RELATIVELY SIMPLE CONFLICT MAINLY BETWEEN TH
E CATTLE RANCHERS AND THE SHEEP RAISERS

a. SOURCE..... "The Rangeland of Western United States" by R. M. Love <u>Scientific American</u> vol. 222, No. 2, pp. 89 February, 1970

b. TOTAL NUMBER CHARACTERS..... 490

c. NUMBER OF CHARACTERS WITHOUT BLANKS AND PERIODS..... 402

d. SET OF ALPHABETS USED IN FORMING THE TEXT..... 127 - 147

REFERENCES

1. R. L. Kashyap, "Algorithm for Pattern Classification", Adaptive, Learnings, and Pattern Recognition Systems Theory and Applications, J. M. Mandle and K. S. Fu Ed., Academic Press, New York, 1970.

2. A. Wald, "Sequential Analysis", John Willey, New York, 1947.

3. D. Blackwell and M. A. Girshick, "Theory of Games and Statistical Decisions", John Willey, New York, 1954.

4. A. Wald and J. Wolfwitz, "Optimum Character of the Sequential Probability Ratio Test", Ann. Math. Statist. Vol. 19, pp. 326-339, 1948.

5. J. J. Bussgang and D. Middleton, "Optimum Sequential Detection of Signals in Noise", IRE Trans. Info. Theory, Vol. IT-1, pp. 5-19, December, 1955.

6. T. G. Birdsall and R. A. Roberts, "Theory of Signal Detectability: Deffered Decision Theory", The J. of Acoust. Soc. of Am., Vol. 37, No. 6, pp. 1064-1074, June, 1965.

7. R. A. Roberts, "Sequential Detection and Composite Hypotheses with Application to a Signal of Uncertain Amplitude", IEEE Trans. Info. Theory, Vol. IT-13, pp. 175-183, April, 1967.

8. K. S. Fu, "A Sequential Decision Model for Optimum Recognition", Biological Prototypes and Synthetic Systems, Bernard and Kare Ed., Vol. 1, Plenum Press, New York, 1962.

9. C. H. Chen, "A Study of Pattern Recognition Systems with a Sequential Learning Procedure", Ph.D. dissertation, School of Electrical Engineering, Pardue University, 1965.

10. ---------, "A Note on Sequential Decision Approach to Pattern Recognition and Machine Learning", Information and Control, Vol. 9, pp. 549-562, 1966.

11. K. S. Fu, "Sequential Methods in Pattern Recognition and Machine Learning", Academic Press, New York, 1968.

12. T. W. Anderson, "A Modification of the sequential Probability Ratio Test to Reduce the Sample Size", Ann. Math. Stat., Vol. 31, March, 1960.

13. J. J. Bussgang and M. B. Marcus, "Truncated Sequential Probability Ratio Test", IEEE Trans. Info. Theory, Vol. IT-13, pp. 512-516, July, 1965.

14. Y. T. Chien and K. S. Fu, "A Modified Sequential Machine Using Time Varying Stopping Boundaries", IEEE Trans. Info. Theory, Vol. IT-12, pp. 206-214, April, 1965.

15. E. Wong and A. Steppe, "Invarient Recognition of Geometric Shapes", Methodologies of Pattern Recognition", S. Watanabe Ed., Academic Press, New York, pp. 535-546, 1969.

16. F.C. Reed, "A Sequential Multidecision Procedure", Proc. of Symp. on Decision Theory and Appli. to Equip. Devp., Rome Air Development Center, Vol. 1, pp. 42-69, May, 1960.

17. R.A. Robert and C.T. Mullis, "A Bayes Sequential Test of m-Hypotheses", IEEE Trans. Info. Theory, Vol. IT PP. 91-94, January, 1970

18. L.C. Palmer, Sequential Test to Select Among M-Hypotheses", IEEE Trans. Info. Theory, Vol. IT-18, pp. 211-214, January, 1972.

19. A. A. Aglintsev and A. P. Ter-Saakov, "Many Alternative Sequential Analysis", Engineering Cybernetics, No. 1, pp. 160-168, 1970.

20. R. Bellman, R. Kalba, and D. Middleton, "Dynamic Programming, Sequential EStimation and Sequential Detection Processes", Proc. Nat. Acad. Sciences, Vol. 47, pp. 338-341, 1961.

