Reprint from

Communication and Cybernetics 10

Digital Pattern Recognition

Edited by K. S. Fu

T. M. Cover and T. J. Wagner

Topics in Statistical Pattern Recognition

Springer-Verlag Berlin Heidelberg New York

---

2. Topics in Statistical Pattern Recognition

T. M. COVER and T. J. WAGNER

Pattern recognition, from the broadest viewpoint, is the study of how one puts abstract objects or patterns into categories in a simple reliable way. We have chosen three areas to review: Nonparametric Discrimination, Finite Memory Learning, and Pattern Complexity. We feel that these statistical areas will ultimately play a role in any global pattern recognition theory which evolves.

2.1 Nonparametric Discrimination

2.1.1 Introduction

A statistician observes a random vector X with values in $\mathbb{R}^d$ and wishes to estimate its *state* $\theta \in \{1, \ldots, M\}$. For this purpose he collects *data* $(X_1, \theta_1), \ldots, (X_n, \theta_n)$ where two possible assumptions, representing extremes in viewpoints, will be considered.

(A) $(X_1, \theta_1), \ldots, (X_n, \theta_n)$ is a sequence of independent, identically distributed random vectors with the distribution of (X, θ) which is given by

$$P\{\theta = j\} = \pi_j, \quad 1 \leqq j \leqq M, \tag{2.1a}$$

$$P\{X \leqq x | \theta = j\} \text{ has a probability density } f_j, \quad 1 \leqq j \leqq M. \tag{2.1b}$$

(B) $(X_1, \theta_1), \ldots, (X_n, \theta_n)$ is a sequence of independent random vectors where, for each $1 \leqq i \leqq n$, X_i has a probability density f_{θ_i}.

Assumption (A), the Bayesian viewpoint, models θ as a random variable with a probability distribution (2.1a) while (B), the deterministic viewpoint, treats $\theta_1, \ldots, \theta_n$ as a deterministic sequence selected by nature. If (X, θ) is independent of the data, the nonparametric discrimination problem is to determine how the statistician should use X and the data to estimate θ when he assumes only (A) or (B) and that $f_1, \ldots, f_M$ are almost everywhere continuous on $\mathbb{R}^d$ [1].

A *discrimination rule*, or, simply, *rule* or *procedure*, is a sequence $\{\delta_n\}$ where, for each n, the decision $\hat{\theta}$ of the rule, is determined by

$$\begin{aligned} &\delta_n : \mathbb{R}^d \times (\mathbb{R}^d \times \{1, \ldots, M\})^n \to [0, 1]^M, \\ &\delta_n = (\delta_{n1}, \ldots, \delta_{nM}), \ \textstyle\sum_{j=1}^{M} \delta_{nj} = 1, \\ &P\{\hat{\theta} = j | X, (X_1, \theta_1), \ldots, (X_n, \theta_n)\} = \delta_{nj}, \quad 1 \leqq j \leqq M. \end{aligned} \tag{2.2}$$

[1] This last assumption essentially restricts the statistician to not only dealing with measures which are absolutely continuous with respect to Lebesgue measure but ones which also have Riemann integrable densities. This restriction is sometimes relaxed or replaced by the assumption that $f_1, \ldots, f_M$ are discrete densities.

Thus $\hat{\theta}$ is a randomized decision drawn according to a distribution δ_n which is allowed to depend on the past observations. For each n, there are M conditional probabilities of error for the rule $\{\delta_n\}$ given the data, namely,

$$L_n^j = P\{\hat{\theta} \neq j | (X_1, \theta_1), \ldots, (X_n, \theta_n), \theta = j\}, \quad 1 \leqq j \leqq M.$$

L_n^j is a random variable whose value is the limiting frequency of errors made when a large number of observations, all with state j, have their states estimated with δ_n and the given data. For the Bayesian problem

$$\begin{aligned} L_n &= P\{\hat{\theta} \neq \theta | (X_1, \theta_1), \ldots, (X_n, \theta_n)\} \\ &= \textstyle\sum_{j=1}^{M} \pi_j L_n^j \end{aligned}$$

is the probability of error for δ_n given the data. Its value is the limiting frequency of errors made when a large number of independent observations, whose states now occur independently with the distribution (2.1a), have their states estimated with δ_n and the given data. The L_n^j's (and L_n, when one can use the Bayesian assumption) measure the performance that the statistician will get when he applies the rule to his data.

It is also convenient to have rules which can make no decision, or "reject" the observation. Here

$$\begin{aligned} &\delta_n : \mathbb{R}^d \times (\mathbb{R}^d \times \{1, \ldots, M\})^n \to [0, 1]^{M+1} \\ &\delta_n = (\delta_{n0}, \ldots, \delta_{nM}); \ \textstyle\sum_{j=0}^{M} \delta_{nj} = 1 \\ &P\{\hat{\theta} = j | X, (X_1, \theta_1), \ldots, (X_n, \theta_n)\} = \delta_{nj}, \quad 0 \leqq j \leqq M \end{aligned}$$

describes a rule where $\hat{\theta} = 0$ means "reject" or make no decision. One now has $2M$ random variables

$$\begin{aligned} L_n^j &= P\{\hat{\theta} \neq j, 0 | (X_1, \theta_1), \ldots, (X_n, \theta_n), \theta = j\}, \quad 1 \leqq j \leqq M \\ R_n^j &= P\{\hat{\theta} = 0 | (X_1, \theta_1), \ldots, (X_n, \theta_n), \theta = j\}, \quad 1 \leqq j \leqq M \end{aligned}$$

representing, respectively, the probability of error for the rule given the data and $\theta = j$, and the probability of reject for the rule given the data and $\theta = j$. With the Bayesian assumption

$$\begin{aligned} L_n &= \textstyle\sum_1^M \pi_j L_n^j \\ R_n &= \textstyle\sum_1^M \pi_j R_n^j \end{aligned}$$

become, respectively, the probability of error and the probability of reject for the rule, given the data, with the analogous frequency interpretations to those used earlier.

One obvious rule that might be employed in the Bayesian problem is to estimate $\pi = (\pi_1, \ldots, \pi_M)$, $f = (f_1, \ldots, f_M)$ from the data and then use these estimates in the Bayes procedure as though they were correct. The class of such rules that

one obtains by using different estimates of $f_1, \ldots, f_M$ from the data are called "two-step" procedures. A similar designation is also used in the deterministic problem.

A rule $\{\delta_n\}$ is termed *symmetric* if the order of the data $(X_1, \theta_1), \ldots, (X_n, \theta_n)$ is immaterial to the value of δ_n (e.g., a permutation of $(X_1, \theta_1), \ldots, (X_n, \theta_n)$ leaves δ_n unchanged). A rule is termed *local* if there is an integer r such that, with probability one, δ_n depends on X and at most its r nearest neighbors and their states. Local rules are, of course, symmetric. A local rule is called *consistent* if, whenever an observation is deleted from the data, $\hat{\theta}$ remains unchanged as long as that observation is not one of the r_0 nearest neighbors of X where r_0 is the smallest integer r for which the rule satisfies the definition of being local.

Examples:

1) The Bayes rule for (2.1) does not depend on the data and is therefore a consistent local rule with $r_0 = 0$.

2) Let h be a positive number and let $\hat{\theta}$ be the state which has the most number of observations within a distance h of X. Ties are broken at random. This rule is a two-step rule which is symmetric but not local.

3) The k-nearest neighbor rule (k-NNR) takes $\hat{\theta}$ to be the state which occurs most often among the k nearest neighbors of X. Two types of ties can occur, ties in voting and ties in distance. In the first case, say with $k = 5$ and $M = 3$, the votes of the nearest neighbors of X may be 1, 1, 2, 2, 3. In this situation the k-NNR breaks the tie at random between states 1 and 2. Ties in distance occur with probability 0 when the conditional probability measures have probability densities as we have assumed. The k-NNR is a consistent local rule with $r_0 = k$.

4) Suppose $d = 1$ and $M = 2$. One looks at the data and finds a threshold t such that one of the two decision procedures

$$\hat{\theta} = \begin{cases} 2, & x \geqq t \\ 1, & x < t \end{cases}$$

$$\hat{\theta} = \begin{cases} 1, & x \geqq t \\ 2, & x < t \end{cases}$$

minimizes the frequency of state 2 errors given that the frequency of state 1 errors is less than or equal to α, $0 < \alpha < 1$. This rule is symmetric but neither two-step nor local.

One of the questions that statisticians, or at least researchers, always ask about a rule is how well it performs in the large sample case. In fact, the bulk of investigation for the nonparametric descrimination problem so far has been devoted to demonstrating asymptotic properties of various rules.

For the Bayesian problem, rules of the form (2.2) are called asymptotically optimal if

$$L_n \xrightarrow{n} L^* \text{ in probability}$$

where $L^* = L^*(\pi, f)$, the Bayes probability of error, is the smallest probability of error possible when the distribution (2.1) is known. Rules which allow rejects are

called asymptotically optimal if

$$L_n \xrightarrow{n} L \text{ in probability}$$
$$R_n \xrightarrow{n} R \text{ in probability}$$

where, when (2.1) is known, R is the smallest probability of reject possible among all rules which have a probability of error less than or equal to L.

For the deterministic problem, asymptotic optimality is difficult to describe except in the case $M=2$. For this case rules of the form (2.2) are asymptotically optimal if

$$L_n^1 \xrightarrow{n} L_1 \text{ in probability}$$
$$L_n^2 \xrightarrow{n} L_2 \text{ in probability}$$

where, assuming f_1, f_2 are known, L_1 is the smallest probability of error possible given $\theta=1$ for all rules with a probability of error, given $\theta=2$, less than or equal to L_2.