21. R. Bellman, "Dynamic Programming", Princeton Univ. Press, Princeton, New Jersey.

22. K. S. Fu and G. P. Cardillo, "An Optimum Finite Sequential Procedure for Feature Selection and Pattern Classification", IEEE Trans. Auto. Control Vol. 12, pp. 588-591, October, 1967.

23. G.P. Cardillo and K. S. Fu, "A dynamic Programming Procedure for Sequentila Pattern Classification and Feature Selection", Intern. J. Math. Biosciences, Vol. 1, No. 3, pp. 463-491, 1967.

24. Y. T. Chien and K. S. Fu, "An Optimal Pattern Classification System Using Dynamic Programming", Intern. J. Math. Biosciences, Vol. 1, No. 3, pp. 439-461, 1961.

25. K. S. Fu, Y. T. Chien and G. P. Cardillo, "A Dynamic Programming Approach to Sequential Pattern Recognition", IEEE Trans. Elect. Comp., Vol. EC-16, pp. 790-803, Dec. 1967.

26. Z. J. Nikolic and K. S. Fu, "On the Selection of Features in Statistical Pattern Recognition", Proc. Ann. Princeton Conf. Inform. Sci, and Syst., pp. 434-438,Mar. 1968.

27. Y. T. Chien, "A Sequential Decision Model for Selection of Feature Subsets in Pattern Recognition", IEEE Trans. Comp. Vol. C-20, No. 3, pp. 282-290, March,1971.

28. G. E. Lasker, "Multivariate Sequential Pattern Recognition Technique", Cybernetica, Vol. XII, No. 2, pp. 67-80, 1969.

29. R. Alter, "Utilization of Contextual Constraints in Automatic Speech Recognition", IEEE Trans. Audio and Electroacoustic, Vol. AV-16, No. 1. pp. 6-11, March, 1968.

30. G. Carlson, "Technique for Replacing Character that are Garbled on Input", AFIPS Conf. Proc., Vol. 128, pp. 189-192, 1966.

31. W.W. Bledsoe and J. Browning, "Pattern Recognition and Reading by Machine", Pattern Recognition, L. Uhr Ed., pp. 301-316, Wiley, New York, 1966.

32. J. Raviv, "Decision Making in Markov Chains Applied to the Problem of Pattern Recognition", IEEE Trans. Info. Theory, Vol. IT-13, pp. 536-551, Oct. 1967.

33. K. Abend, "Compound Decision Procedures for Unknown Distributions and for Dependent States of Nature", Pattern Recognition, L. Kanal Ed., Thomson Book Co., Washington, D.C., 1968.

34. K. Abend, "Compound Decision Procedure for Pattern Recognition", Proc. National Elect. Confer., Vol. 22, pp. 777-779, 1966.

35. R. W. Chang and J. C. Hancock, "On Receiver Structure of Channel having Memory", IEEE Trans. Inform. Theory, Vol. IT-12, pp. 463-468, Oct. 1966.

36. R. O. Duda and P. E. Hart, "Experiments in the Recognition of Hand-Printed Text Part II-Context Analysis", AFIPS Conf. Proc., pp. 1139-1149, 1968.

37. A. W. Drake, "Observation of a Markov Source Through Noisy Channel", IEEE Symp. Signal Trans. and Processing, pp. 12-18, 1965.

38. R. W. Donaldson and G. T. Toussaint, "Use of Contextual Constraint in Recognition of Contour-Traced Hand-Printed Characters, IEEE Trans. Computers, Vol. C-19, No.22, pp. 1096-1099, Novem. 1970.

39. E. M. Riseman and R. W. Ehrich, "Contextual Word Recognition Using Binary Diagrams", IEEE Trans. Comp., Vol. C-20, No. 4, pp. 397-403, April 1971.

40. C. G. Hilborn, Jr. and D. G. Lainiotis, "Unsupervised Learning Minimum Risk Pattern Classification for Dependent Hypotheses and Dependent Measurements", IEEE Trans. System, Sciences and Cyber , Vol. SSC-5, pp. 109-115, April, 1969.

41. C. G. Hilborn and D. G. Lainiotis, "Optimal Unsupervised Learning Multicategory Dependent Hypotheses Pattern Recognition", IEEE Trans. Info. Theory, Vol. IT-14, pp. 468-470, May, 1968.

42. E. Samuel, "On Simple Rules for the Compound Decision Problem", Journal of the Royal Statistical Society, Series B., Vol. 27, No. 2, pp. 238-240, 1965.

43. E. Samuel, "Asymptotic Solution of the Sequential Compound Decision Problem", Ann. of Math. Statist. Vol. 34, No. 3, pp. 1079-1094, 1963.