Other notions of asymptotic optimality are possible in the deterministic case. One can think of the data unfolding, one observation at a time, yielding a frequency of errors after n observations equal to

$$\frac{1}{n-1}\sum_1^{n-1} I_{[\hat{\theta}_{j+1} \neq \theta_{j+1}]}$$

where $\hat{\theta}_{j+1}$ is the estimate of θ_{j+1} from X_{j+1} and $(X_1, \theta_1), \ldots, (X_j, \theta_j)$ using δ_j. If $L^*(\hat{\pi}, f)$ represents the Bayes probability of error computed with the empirical probabilities $\hat{\pi}_1, \ldots, \hat{\pi}_M$ (from n observations), one can now consider rules asymptotically optimal if

$$\frac{1}{n-1}\sum_1^{n-1} I_{[\hat{\theta}_{j+1} \neq \theta_{j+1}]} - L^*(\hat{\pi}, f) \xrightarrow{n} 0 \text{ in probability}$$

uniformly in the sequences $\{\theta_i\}$. Such a rule has a robust character in that it drives the difference between the observed frequency of errors and its expected empirical counterpart to zero regardless of the sequence $\{\theta_i\}$. Nevertheless, the quantity $L^*(\hat{\pi}, f)$ which is being tracked is irrevocably linked to the past in that it tells us nothing about future performance unless the future empirical distribution of the sequence $\{\theta_i\}$ is the same as that in the past.

In the following two subsections we review recent work on the Bayesian problem and all of the work of which we are aware for the deterministic problem. The emphasis in these subsections is on describing rules and their asymptotic properties, if any. The next subsection is devoted to error rate estimation. That is, given the data and a particular rule, how does the statistician estimate the current performance of his rule? The final subsection is devoted to recent results in nonparametric density estimation.

2.1.2 The Deterministic Problem

In this subsection we review the known results for the deterministic problem where, for simplicity, we consider only $M=2$. The standard asymptotic result for two-step procedures is due to FIX and HODGES [2.1]. In particular, if $\hat{f}_i$ is a consistent estimate of f_i from the data[2] for each i and if $\{x: f_1(x)=t f_2(x)\}$ has zero probability with each density, then

$$\hat{\theta}=\begin{cases}1, & \hat{f}_1(X)/\hat{f}_2(X)\geqq t\\ 2, & \hat{f}_1(X)/\hat{f}_2(X)<t\end{cases}$$

yields[3]

$$L_n^i \to P\{\theta' \neq i \mid \theta=i\} \triangleq L_i \text{ in probability } i=1,2$$

where

$$\theta'=\begin{cases}1, & f_1(X)/f_2(X)\geqq t\\ 2, & f_1(X)/f_2(X)<t\end{cases}$$

and $0<t<\infty$ is fixed but arbitrary. The Neyman-Pearson lemma guarantees that L_1 is the smallest possible probability of error, given $\theta=1$, for any rule which has a probability of error, given $\theta=2$, less than or equal to L_2. Thus if the set $\{x: f_1(x)=t f_2(x)\}$ has zero probability with each density, then two-step procedures using consistent estimates are asymptotically optimal. The convergence of the L_n^i can be strengthened to convergence with probability one if the requirement for consistent estimates is replaced by strongly consistent estimates. While the above two-step rules are asymptotically optimal, it is nevertheless assumed that one has picked a suitable t for the given application. For example, one has to be satisfied with the value of L_2 obtained. Finally, if $\hat{f}_i$ is any one of the usual density estimates from the observations having state i we need only have an infinite number of observations from each state in order to satisfy the given consistency requirements.

The rule suggested by FIX and HODGES was the forerunner of the k-nearest neighbor rule of COVER and HART [2.2] (see Ex. 3). In the context of the density estimation approach here, their rule is obtained by setting $\hat{f}_i(x)=M_i/n_i V$ where

a) n_i is the number of observations in the data with state i,

b) V is the volume of the smallest sphere S about x containing at least k observations from the data,

c) M_i is the number of observations having state i which are contained in S, and

d) $k=k(n_1,n_2)\to\infty$, $k(n_1,n_2)/n_i\to 0$, $i=1,2$, when $n_1,n_2\to\infty$ with n_1/n_2 bounded away from 0 and ∞.

The metric used in proving consistency of these densities was the ordinary Euclidean metric although the result holds for other metrics. There is, of course,

[2] $\hat{f}_i$ is a consistent estimate of f_i if $\hat{f}_i(x) \to f_i(x)$ in probability for each x at which f_i is continuous. The estimate is called strongly consistent if the convergence is with probability one.

[3] The result in [2.1] is somewhat stronger than stated here.

a very slight additional restriction put on the sequences $\{\theta_i\}$ in order to obtain the consistent density estimates of the above rule.

FIX and HODGES [2.3], assuming simple cases for normal densities, computed the average probability of error of each type for $n_1 = n_2$ and $k=1$, $k=3$. For the univariate case these probabilities were compared with the corresponding ones obtained by estimating the parameters in the optimal linear discriminant function. Their conclusion was

> "If the populations to be discriminated are well known, and have been investigated to establish that the normal distribution gives a good fit and that the variances and correlations do not change much when the means are changed, and if the classification to be made warrants the labor of matrix inversion, then the linear discriminant function should certainly be used. If on the other hand, the populations are either not well known, or are known not to be approximately normal, or to have very different covariance matrices; or if the discrimination is one in which small decreases in probability of error are not worth extensive computations, then the simple nonparametric rule, perhaps with $k \geqq 3$, seems to have the edge."

KENDALL [2.4] discussed two procedures, the first one called the convex-hull method. Here

$$\hat{\theta} = \begin{cases} 1 & \text{if} \quad X \in A_1 \bar{A}_2 \\ 2 & \text{if} \quad X \in A_2 \bar{A}_1 \\ 0 & \text{otherwise}, \end{cases}$$

where A_i is the convex-hull of the observations with state i. Despite the simplicity of this procedure the statistician will, with probability one, be making no decision as $n_1, n_2 \to \infty$ for typical densities f_1, f_2 which have a common support. KENDALL also described a procedure termed the order-statistic method. Here one looks at the coordinate with the best discrimination (e.g., most decisions) using the convex-hull method. The remaining observations which are not discriminated are used with another coordinate in the same manner and so on. At the end of the process one has a sequence of tests on each coordinate which are applied until X is discriminated or until one exhausts the coordinates and makes no decision. This rule, while also simple, has the same difficulty as the convex-hull method.

QUESENBERRY and GESSAMAN [2.5] introduced a procedure using the theory of coverages which, for the $M=2$ case, can be described as follows. Let $P_i(\cdot)$ denote the probability measure on Borel subsets of $\mathbb{R}^d$ corresponding to the density f_i. If the data contains n_i observations with state i, then, given α_1, α_2 with $0 \leqq \alpha_1, \alpha_2 \leqq 1$, and using any ordering function which satisfies mild conditions, sets A_1, A_2 can be found for which $P_j(A_j)$ is a random variable with a Beta distribution Beta $[a_j, n_j - a_j + 1]$, a_j being the integer part of $\alpha_j(n_j+1)$. Letting

$$\hat{\theta} = \begin{cases} 1 & \text{on} \quad \bar{A}_1 A_2 \\ 2 & \text{on} \quad \bar{A}_2 A_1 \\ 0 & \text{on} \quad \bar{A}_1 \bar{A}_2 \cup A_1 A_2 \end{cases}$$

then yields the inequalities

$$L_n^1 \leqq P_1(A_1)$$
$$L_n^2 \leqq P_2(A_2)$$

where, in addition, $P_i(A_i) \xrightarrow{n_i} \alpha_i$ in probability, $i=1,2$. Thus one can control the two types of error probabilities asymptotically to be less than or equal to α_1, α_2, respectively. The difficulty with this procedure is that one cannot be sure that the reject region has minimum probability for each state value and no procedure for constructing A_1, A_2 was given which guaranteed this unless f_1, f_2 were known to belong to special parametrically described families of densities. Even if one were willing to make one of these parametric assumptions the rejection rates may still be too high for the given application. Finally, the method given, when extended to $M>2$ states, requires regions of partial decisions, that is, if X falls in a set $S_{i_1,\ldots,i_k}$, then one merely asserts that X has one of the states $i_1, \ldots, i_k$.

In [2.5] the regions A_1 and A_2 are not specified even though one has already picked the n_i+1 blocks of each type. ANDERSON and BENNING [2.6] described a method for picking the ordering functions and A_1, A_2 such that the region of no decision tends to have a small probability for each state value while, at the same time, keeping the properties of [2.5] for controlling the individual error probabilities.

BEAKLEY and TUTEUR [2.7] eliminated the no decision region altogether. Methods for determining a sequence of statistically equivalent blocks $B_1, \ldots, B_{n_1+1}$ for state 1 were given with the property that *all* n_2 observations from state 2 were contained in B_1. Thus if

$$\hat{\theta} = \begin{cases} 2, & X \in \bigcup_1^m B_i \\ 1, & X \in \bigcup_{m+1}^{n_1+1} B_i \end{cases}$$

then $L_n^1 = P_1(\bigcup_1^m B_1)$ is a random variable with Beta $[n, n-m+1]$ distribution, and L_n^1 can be controlled as in the QUESENBERY and GESSAMAN case. $L_n^2 = P_2(\bigcup_{m+1}^{n_1+1} B_i) = 1 - P_2(\bigcup_1^m B_i)$ intuitively is being minimized since all of the state 2 observations are in B_1. The conditions needed to insure the asymptotic optimality of this procedure are unknown, however.

HENRICHON and FU [2.8, 9] and ANDERSON [2.10] have also discussed various discrimination rules but little is known analytically about their proposed procedures. OWEN et al. [2.11], for the two-state univariate case, described a scheme which asymptotically locates the extrema of $F_2(x) - tF_1(x)$ where F_i is the distribution function corresponding to f_i. The procedure requires that (a) the set of extrema is finite and (b) the smallest distance between the extrema is known. Thus the regions $\{x: f_1(x)/f_2(x) \geqq t\}$ and $\{x: f_1(x)/f_2(x) < t\}$ can asymptotically be found for these assumptions and, in fact, the procedure is asymptotically optimal in this case. GESSAMAN and GESSAMAN [2.12] experimentally compared various rules for three different sample sizes and three different choices of bivariate normals for f_1, f_2.

In [2.13] VAN RYZIN, using kernel estimates for the densities and empirical frequencies for the states in Bayes rule, showed that estimating θ_{n+1} from X_{n+1} and $(X_1, \theta_1), \ldots, (X_n, \theta_n)$ yields

$$\frac{1}{n} \sum_1^n I_{[\hat{\theta}_{i+1} \neq \theta_{i+1}]} - L^*(\hat{\pi}, f) \xrightarrow{n} 0 \text{ in probability}$$

uniformly in all sequences $\{\theta_i\}$.

2.1.3 The Bayesian Problem

In this subsection we review the recent work on the Bayesian problem where, for the most part, we use the review paper of COVER [2.14] as our starting point.

Two-step rules are asymptotically optimal if one uses reasonable estimates of $\pi_1, \ldots, \pi_M, f_1, \ldots, f_M$. In particular, if

$$\hat{\theta} = \begin{cases} j, \hat{\pi}_j \hat{f}_j(X) \geqq \hat{\pi}_i \hat{f}_i(X), & 1 \leqq i \leqq M \\ \text{ties are broken arbitrarily} \end{cases}$$

then

$$L_n \xrightarrow{n} L^* \text{ in probability}$$

when $\hat{\pi}_1, \ldots, \hat{\pi}_M, \hat{f}_1, \ldots, \hat{f}_M$ are consistent estimates of $\pi_1, \ldots, \pi_M, f_1, \ldots, f_M$ from the data. In fact,

$$L_n \xrightarrow{n} L^* \text{ with probability one}$$

if the estimates are strongly consistent. This Bayesian analogue of the FIX-HODGES result for the deterministic problem does not appear specifically in the literature (see, however, ROBBINS [2.15], VAN RYZIN [2.16], SCHWARTZ [2.17] and GLICK [2.18] for similar versions).

If $\{\delta_n\}$ is a symmetric rule then L_n and R_n, if rejects are allowed, are symmetric functions of the data. The HEWITT-SAVAGE 0–1 law [2.19, 20] implies that such sequences of random variables converge with probability one to a constant or diverge with probability one. This appealing all-or-nothing behavior for symmetric rules appears to be difficult to take advantage of, however, in analyzing any particular rule.

In some situations the statistician wants to estimate the state of a specific X where it is demanded that a decision be made after each new observation in the data. If he uses a two-step procedure, then he wants one which, in some sense, can be recursively updatable for each value of X as the data unfolds, one observation at a time. For $M=2$, VAN RYZIN [2.21] demonstrated the asymptotic optimality of a two-step procedure which essentially uses recursive versions of kernel density estimates for f_1, f_2. For a similar procedure, WOLVERTON and WAGNER [2.22] showed that $L_n \xrightarrow{n} L^*$ with probability one and that

$$P[\bigcup_{k=n}^{\infty} \{L_k - L^* \geqq \varepsilon\}] \leqq A/n^{(2d+1)}$$

for a particular choice of window width in the density estimates. REJTÖ and RÉVÉSZ [2.23], with additional assumptions on the derivatives of f_1, f_2, have shown that

$$P[L_n - L^* \geqq \varepsilon] \leqq e^{-An^{(d+1)/(2d+1)}}$$

for a scheme similar to [2.22]. In each of the above procedures symmetry has been sacrificed to obtain a recursive procedure.

One of the difficulties in applying the rule of example 2) in Subsection 2.1.1 is that the statistician does not know how to choose the value of h[4]. A nice idea was investigated by PELTO [2.24] who, for $M=2$ and known π_1, π_2, examined different ways of choosing h from the data. One method consists of choosing an h which minimizes

$$\pi_1 \hat{L}_n^1 + \pi_2 \hat{L}_n^2$$

where $\hat{L}_n^1, \hat{L}_n^2$ are the deleted estimates of L_n^1, L_n^2 from the data (see Subsection 2.1.4). An argument was given which indicates that this method yields an asymptotically optimal rule. Unfortunately the argument is incomplete since it does not take into consideration the fact that the h chosen is now a random variable. Also discussed was a second method for choosing h, which heuristically has a smaller bias in estimating $\pi_1 L_n^1 + \pi_2 L_n^2$.

The k-nearest neighbor rule [see example 3) of Subsection 2.1.1] was initially investigated by COVER and HART [2.2] who, among other things, showed that the single nearest neighbor rule asymptotically "plays nature against itself". For example, for $M=2$,

$$EL_n \to \int_{R^d} \frac{2\pi_1 f_1 \pi_2 f_2}{\pi_1 f_1 + \pi_2 f_2}\, dx \triangleq L(1)$$

which is the same probability of error obtained using the rule

$$\delta(X)=j \text{ with probability } \pi_j f_j(X)/(\pi_1 f_1(X)+\pi_2 f_2(X)), \quad j=1,2,$$

that is, the rule that chooses state j with its *a posteriori* probability given X. However, for a finite amount of data, the single nearest neighbor rule is a deterministic rule (with probability one) which achieves its effective randomization in the large sample case by partitioning each small neighborhood of x into many small subregions where, on each subregion, the decision for a given state is constant and where the total fraction of subregion areas for state j is approximately $\pi_j f_j(x)/\sum_1^M \pi_j f_j(x)$. What, of course, makes the rule interesting is the tight inequality [2.2]

$$L^* \leqq L(1) \leqq 2L^*(1-L^*).$$

Beyond this COVER and HART showed, for the k-nearest neighbor rule, that

$$EL_n \xrightarrow{n} L(k)$$

and gave tight bounds for $L(k)$ in terms of L^*, WAGNER [2.25] demonstrated that

$$L_n \xrightarrow{n} L(k) \text{ in probability}$$

[4] This rule is the two-step rule obtained by using the empirical estimates of $\pi_1, \ldots, \pi_M$ and kernel estimates of $f_1, \ldots, f_M$ with a kernel width h and a kernel which is uniform over the unit sphere centered at the origin (see Subsection 2.1.5).

with convergence being with probability one if all components of X have a finite first moment with the density $\sum_1^M \pi_i f_i$. FRITZ [2.26] has shown that the above convergence for the single nearest neighbor rule is always with probability one and that $P\{|L_n - L(1)| \geqq \varepsilon\}$ is asymptotically dominated by $A\exp(-B\sqrt{n})$ where A and B are functions of ε and d but not of the distribution of (X, θ).

Suppose the statistician is looking at a large amount of data with $\pi_1 = \pi_2 = 1/2$ and $f_1 = N(-1, 1)$, $f_2 = N(1, 1)$. For positive values of x, state 2 observations will be denser than state 1 observations, while for negative x the reverse will be true. For example, for x greater than 0, one will tend to see each state 1 observation surrounded by strings of state 2 observations. One reasonable thing to try to do is "edit out" the weaker class. WILSON [2.27] took this approach by first editing the data with the k-nearest neighbor rule and then using the nearest neighbor rule on the edited data set. For example,

1) let $\hat{\theta}_j$ be the k-NNR estimate of θ_j from X_j and the data with (X_j, θ_j) *deleted*, $1 \leqq j \leqq N$, and
2) edit (X_j, θ_j) from the data if $\hat{\theta}_j \neq \theta_j$.

The single nearest neighbor rule is now used with the edited data set to estimate the state θ of an unclassfied observation X. WILSON showed that

$$EL_n \xrightarrow{n} L^E(k)$$

where $L^E(k) \leqq L(k)$ for all k and, for the first 3 odd values of k,

$$L^* \leqq L^E(1) \leqq 1.2L^*$$
$$L^* \leqq L^E(3) \leqq 1.149L^*$$
$$L^* \leqq L^E(5) \leqq 1.10L^*\ ^5.$$

where, for comparison, the k-nearest neighbor rule has

$$L^* \leqq L(1) \leqq 2L^*$$
$$L^* \leqq L(3) \leqq 1.31L^*$$
$$L^* \leqq L(5) \leqq 1.2L^*.$$

Thus, by editing out the less dense observations with the k-NNR, one not only reduces the storage requirements for future classification but improves the asymptotic performance over the k nearest-neighbor rule as well. This rule is symmetric but not local since, for example, an arbitrary number of nearest neighbors of X can be deleted in the editing process. Also it is not known if $L_n \xrightarrow{n} L^E(k)$ in any sense.

Other rules have been suggested whose primary purpose is to reduce the data before the nearest neighbor rule is applied, for example, the condensed nearest neighbor rule and the reduced nearest neighbor rule [2.28–31]. CHANG [2.32] has considered reducing the data by merging two nearest vectors of the same

[5] There is a flaw in WILSON's argument which may make these bounds only approximate.

state into a weighted average of the two as long as the recognition rate on the data is not lowered. All three rules are nonsymmetric and nothing is known analytically about their finite or asymptotic performance.