44. A. B. S. Hussain, G. T. Toussaint and R. W. Donadlson, "Results Ob-

tained Using a Simple Character Recognition Procedure on Munson's Handprinted Data", IEEE Trans. Computers, Vol. C-21, pp. 201-205, February, 1972.

45. C. K. Chow, "An Optimum Character Recognition System Using Decision Functions", IRE Trans. Elect. Computers, Vol. 6, pp. 247-254, December 1957.

46. J. C. Hancock and P. A. Wintz, "Signal Detection Theory", McGraw-Hill, New York, 1966.

47. E. E. Beckenback and R. Bellman, "Inequalities", 2nd Ed., Springer-Verlag, New York, 1965.

48. W. H. Highleyman, "The Design and Analysis of Pattern Recognition Experiments", Bell Syst. Tech. J., pp. 723-744, March, 1962.

49. P. A. Lachenbruch and M. R. Mickey, "Estimation of Error Rates in Discriminant Analysis", Technometrics, Vol. 10, pp. 1-11, Feb., 1968.

50. L. Kanal and B. Chandasekaran, "On Dimensionality and Sample Size in Statistical Pattern Classification", Proc. Nat. Electronics Conf., pp. 2-7, 1968.

51. J. H. Munson, "Experiments in the Recognition of Hand-Printed Texts", 1968 Fall Joint Computer Conference, AFIPS Proc., Vol. 33, pt. 2, Wash. D. C., Thompson, pp. 1125-1138, 1968.

52. J. H. Munson, "The Recognition of Hand-Printed Text", Pattern Recognition, L. Kanal, Ed., Thompson Book Co., Wash. D. C., pp. 115-140, 1968.

53. A. L. Knoll, "Experiments with 'Characteristic Loci' for Recognition of Hand-Printed Characters", IEEE Trans. Comp., Vol. EC-18, pp. 366-372, April, 1969.

54. J. K. Clemens, "Optical Character Recognition for Reading Machine Applications", Ph.D. Thesis, Dept. of Elect. Engg., MIT, 1965.

55. M. D. Levine, "Feature Extraction: A Survey", Proc. of IEEE, Vol. 57, No. 8, pp. 1391-1407, August, 1969.

56. G. F. Hugehes, "On the Mean Accuracy of Statistical Pattern Recognizers", IEEE Trans. Info. Theory, Vol. IT-14, pp. 55-63, Jan., 1968.

57. B. Chandrasekaran, "Independence of Measurements and Mean Recognition Accuracy", IEEE Trans. Info. Theory, Vol. IT-17, pp. 452-456.

58. G. T. Toussaint, "Machine Recognition of Independent and Contextually Constrained Contour-Traced Hand-Printed Characters", M.A.Sc. Thesis, Dept. of Electrical Engineering, Univ. British Columbia, 1969.

59. G. T. Toussaint and R. W. Donaldson, "Algorithm for Recognizing

Contour-Traced Hand-Printed Characters", IEEE Trans. Comp., Vol. C-19, pp. 541-546, June, 1970.

60. A. B. S. Hussain, "A Sequential Test for Determining the Dependency Structure of Binary Features", submitted for Publication in IEEE Trans. System, Man, and Cyber.

61. I. J. Good, The Estimation of Probabilities, Research Monograph 20, Cambridge, Mass. MIT Press, 1965.

62. B. Russell, "Education and Social Order", London, G. Allen and Unwin Ha., 1933.

63. C. A. Dykema, "A Computer Simulation Study of the Application of Contextual Constraints to Character Recognition", M.A.Sc. Thesis, Dept. of Elect. Engg., University of British Columbia, September, 1968.

64. B. Chandrasekaran and L. Kanal, "On Linguistic, Statistical, and Mixed Models for Pattern Recognition", Technical Report OSU-CSRC-TR-71-3, Computer and Information Sciences Research Center, Ohio State University, March, 1971.

65. G. P. Cardillo and K. S. Fu, "On Suboptimal Sequential Pattern Recognition", IEEE Trans. Comp., Vol. C-17, No. 8, pp. 789-792, August, 1968.

66. B. E. Boyle, "Model for Therapeutic Diagnosis of Disease", MIT Quat. Report, QPR No. 97, pp. 158-165, April, 1969.

67. Y. L. Barabash, "On Properties of Symbol Recognition", Eng. Cybern. (USSR), pp. 71-77, September-October, 1965.

68. A. B. S. Hussain, "On the Correctness of Some Sequential Classification Schemes in Pattern Recognition", IEEE Trans. Comp., Vol. C-21, pp. 318-320, March, 1972.