HELLMAN [2.33] considered modifying the k-nearest-neighbor rule to allow rejects. The modified rule makes the decision of the k-nearest-neighbor rule only if all k nearest neighbors agree; otherwise it makes no decision. It was shown that

$$EL_n \xrightarrow{n} L$$
$$ER_n \xrightarrow{n} R$$

where, for the given asymptotic reject rate R, the corresponding probability of error L satisfies

$$L \leqq (1+k/2)L_R^* .$$

Here L_R^* is the smallest possible probability of error for all rules with a given probability of reject less than or equal to R and the constant $(1+k/2)$ is the smallest possible. Thus, when $k=2$ and when errors are twice as costly as rejects, the above rule always performs better than the single nearest-neighbor rule. The modified rule is symmetric, local and consistent.

CHOW [2.35], in one of the earliest applications of decision theory to pattern recognition, modified the Bayes rule to allow rejects. In particular, if

$$r(x)=1-\max_{1\leqq i\leqq M}\{\pi_i f_i(x)/f(x)\}$$
$$f(x)=\sum_1^M \pi_i f_i(x)$$

then the rule

$$\hat{\theta}=\begin{cases} j \text{ when } 1-(\pi_j f_j(X)/f(X))=r(X)\leqq t \\ 0 \text{ when } r(X)>t \end{cases}$$

yields a probability of error L and a probability of reject R which are functions of t, $0\leqq t\leqq 1$, and, furthermore, for each t, $R(t)$ is the smallest reject probability of any rule with an error probability less than or equal to $L(t)$. In [2.34] CHOW exhibited a simple relationship between L and R, namely,

$$L(t)=-\int_0^t t'\,dR(t')=\int_0^t R(t')dt'-tR(t) .$$

2.1.4 Probability of Error Estimation

Estimating the probability of error that a statistician has with a particular rule and his data, sometimes called error estimation in the literature, is the topic of this subsection. The emphasis is on the existence of distribution-free bounds for these estimates and, because such bounds occur infrequently in statistical problems, we first describe a distribution-free bound for estimating distribution

functions which will be used later for probability of error estimates with a particular class of rules.

Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables with a common distribution function F. Suppose F is unknown to us and we wish to estimate it from $X_1, \ldots, X_n$. If F_n is the empirical distribution function for $X_1, \ldots, X_n$ [e.g., $F_n(x)$ is the frequency of the first n observations less than or equal to x] then the Glivenko-Cantelli theorem asserts that

$$D_n \triangleq \sup_x |F_n(x) - F(x)| \xrightarrow{n} 0 \text{ with probability one.}$$

While this reassures the statistician that, for large n, F_n should be a good estimate of F in the sense of making D_n small, he cannot infer, say, how large n should be in order to insure that

$$P[D_n \geqq \varepsilon] \leqq \delta,$$

where ε, δ are two pre-assigned arbitrary positive numbers. What is being sought here is very similar to finding confidence intervals for estimating an unknown parameter, the main difference being that this problem is nonparametric and we want an n here which will work for all F. A nice result of DVORETSKY et al. [2.36] is that

$$P[D_n \geqq \varepsilon] \leqq C_0 e^{-2\varepsilon^2 n},$$

where C_0 is a universal constant which does not depend on F. Choosing the smallest n such that

$$C_0 e^{-2\varepsilon^2 n} \leqq \delta$$

satisfies our requirements. The bound above, which works for all F, is called *distribution free* because it does not depend on F.

Probability of error estimation is the problem of how the statistician estimates the *current* performance of his rule, that is, the performance he will get when he applies the rule to his data. In particular, in the Bayesian problem, the statistician would like to estimate L_n from the given data while in the deterministic problem he would like to estimate L_n^j, $1 \leqq j \leqq M$, from the data, and, indeed, he may wish to estimate these latter quantities in the Bayesian problem as well. In addition, the statistician may wish to estimate the ultimate performance of the rule, if any, with the data. In the Bayesian problem, for example, suppose he has a rule for which

$$L_n \xrightarrow{n} L \text{ in probability},$$

where L is not necessarily equal to L^*. L then is the performance of the rule with an infinite amount of data. In those situations where there is the possibility of gathering more data, the statistician would be interested in L if only because a large value of L would indicate that the proposed discrimination is unfeasible.

Generally, though, the interest is in L_n since the data is always finite. We will comment on another aspect of these two different problems later in this subsection.

TOUSSAINT [2.37] has recently published an extensive bibliography on the estimation of error probabilities to which we refer the reader for an historical perspective and a complete survey, while we concentrate here on distribution-free aspects of the error estimation problem. The recent review by KANAL [2.38], particularly the section on error estimation, is also highly recommended. The estimates[6] which have been considered extensively are as follows.

(A) *Resubstitution Estimate:* $\hat{L}_n \triangleq \frac{1}{n}\sum_1^n I_{[\hat{\theta}_j \neq \theta_j]}$ where $\hat{\theta}_j$ is the estimate of θ_j using δ_n with X_j and all of the data.

(B) *Holdout Estimate:* $\hat{L}_n \triangleq \frac{1}{n\alpha}\sum_{n-n\alpha+1}^n I_{[\hat{\theta}_j \neq \theta_j]}$ where $\hat{\theta}_j$ is the estimate of θ_j using $\delta_{n-n\alpha}$ with X_j and $(X_1, \theta_1), \ldots, (X_{n(1-\alpha)}, \theta_{n(1-\alpha)})$, $n-n\alpha < j \leqq n$. The fraction of the data "held out" is α and we have assumed that αn is an integer.

(C) *Deleted Estimate:* $\hat{L}_n \triangleq \frac{1}{n}\sum_1^n I_{[\hat{\theta}_j \neq \theta_j]}$ where $\hat{\theta}_j$ is the estimate of θ_j using δ_{n-1} with X_j and the data with (X_j, θ_j) deleted.

(D) *Rotation Estimate:* $\hat{L}_n \triangleq \frac{1}{n}\sum_1^n I_{[\hat{\theta}_j \neq \theta_j]}$ where $\hat{\theta}_j$ is the estimate of θ_j using δ_{n-l} with X_j and the data where the l-block that contains (X_j, θ_j) is deleted. Here, l is an integer which divides n (e.g., $n = ml$) and the data is partitioned into m consecutive l-blocks: $(X_{(i-1)l+1}, \theta_{(i-1)l+1}), \ldots, (X_{il}, \theta_{il})$, $1 \leqq i \leqq m$.

Obviously the deleted estimate is a special case of the rotation estimate with $l = 1$. The rotation estimate and the holdout estimate are, in general, not symmetric since they depend on the order of the observations. With $l = n/2$ the rotation estimate is just the average of two holdout estimates, the regular one and the one with the data reversed. The re-substitution estimate is frequently an optimistic estimate of L_n, but has been shown by FRALICK and SCOTT [2.39] to be a consistent estimate of L^* for a wide class of asymptotically optimal two-step procedures. With the nearest neighbor rule, for example, (A) always yields an estimate of 0 for L_n. The deleted estimate can require considerable computation but, with local rules, the computation is reasonable and its intuitive appeal can be taken advantage of.

The aspect which interests us here is whether these estimates have distribution-free performance bounds for any types of rules. In [2.40], for the Bayesian problem, it is shown that for any consistant local rule using k nearest neighbors

$$E(L_n - \hat{L}_n)^2 \leqq \frac{(2k+\frac{1}{4})}{n} + \frac{2k(2k+\frac{1}{4})^{\frac{1}{2}}}{n^{3/2}} + \frac{k^2}{n}, \tag{2.3}$$

where $\hat{L}_n$ is the deleted estimate of L_n. Thus, for a given $\varepsilon, \delta > 0$, an n can be specified using (2.3) and Chebychev's inequality to insure that

$$P[|\hat{L}_n - L_n| \geqq \varepsilon] \leqq \delta$$

[6] For simplicity we consider these estimates for the Bayesian problem. Their counterparts for the deterministic problem are usually obvious.

regardless of the nonparametric discrimination problem considered. The n obtained through the use of (2.3), however, is probably far from the smallest possible.

Several comments are appropriate. First, the variance of $\hat{L}_n$, usually a focus of study in the literature, is more or less irrelevant since the pertinent mean-square quantity is $E(\hat{L}_n - L_n)^2$. Second, the result (2.3) seems surprising since it does not depend on d! Thus, if the statistician uses a local rule he does not need to be concerned about extraneous or irrelevant components affecting the accuracy of the deleted estimate. Finally (2.3) is valid for arbitrary conditional measures, which may include atoms, provided ties are broken in a satisfactory way. For example, suppose $Z_1, Z_2, \ldots$ is a sequence of independent random numbers from $[0, 1]$ where Z_i is "attached" to (X_i, θ_i) for each $i = 1, 2, \ldots$. If ties in distance are broken by choosing the observations with the smallest attached numbers then the ordering of the neighbors of X is specified with probability one and (2.3) holds for local rules using this ordering.

This result can also be used to select the best set of l out of d components for use with a local rule. For example, if one has m different subsets of the d components and picks that subset with the smallest deleted risk estimate, then the probability that he has picked a subset whose L_n is within 2ε of that of the best subset of those tested is $\geqq 1-(mA/\varepsilon^2)$ where A is given by the right-hand-side of (2.3). Putting $m = \binom{d}{l}$ then allows the selection of the best set of l components, although the n required in the bound, to be useful, will probably be quite large. While it seems to be true that deleted estimates work well for local rules, one should not conclude that they will necessarily work well for all rules or that any other estimate will work well for local rules. One might wonder why it is that local rules, using k-nearest neighbors, have a performance bound which does not depend on d but only on k. Intuitively it seems that local rules exchange k for d; for example, the nearest neighbor rule essentially reduces the discrimination procedure to one dimension since only distances from the observations to x are factors in the decision.

The re-substitution estimate, often criticized in the literature, also yields a distribution-free performance bound on linear rules. First consider the deterministic problem for $M = 2$ and $d = 1$ with rules of the form

$$\hat{\theta} = \begin{cases} 2, & X > t \\ 1, & X \leqq t \end{cases}$$

where t is some function of the data. If n_i observations have state i, and if $\hat{F}_i$ represents the empirical distribution function for the observations with state i, then $\hat{F}_2(t)$ is the re-substitution estimate of $P\{\hat{\theta} \neq 2 | \theta = 2\} = F_2(t) = L_n^2$ while $1 - \hat{F}_1(t)$ is the re-substitution estimate of $P\{\hat{\theta} \neq 1 | \theta = 1\} = 1 - F_1(t) = L_n^1$. Thus, for *all* functions t,

$$P[|\hat{L}_n^1 - L_n^1| \geqq \varepsilon] \leqq C_0 e^{-2n_1\varepsilon^2}$$
$$P[|\hat{L}_n^2 - L_n^2| \geqq \varepsilon] \leqq C_0 e^{-2n_2\varepsilon^2}.$$

This result may generalize to higher dimensions but, for the present, remains an open question. For example let $Z, Z_1, Z_2, \ldots$ be a sequence of independent,

identically random vectors with values in $\mathbb{R}^m$ and suppose that we can extend the KIEFER and WOLFOWITZ [2.41] result from semi-infinite rectangles to half planes, namely, to

$$P\left\{\sup_a \left|\frac{1}{n}\sum_1^n I_{[a^T Z_i \leqq 0]} - P[a^T Z \leqq 0]\right| \geqq \varepsilon\right\} \leqq C_0 e^{-Cn\varepsilon_2} \tag{2.4}$$

where C_0, C depend on m but not on the distribution of Z. Consider rules of the form

$$\hat{\theta} = \begin{cases} 2, & a^T \Phi(X) \geqq 0 \\ 1, & a^T \Phi(X) < 0 \end{cases}$$

where

$$a = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix}, \quad \Phi(X) = \begin{pmatrix} \varphi_1(X) \\ \vdots \\ \varphi_m(X) \end{pmatrix}$$

and X takes values in $\mathbb{R}^d$. The "φ-functions" are fixed and the vector a is a function of the data. (See COVER [2.42] for the development of the φ-function approach and related capacity theorems.) By taking $m=2$, $d=1$ and $\varphi_1(x)=x$, $\varphi_2(x)=-1$ we obtain the previous rules with $a_1=1$ and $a_2=t$. As before, we recognize that the first term on the left-hand side of (2.4) corresponds to the re-substitution estimate of L_n^i so that for all vectors a

$$P[|\hat{L}_n^i - L_n^i| \geqq \varepsilon] \leqq C_0 e^{-C n_i \varepsilon^2} \qquad i=1,2.$$

The actual dependence of C and C_0 on m in [2.41] is unknown although it is known that C decreases with m. The performance bound which might be obtained here, while distribution-free, requires that the amount of data needed per state depend on m, the number of features selected. This has been observed before using linear discriminant functions with Gaussian data, the latest analysis being FOLEY [2.43].

Returning to an earlier discussion suppose that the statistician wishes to estimate the ultimate performance of his rule from the data where, for simplicity, we will assume that the ultimate performance is L^*. We now give an example that shows that no reasonable estimate of L^* will have a distribution-free performance. If $\hat{L}_n^*$ is any function of the data used to estimate L^* then its performance is distribution-free if, for each $\varepsilon>0$,

$$\sup_{\pi,f} P[|\hat{L}_n^* - L^*| \geqq \varepsilon] \xrightarrow{n} 0$$

where the supremum is taken over all distributions for which $f_1, \ldots, f_M$ are almost everywhere continuous. It is easy to see that this is equivalent to

$$\sup_{\pi,f}(E|\hat{L}_n^* - L^*|) \xrightarrow{n} 0. \tag{2.5}$$

Taking $M=2$, $\pi_1=\pi_2=1/2$ and

$$f_1=\begin{cases}2, & \frac{2j-2}{2k}\leqq x<\frac{2j-1}{2k}, \quad j=1,\ldots,k\\ 0 \text{ elsewhere}\end{cases}$$

$$f_2=\begin{cases}2, & \frac{2j-1}{2k}\leqq x<\frac{2j}{2k}, \quad j=1,\ldots,k\\ 0 \text{ elsewhere}\end{cases}$$

we have $L^*=0$. Furthermore, we see that if, for each fixed $\theta_1,\ldots,\theta_n$, $\hat{L}_n^*$ is an almost everywhere continuous function of $x_1,\ldots,x_n$, then

$$\begin{aligned}E\hat{L}_n^* &= \textstyle\sum_{\theta_1,\ldots,\theta_n}(\frac{1}{2})^n\int\ldots\int \hat{L}_n^*((x_1,\theta_1),\ldots,(x_n,\theta_n))\, f_{\theta_1}(x_1)\ldots\\ &\quad \ldots f_{\theta_n}(x_n)dx_1\ldots dx_n\\ &\xrightarrow{k} \textstyle\sum_{\theta_1,\ldots,\theta_n}(\frac{1}{2})^n\int_0^1\ldots\int_0^1 \hat{L}_n^*((x_1,\theta_1),\ldots,(x_n,\theta_n))dx_1\ldots dx_n.\end{aligned}$$

Thus, as $k\to\infty$, $E\hat{L}_n^*$ tends to its average estimate of L^* for the "no information" experiment (e.g., $\pi_1=\pi_2=1/2$ and f_1, f_2 both uniform on $[0,1]$). For reasonable estimates of L^* (for example, consistent estimates of L^*) this last quantity does not tend to 0 with n. Thus (2.5) is violated. Bounds on performance of estimates of L^* thus necessarily involve quantities which require more *a-priori* information than is available for the nonparametric discrimination problem.

Estimating L^* from the data can be done in many ways. For example, one could take any two-step procedure and use the estimates considered earlier for L_n. For $M=2$ and assuming that

$$\int_{\{x:\pi_1 f_1(x)=\pi_2 f_2(x)\}}[\pi_1 f_1(x)+\pi_2 f_2(x)]dx=0$$

all of these estimates are consistent estimates of L^* when consistent estimates of π_2, π_1, f_2, f_1 are used [2, 39, 44]. Also, GLICK [2.18] has shown that if one uses consistent estimates $\hat{\pi}_1,\ldots,\hat{\pi}_M, \hat{f}_1,\ldots,\hat{f}_M$ then

$$\hat{L}^*=1-\int\max(\hat{\pi}_1 f_1,\ldots,\hat{\pi}_M\hat{f}_M)dx$$

is a consistent estimate of L^*. Other techniques for estimating L^* have utilized unlabeled observations, an idea first proposed by CHOW [2.34]. For example, $L^*=1-Er(X)$ where, as in Subsection 2.1.3,

$$\begin{aligned}r(x)&=\max\{\pi_1 f_1(x),\ldots,\pi_M f_M(x)\}/f(x),\\ f(x)&=\textstyle\sum_1^M \pi_i f_i(x).\end{aligned}$$

Thus one can estimate $r(x)$ from the labeled observations with, say, $\hat{r}(x)$ and then estimate $Er(X)$ by

$$\frac{1}{T}\textstyle\sum_1^T \hat{r}(Y_i)$$

where $Y_1, \ldots, Y_T$ are the unlabeled observations. Various techniques for estimating $r(\cdot)$ from the labeled observations have been considered by FUKUNAGA and KESSELL [2.45], and FUKUNAGA and HOSTETLER [2.46].

Finally one may be interested in recursive estimates of L^*. One possibility is the estimate used earlier by VAN RYZIN [2.13], namely,

$$\frac{1}{n-1}\sum_1^{n-1} I_{[\hat{\theta}_{j+1} \neq \theta_{j+1}]}$$

where $\hat{\theta}_{j+1}$ is the estimate of θ_{j+1} using δ_j with X_{j+1} and $(X_1, \theta_1), \ldots, (X_j, \theta_j)$. Straightforward arguments (e.g., WOLVERTON [2.47] show that if

$$L_n \xrightarrow{n} L \text{ in probability (or with probability one)}$$

then

$$\frac{1}{n-1}\sum_1^{n-1} I_{[\hat{\theta}_{j+1} \neq \theta_{j+1}]} \xrightarrow{n} L \text{ in probability (or with probability one)}.$$

2.1.5 Density Estimation

Nonparametric density estimation has its raison d'etre rooted in applications like the two-step rules for nonparametric discrimination. After looking at density estimation from a more basic point of view we discuss the results which have been presented since the two paper review of WEGMAN [2.48, 49] and the paper of COVER [2.50].

Density estimation can be viewed as a special case of learning the law of a sequence. This rather basic viewpoint, though certainly not original, singles out a natural error criterion from the seemingly many ad hoc ones which have been considered [2.48]. Suppose $X_1, X_2, \ldots$ is a sequence of independent, identically distributed random vectors with values in $\mathbb{R}^d$ and a common probability measure μ on the Borel σ-algebra β^d of $\mathbb{R}^d$. Estimating the law of the sequence is the problem of finding an estimate μ_n of μ from $X_1, \ldots, X_n$. If the Borel set is *fixed* then

$$\mu_n(B) \triangleq \frac{1}{n}\sum_1^n I_{[X_i \in B]}$$

is the obvious choice as an estimate for the value $\mu(B)$ since it is the minimum variance unbiased estimate of $\mu(B)$. As a function on β^d, μ_n is called the empirical probability measure for $X_1, \ldots, X_n$. The Glivenko-Cantelli theorem states that it is a uniformly good estimate of intervals $(-\infty, x]$, that is,

$$\sup_{\{B: B=(-\infty, x], x \in \mathbb{R}^d\}} |\mu_n(B) - \mu(B)| \to 0 \text{ with probability one}.$$

While it is true that the supremum above can be extended to a larger class of subsets of β^d it nevertheless cannot be extended in general to all Borel sets, see RAO [2.51].