69. A. B. S. Hussain and R. W. Donaldson, "Some Sequential Sampling Techniques for Multicategory Dependent Hypotheses Pattern Recognition", Proc. Fifth Hawaii Inter. Conf. on System Sciences, pp. 40-43, January, 1972.

70. A. B. S. Hussain, "Sequential Decision Schemes with On-Line Feature Ordering", Proc. Ann. Canadian Computer Conference, SESSION'72, June 1-3, 1972.

71. A. Bhattacharyya, "On a Measure of Divergence Between Two Multinomial Populations", Sankhya, Vol. 6, pp. 401-406, 1964.

72. T. Kailath, "The Divergence and Bhattacharyya Distance Measure in Signal Selection", IEEE Trans. Comm. Tech., Vol. COM-15, pp. 52-60, Febuary, 1967.

73. J. T. Tou and R. P. Heydorn, "Some Approaches to Optimum Feature

Extraction", Computer and Information Sciences II, J. T. Tou, Ed. New York, Academic Press, pp. 57-89, 1967.

74. H. F. Ryan, "The Information Content Measure as a Performance Criterion for Feature Selection", IEEE Proc. of Seventh Symposium on Adaptive Processes, Los Angles, Calif., 16-18 December, 1968.

75. P. M. Lewis, "The Characteristic Selection Problem in Recognition Systems", IRE Trans. Inform. Theory, IT-8, pp. 171-178, Feb., 1962.

76. P. J. Min, D. A. Landgrebe and K. S. Fu, "On Feature Selection in Multiclass Pattern Recognition", Proc. 2nd Ann. Princeton Confer. on Info., Sciences and Systems, IEEE 68, CT-1, pp. 453-457, 1968.

77. A. N. Mucciardi and E. E. Gose, "A Comparison of Seven Techniques for Choosing Subsets of Pattern Recognition Properties", IEEE Trans. Comp. Vol.,C-20, pp. 1023-1031, September, 1971.

78. G. D. Nelson and D. M. Levy, "A Dynamic Programming Apporach to the Selection of Pattern Features", IEEE Trans. Syst. Sci. and Cyber., Vol. SSC-4, No. 2, pp. 145-151, July 1968.

79. G. T. Toussaint and A. B. S. Hussain, "Comment on 'A Dynamic Programming Approach to the Selection of Pattern Features", IEEE Trans. Syst. Man and Cyber, Vol. SMC-1, No. 2, pp. 186, April, 1971.

80. H. V. Pipeberger and F. W. Stallman, "computation of Differential Diagnosis in Electrocardigraphy", Ann. N.Y. Acad. Sci., Vol. 115, pp. 1115-1125, July, 1964.

81. P. S. R. S. Rao, "On Selecting Variables for Pattern Classification", Tech. Report No. 11, Information Research Associates, Inc., Lexington, Massachusetts, June, 1967.

82. G. S. Sebestyen and J. Edie, "Pattern Recognition Research", Litton Systems, Waltham, Mass., Final Rep., Contract AF 19 (628)-1604 AD620236, July 1965.

83. J. D. Elashoff, R. M. Elashoff and G. E. Goldman, "On the Choice of Variables in Classification Problem with Dichotomans Variables", Biometrika, Vol. 54, pp. 668-670, 1967.

84. G. T. Toussaint, "Note on Optimal Selection of Independent Binary valued Features for Pattern Recognition", IEEE Trans. Info. Theory pp. 618, September 1971.

85. S. Siegel, "Nonparametric Statistics for the Behaviourial Sciences", McGraw-Hill, New York, 1956, Chapter 9.

86. W. L. Hays, "Statistics", Holt, Rinehart and Winston, Inc., New York, 1963, Chapter 18.

87. E. Persoon and K. S. Fu, "Medical Diagnosis Using Dynamic Sequential Pattern Recognition", Proc. Fifth Hawaii Int. Conf. on System Sciences, Suppl. Computer in Biomedicine, pp. 161-163, January, 1972.

88. K. S. Fu, and M. H. Loew, "Automatic Medical Diagnosis Using Non-parametric Sequential Classification Procedures", AGARD Avionics 21st Technical Symposium on Artificial Intelligence, May 24-28, 1971, Rome, Italy.