Now, however, assume that we know that μ is absolutely continuous with respect to Lebesgue measure with an almost everywhere continuous probability density f. The empirical probability measure seems inappropriate as an approximation to μ since it is atomic with mass $1/n$ at $X_1, \ldots, X_n$ and, necessarily,

$$\sup_{B \in \mathcal{B}^d} |\mu_n(B) - \mu(B)| = 1 .$$

Is there then an estimate μ_n in the absolutely continuous case for which

$$\sup_{B \in \mathcal{B}^d} |\mu_n(B) - \mu(B)| \xrightarrow{n} 0 \text{ with probability one} . \tag{2.6}$$

Suppose we choose some estimate μ_n which is absolutely continuous with a Radon-Nikodym derivative f_n so that we may think of f_n as an estimate of f from $X_1, \ldots, X_n$. Because

$$\begin{aligned} |\mu_n(B) - \mu(B)| &= \left| \int_B f_n(x)dx - \int_B f(x)dx \right| \\ &\leqq \int_B |f_n - f| dx \leqq \int_{R^d} |f_n(x) - f(x)| dx \end{aligned}$$

we conclude that (2.6) follows whenever

$$\int |f_n(x) - f(x)| dx \xrightarrow{n} 0 \text{ with probability one} . \tag{2.7}$$

From the viewpoint of estimating the law of a sequence the natural distance between f and its estimate f_n thus appears to be $\int |f_n(x) - f(x)| dx$. GLICK [2.52], extending SCHEFFÉ's theorem [2.53], showed that (2.7) follows whenever f_n is an appropriately measurable probability density for all n which is strongly consistent. In addition, he shows that if f_n is a probability density for all n which is just consistent, then

$$\int |f_n(x) - f(x)| dx \to 0 \text{ in probability} . \tag{2.8}$$

If f_n is not a probability density for each n it is also possible to give conditions which insure (2.7) or (2.8), although they are not always easy to apply (see GLICK's corollary B).

The kernel estimate of f from $X_1, \ldots, X_n$ is given by

$$f_n(x) = \sum_1^n K((x - X_i)/r_n)/nr_n^d$$

where K, the kernel, is an arbitrary probability density and $\{r_n\}$ is a sequence of positive numbers. The kernel estimate is a probability density for each n and always satisfies the measurability requirements of GLICK referred to above. Depending on the conditions put on K and $\{r_n\}$ [2.54–57] the kernel estimate is consistent or strongly consistent and (2.7) or (2.8) then follow.

One of the disadvantages of the kernel estimate of f is that the sequence $\{r_n\}$ is chosen without regard to the data. Since r_n controls the degree of smoothing

of the kernel estimate f_n about the observations $X_1, \ldots, X_n$ it seems desirable to have the data itself play a role in the amount of smoothing. If, in the deterministic discrimination problem, one lets $f_1 = f_2 = f$ then the FIX-HODGES estimate of f_1 and f_2 can be combined to yield

$$\hat{f} = \hat{f}_1 + \hat{f}_2 = \frac{k(n_1, n_2)}{nV} = \frac{k(n)}{nV} \text{ where } n_1 = [n/2],\ n_2 = n - n_1$$

and V is the volume of the smallest sphere, centered at x, which contains at least k of the n observations. In particular $\hat{f}$ is a consistent estimate of f whenever

$$k(n) \xrightarrow{n} \infty$$
$$k(n)/n \longrightarrow 0\,.$$

This result has usually been attributed to LOFTSGAARDEN and QUESENBERRY [2.58]. Additional conditions on k insure strong consistency [2.59]. The nice feature of the FIX-HODGES estimate is that the data does play a part in the smoothing of the estimate about $X_1, \ldots, X_n$. However,

$$\int_{R^d} \hat{f}(x)dx = \infty$$

for all n so that (2.6) is not possible for this estimate. WAGNER [2.60] has considered letting r_n be a function of the data in the kernel estimate of f [e.g., $r_n = r_n(X_1, \ldots, X_n)$] and has, for $d = 1$, given general conditions for consistency. One particular example of these conditions combines the features of the FIX-HODGES estimate by letting r_n be the average of the distances of X_j to its $k(n)$ nearest neighbor from $X_1, \ldots, X_n$.

Recently, spline methods have been investigated for density estimation [2.61–63]. Here, for $d = 1$, one fits a spline, usually a cubic spline, to the empirical distribution function of $X_1, \ldots, X_n$ with the derivative of the spline being the resulting density estimate. How one chooses the degree of the spline, the points for the knots and the boundary conditions are part of the investigations mentioned.

In [2.64] WAHBA establishes the best possible convergence rate for

$$E[f(x) - \hat{f}_n(x)]^2$$

where $\hat{f}_n(x)$ is any estimate of $f(x)$ and where f is assumed to belong to the Sobolev space of functions $W_p^{(m)}$. In addition, conditions are given for the various types of density estimates to achieve this rate.

2.2 Learning with Finite Memory

There was great emphasis in the early 1960's on recursive or adaptive methods for learning. The work of ABRAMSON and BRAVERMAN [2.65] initiated interest in recursive learning algorithms that were computationally simple. For example, if

$X_1, X_2, \ldots$ are i.i.d. normal random variables $N(\theta, 1)$, with unknown mean θ, and if θ has a normal prior distribution, then $p(\theta|X_1, X_2, \ldots, X_n)$ is also normal with a mean given by a simple recursive function of the sufficient statistic $\bar{X}_n = \frac{1}{n}\sum X_i$. This work was generalized by FRALICK [2.66] and imbedded in the structure of conjugate prior distributions (RAIFFA and SCHLAIFER [2.67]) by SPRAGINS [2.68]. Subsequently, however, this work was misinterpreted because of the emphasis on the interpretation that distributions $f(x|\theta)$ with finite dimensional sufficient statistics require less memory than those that do not. However, when investigated closely, this interpretation is difficult to support. While it is true, for the example, that $\bar{X}_n$ is a sufficient summary of $X_1, X_2, \ldots, X_n$ it is also true that the interleaved decimal expansion $\tilde{X}$ of $X_1, X_2, \ldots, X_n$ represents all the information in the sample, and $\tilde{X}$ and $\tilde{X}_n$ are both single real numbers.

One way to make sense out of what is meant by less memory of the past is to restrict the memory to be finite and pursue the question of which problems are easiest under this constraint. But how are we to know which of our real valued statistics are well behaved when quantized, and how should these quantizations be updated? These questions led COVER [2.69] to investigate learning algorithms of the form $T_{n+1} = f(T_n, X_{n+1}, n)$, where $T_n \in \{1, 2, \ldots, m\}$ is the current state of memory. If f is independent of n, the algorithm is said to be *time-invariant*, otherwise the algorithm is *time-varying*. The *memory size* is m, and T_n is the *state of memory* at time n. It was found that ease of learning has nothing to do with the existence of sufficient statistics.

In this section we shall survey past work on these questions and present some of the current difficulties and open problems.

We are given a sequence of independent, identically distributed observations $\{X_n\}_{n=1}^{\infty}$ where each observation X_n is drawn according to the probability measure P. There are two hypotheses H_0 and H_1 with a priori probabilities π_0 and $\pi_1 = 1 - \pi_0$, where under H_t, $P = P_t$ for $t = 0, 1$. We assume that π_0, π_1, P_0 and P_1 are known, and that $P_0 \neq P_1$.

Let $d_n \in \{H_0, H_1\}$ denote the decision made at time n. If d_n is allowed to depend on $X_1, X_2, \ldots, X_n$ then a standard likelihood ratio test yields a probability of error tending exponentially to zero in the sample size n. However, the likelihood ratio is real valued. It requires infinite memory to store it exactly. We could try to estimate the degradation introduced in this method by the use of finite memory, but we prefer to take the more fundamental viewpoint discussed below.

We shall consider algorithms $\mathscr{A}$ of the form

$$\begin{aligned} &T_n = f(T_{n-1}, X_n, n), \\ &d_n = d(T_n), \\ &T_n \in \{1, 2, \ldots, m\}, \\ &d_n \in \{H_0, H_1\}, \forall n. \end{aligned} \tag{2.9}$$

If f is a single valued mapping, then $\mathscr{A}$ is said to be a *deterministic* rule, if f is a randomized mapping, then $\mathscr{A}$ is called a *randomized* or stochastic rule.

Elementary decision theoretic considerations show that the error probability cannot be lowered by randomization in d or T_0.

The probability of error at time n is defined by

$$P(m, n) = Pr\{d(T_n) \neq H\}, \tag{2.10}$$

where H is the true hypothesis, $d(T_n)$ is the decision at time n, and m is the memory size. We shall denote by a star the optimal probability of error

$$P^*(m, n) = \inf_{\mathscr{A}} P(m, n), \tag{2.11}$$

where the class of automata $\mathscr{A}$ (e.g., time-invariant, time-varying) will be clear from the context. Finally, consider the limit for an infinite number of samples

$$P(m, \infty) = \lim_{n \to \infty} \sup P(m, n); \quad P^*(m, \infty) = \inf_{\mathscr{A}} P(m, \infty). \tag{2.12}$$

The objective is, for given m, n, P_0, P_1, π_0, π_1 to find the algorithm (f, d, T_0) which minimizes the probability of error $P(m, n)$ or $P(m, \infty)$.

Parts of the following review of the finite memory literature have been contributed by M. HELLMAN.

2.2.1 Time-Varying Finite Memory

In the infinite sample, time-varying, two hypothesis problem, COVER [2.69] has shown that a four-state memory (two bits) is sufficient to insure that the probability of error tends to zero. Basically, one bit is used to remember the current favorite hypothesis and one bit to keep track of the success or failure of test blocks which become increasingly larger. KOPLOWITZ [2.70] has recently shown that COVER's rule can be reduced to a 3-state form. He also demonstrated that an m-state, time-varying memory has an asymptotic error probability of zero for any $(m-1)$-hypothesis problem. Further, KOPLOWITZ proved that, in general, m-states are necessary for this behavior. HIRSCHLER and COVER [2.71] showed that eight states are sufficient to determine the rationality or irrationality of the parameter of a coin, given independent coin flips.

MULLIS and ROBERTS [2.72] investigated a sequential decision problem with time-varying finite memory. The cost for an observation and the cost for each type of error are variable. They find necessary conditions for an optimal design and used an iterative technique to find an approximation to the optimal rule.

WAGNER [2.73] applied time-varying rules for estimating the mean of a distribution. For Bernoulli observations WAGNER's scheme is very close to optimal since its maximum absolute error is at most $1/m$ with $2m$ states in memory.

For the finite sample problem MUISE and BOORSTYN [2.74] establish that the optimal time-varying rule essentially stores a quantized version of the likelihood ratio, although the quantization is not of any simple form. Using detectors of the form given by MUISE and BOORSTYN will result in the fastest decay of error probability with increasing sample size. The result that 4 states allows the error

probability to decay to zero cannot be (or at least to date has not been) inferred from their work.

ROBERTS and TOOLEY [2.75] attacked the problem of estimating a parameter with a time-varying finite memory. They restrict their rules to be of a special form which, although not optimal in general, does make sense (and is probably optimal) for many problems of interest.

KOPLOWITZ and ROBERTS [2.76] unified and extended this work. In particular, their demonstration of necessary and sufficient conditions for the optimal state transition function should prove valuable.

TOOLEY and ROBERTS [2.77] extended these ideas to estimating random processes with finite memory. BAXA and NOLTE [2.78] used rules similar to those of ROBERTS and TOOLEY for the detection problem. Their rules, while suboptimal, show good performance for even three bits of memory.

COVER et al. [2.79] established the existence of an optimal rule for the finite sample problem. This work also demonstrates that knowledge of the sample size can be of use in lowering the error probability. In particular, there is a problem (testing whether the bias of a coin is 10^{-10} vs. 10^{-20} with a two state memory) for which the ε-optimal infinite sample solution has $P^*(2, \infty) = 10^{-5}$, while the optimal finite sample solution has $P^*(2, n) < 2 \times 10^{-9}$ for sample sizes $n \approx 2 \times 10^{11}$. This paper then goes on to examine the structure of optimal time-varying algorithms for finite n. First it shows that the optimal rule is deterministic. The proof notes that a randomized rule can be thought of as a collection of deterministic rules indexed by a random variable ω. The error probability of the randomized rule $P_n(e)$ equals $EP_n(e|\omega)$ where the expectation is over ω. There then must be at least one value ω_0 such that $P_n(e|\omega_0) \leqq P_n(e)$. By using the deterministic time-varying rule induced by ω_0 we thus suffer no greater loss.

This reasoning fails to go through for time-invariant rules because the deterministic rule induced by ω_0 need not be time-invariant, even though the original randomized rule was. Even so, it indicates that the source of randomization need not be a true random number generator.

Still considering the time-varying problem, this paper then shows that the optimal rule is likelihood ratio in form. That is, under an appropriate renumbering of the states of memory, higher likelihood ratio observations cause transitions to higher numbered states. This result simplifies the earlier proof established by MUISE and BOORSTYN [2.74].

ROBBINS [2.80], SAMUELS [2.81], COVER [2.82], TANAKA [2.83–91] and TARUMI [2.92] have proved similar results [such as $P^*(m, \infty) = 0$, for $m \geqq 4$] for the two-armed bandit problem with a time-varying finite memory constraint.

2.2.2 Time-Invariant Finite Memory

In [2.93] HELLMAN and COVER demonstrated that

$$P^*(m, \infty) = \min \left\{ \frac{2\sqrt{\pi_0 \pi_1 \gamma^{m-1}} - 1}{\gamma^{m-1} - 1}, \pi_0, \pi_1 \right\} \tag{2.13}$$

where γ is a measure of the distance between H_0 and H_1. When $\pi_0=\pi_1=1/2$ we have

$$P^*(m,\infty)=\frac{1}{\gamma^{(m-1)/2}+1}. \tag{2.14}$$

The parameter γ is defined by

$$\gamma=\bar{l}/\underline{l}>1 \tag{2.15}$$

where $\bar{l}$ is the essential supremum on the likelihood ratio $l(x)$ and $\underline{l}$ is the essential infimum. Clearly $P_0\neq P_1$ implies $\bar{l}>1$, $\underline{l}<1$ and $\gamma>1$. Since $\gamma>1$, we see that $P^*(m,\infty)$ goes to zero exponentially in m.

The form of the optimal machine was derived in [2.93]. Here we will merely examine its structure. Let

$$\mathcal{K}_\varepsilon=\{x: l(x)\geqq[(1/\bar{l})+\varepsilon]^{-1}\} \tag{2.16}$$

and

$$\mathcal{T}_\varepsilon=\{x: l(x)\leqq \underline{l}+\varepsilon\}. \tag{2.17}$$

Thus for small ε, $\mathcal{K}_\varepsilon$ and $\mathcal{T}_\varepsilon$ have likelihood ratios close to $\bar{l}$ and $\underline{l}$, respectively. Furthermore $P_0(\mathcal{K}_\varepsilon)>0$ and $P_1(\mathcal{T}_\varepsilon)>0$ by the definitions of $\bar{l}$ and $\underline{l}$.

Consider the machine which transits from state i to $i+1$ if $X\in\mathcal{K}_\varepsilon$ and $i\leqq m-1$; from i to $i-1$ if $X\in\mathcal{T}_\varepsilon$ and $i\geqq 2$; and stays in the same state otherwise. This machine changes state only on a subsequence of high information observations, thereby making maximal use of its limited memory to store information. However, it is seen that states 1 and m are the states in which we are most certain of our decisions. Therefore once in an end state we would like the machine to stay there for a long time before leaving. Randomization achieves this.

If in state 1 and $X\in\mathcal{K}_\varepsilon$, move to state 2 with small probability δ (and stay in state 1 with probability $1-\delta$). If in state m and $X\in\mathcal{T}_\varepsilon$, move to state $m-1$ with probability $k\delta$ (and stay in state m with probability $1-k\delta$). Leave all other transitions as they were.

The purpose of not fixing $k=1$ is to allow asymmetries in the structure of the machine to compensate for asymmetries in the statistics (e.g., $\pi_0\neq\pi_1$, etc.). For symmetric problems the optimal value is $k=1$.

In [2.93] it is shown that with k properly chosen, as ε, $\delta\to 0$, the probability of error tends to $P^*(m,\infty)$, so that this is an optimal class of algorithms. The simple structure of this class is pleasing, and somewhat unexpected, since no constraints were placed on the "complexity" of the mapping f.

Randomization is generally required to achieve $P^*(m,\infty)$ [2.93]. In fact, for discrete distributions, HELLMAN and COVER [2.94] showed that there can be arbitrarily large discrepancies between the performance of randomized and deterministic rules for a fixed memory size. On the other hand, HELLMAN [2.95]

demonstrated that deterministic rules are asymptotically optimal for large memories, if memory size is measured in bits.

FLOWER and HELLMAN [2.96] examined the finite sample problem for Bernoulli observations. They found that most properties of the infinite sample solution carry over. For optimal designs, transitions are made only between adjacent states, and randomization is needed. However, in the finite sample problem randomization is needed on all transitions toward the center state (i.e., on transitions from states of low uncertainty to states with higher uncertainty). SAMANIEGO [2.97] proved that this structure is optimal for $m=3$ when attention is restricted to symmetric machines and problems.

LYNN and BOORSTYN [2.98] examined the finite sample problem for observations with continuous symmetric distributions. They calculated the probability of error for algorithms of a particular form which they call finite memory linear detectors. For this type of detector a transition occurs from state i to $i-1$ if $i \leqq m-1$ and $X_n > D$; a transition occurs from state i to $i-1$ if $i \geqq 2$ and $X_n < -D$; and the transition is from state i to itself in all other cases. The threshold D is optimized over the non-negative real line. The authors noted that this form of machine is somewhat restrictive, but that its simplicity makes it attractive. It resembles the ε-optimal solution to the infinite sample problem in many respects.

SHUBERT and ANDERSON [2.99] studied a form of generalized saturable counter and found performance to be close to optimal. The simplicity of this class of rules makes it attractive for implementation on binary data. SHUBERT [2.100] also studied a variant of the Bernoulli hypothesis testing problem in which the machine observes not only $\{X_n\}$, but also two reference sequences $\{Y_n\}$ and $\{Z_n\}$ with biases p_1 and p_2 respectively. He showed that if memory is increased by one bit then a deterministic machine can perform better than the optimal randomized machine.

SAMANIEGO [2.101] investigated the problem of estimating the parameter of a Bernoulli distribution and, restricting attention to a certain form of machine, finds minimax solutions using a variant of the mean-square-error loss criterion. If p is the true value of the parameter and $\hat{p}$ is the estimate, his loss function is $(p-\hat{p})^2/(p(1-p))$. The machine is restricted to make transitions only between adjacent states, and to move up on heads and down on tails.

HELLMAN [2.102] examined the infinite sample, Gaussian estimation problem and showed that the problem can be reduced to a quantization problem. This result also applies to a larger class of infinite sample estimation problems.

SHUBERT [2.103] has recently established some interesting results on the structure of optimal finite memory algorithms for testing k hypotheses, for $k \geqq 3$. This problem remains unsolved, but SHUBERT was able to exhibit a counter-example to the natural conjecture that the optimal algorithm has a tree structure for its transition rule. Other relevant references in finite memory include [2.14, 104–111].

The outstanding open problems in this area are

i) k-hypothesis testing with finite memory
ii) estimation with finite memory
iii) establishing optimality of likelihood ratio transition rules for finite-sample time-invariant rules.

2.3 Two-Dimensional Patterns and Their Complexity

Human visual pattern recognition deals with two- and three-dimensional scenes. What we need for a systematic investigation of statistical pattern recognition applied to two- and three-dimensional scenes are the following:

1) A systematic description of two-dimensional scenes. This description should allow for degrees of resolution. It is very clear in practice that many scenes that differ point to point may in fact be perceived by a visual system as the same scene. Thus a suitable metric on the set of all two-dimensional scenes must be developed.

2) A sequence of sequentially more refined partitions of the pattern space. For example, one standard scheme of representing a two-dimensional image is to quantize it into small squares, each square of which has associated a brightness level belonging to some finite set. Alternatively, one might have a metric defined on the pattern space and describe the scene to ε-accuracy by finitely describing some pattern that lies within ε of the true perceived pattern. Thus the representation of a pattern would consist of a canonical pattern P and a real number $\varepsilon > 0$.

3) There should be an algebra on the pattern space corresponding to the usual manipulations—union, intersection, obscuring of images, complementations, rotations, translations, etc. Some notions on developing an algebra on the space of man's senses are discussed in COVER [2.14].

4) There should be a notion of a universal intrinsic complexity of patterns. At first the idea seems absurd that a Martian, a Human, and a member of some far away galaxy would all perceive the image of a cabin by a lake with smoke coming out of the chimney as having the same intrinsic complexity. While it is certainly true that there are certain familiar stored scenes which facilitate an efficient description of the image, it turns out that this cultural bias manifests itself as an additive constant which is washed out by scenes of sufficient complexity. Indeed, using a modified notion of KOLMOGOROV-SOLOMONOFF-CHAITIN [2.112–114] complexity, in which the complexity of an image is defined to be the length of the shortest binary computer program that will cause a computer to print out the image to the desired degree of accuracy, we can prove that the intrinsic complexity of an image is universal. We shall develop some of the properties of the complexity notion for patterns.

The intrinsic complexity of patterns sits by itself as the most efficient description of the pattern, but we wish to invest this notion with further operational significance. For example, of what use is this information in allowing one to infer the classification of new, as yet unseen patterns? We find that by putting this problem into a gambling context in which one's degree of belief is immediately reflected in the amount of money gambled, that there is an absolute duality between the amount of money which can be won in sequential gambling schemes on a pattern and the number of bits of compression that can be achieved in the pattern by taking it from its raw form to its shortest program.

Furthermore, in addition to being able to make an exponential amount of money on the inference of patterns exhibited bit by bit, we find that we can also infer the classification of new patterns. We show here that the amount of money that can be made corresponds to the complexity of the classification function

which classifies each pattern into its appropriate category. So although the underlying classification function is not known, sequential inferences can be made in a universal way about this unknown classification function in such a manner that the amount of money S_n achieved after n patterns have been presented is given by $2^{n-K(f)}$, where $K(f)$ is the length in bits of the shortest binary program describing the classification function f. As a particular application, we shall show that the standard linearly separable, quadratically separable, and spherically separable pattern sets yield an inference procedure generating an amount which is given by $S_n \approx 2^{n[1-H(d/n)]}$, where d is the number of degrees of freedom of the classification surface, n is the number of points being classified, and $H(p) = -p \log p - q \log q$ is Shannon's entropy function.

Technically, the approach given in this section is not statistical in that we make deterministic statements about the intrinsic complexity of a sequence and the amount of money that a gambling scheme will surely have earned on this sequence. However, the universal gambling schemes have the property that they contain all known finitely describable statistical inference procedures as a special case. In addition, they have an optimality property which guarantees that the amount of money earned is earned at a rate which is the maximum possible, given known statistics.

2.3.1 Pattern Complexity

In this section we shall consider the intrinsic complexity of patterns and the extent to which this notion is well defined. Past work in the computer science and artificial intelligence literature on the decomposition of pictures into their basic building blocks is consistent with our motivation, but we shall look at the simplest description over all possible descriptions. Also, work on computational geometry and Perceptron complexity by MINSKY and PAPERT [2.115] is an attempt to measure picture complexity with respect to Perceptrons. We shall consider picture complexity with respect to so-called universal computers.

Kolmogorov Complexity

Let N denote the natural numbers $\{0, 1, 2, \ldots\}$. Let $\{0,1\}^*$ denote all binary sequences of finite length. Let $x \in \{0,1\}^*$ denote a finite binary sequence and let $x(n) = (x_1, x_2, \ldots, x_n)$ denote the first n terms. Let A be a partial recursive function A: $\{0,1\}^* \times N \to \{0,1\}^*$. Let $l(x)$ denote the length of the sequence x. Then $K_A(x(n)|n) = \min_{A(p,n)=x(n)} l(p)$ is the program complexity of KOLMOGOROV et al. [2.112–114]. We know that

$$\begin{aligned} &\text{i)} \quad K(x(n)|n) \leqq K_B(x(n)|n) + C_B \\ &\qquad \text{for all } n \in N, \quad \forall x, \quad \forall B \\ &\text{ii)} \quad |\{x \in \{0,1\}^* : K(x) = k\}| \leqq 2^k, \\ &\qquad \forall k \in N. \end{aligned} \tag{2.18}$$

Now we define a complexity measure for functions $f: D \to \{0, 1\}$, where the domain D is some finite set. Let A be a universal partial recursive function.

Definition. $$K_A(f|D) = \min_{\substack{A(p,x)=f(x) \\ x \in D}} l(p). \tag{2.19}$$

We assume throughout that the inputs (p, n) and (p, x) for A are suitably presented to the computer (partial recursive function) A. Thus the complexity of f given the domain D is the minimum length program p such that a Turing machine A, or equivalently a mechanical algorithm A, can compute $f(x)$ in finite time, for each $x \in D$.

Example 1. Let $D = \{0, 1\}^d$.

$$\text{Let } f(x) = \begin{cases} 1, & \sum x_i = \text{odd} \\ 0, & \sum x_i = \text{even} \end{cases}.$$

Then $K(f|D) = c$, where c is some small constant independent of d. The parity function above is easy to describe and thus has essentially zero complexity.

Pattern Complexity

Let $D = \{1, 2, \ldots, m\} \times \{1, 2, \ldots, m\}$ denote a *retina* and x denote a *pattern*, $x: D \to \{0, 1\}$. The interpretation is that the cell (i, j) of the retina has brightness level $x(i, j)$. Let A be a universal partial recursive function. We shall consider A fixed and henceforth drop the subscript in K_A.

Definition: The *pattern complexity* of the pattern x is given by

$$K(x) = \min_{\substack{A(p,(i,j))=x(i,j) \\ \forall (i,j) \in D}} l(p)$$

We have the following properties:

i) $K(x) \leqq K_B(x) + c, \quad \forall x, \quad \forall B,$
ii) $K(x) \leqq m^2 + c, \quad \forall x.$

Here are some examples without proof:

i) The blank pattern $x_0 \equiv 0$ has $K(x_0) = 0(1)$, where $0(1)$ denotes a (small) constant independent of the retina size m.

ii) The single spot pattern

$$x(i, j) = \begin{cases} 1, & (i, j) = (i_0, j_0) \\ 0, & \text{otherwise} \end{cases}$$

has

$$K(x) \leqq 2 \log m + 0(1).$$

iii) If x is the pattern corresponding to a rectangular subset of D, then $K(x) \leqq 6 \log m + 0(1)$.

iv) Let C be a circle thrown down on the retina, and let $x(i, j) = 1$ if and only if $(i, j) \in C$. Then $K(x) \leqq 6 \log m + 0(1)$.

2.3.2 Inference of Classification Functions

This section follows the presentation in COVER [2.116]. Given a domain D of patterns $D=\{x_1, x_2, \ldots, x_n\}$ and an unknown classification function $f: D \to \{0, 1\}$ assigning the patterns to two classes, we ask for an intelligent way to learn f as the correctly classified elements in D are presented one by one. We ask this question in a gambling context in which a gambler, starting with one unit, sequentially bets a portion of his current capital on the classification of the new pattern. We find the optimal gambling system when f is known a priori to belong to some family F. We also exhibit a universal optimal learning scheme achieving $\exp_2[n-K(f|D)-\log(n+1)]$ units for each f, where $K(f|D)$ is the length of the shortest binary computer program which calculates f on its domain D. In particular it can be shown that a gambler can double his money approximately $n[1-H(d/n)]$ times, if f is a linear threshold function on n patterns in d-space.

Let F denote a set of (classification) functions $f: D \to \{0, 1\}$. For example, F might be the set of all linear threshold functions. Let $|F|$ denote the number of elements in F. The interpretation will be that D is the set of patterns, and $f(x)$ is the classification of the pattern x in D.

Consider the following gambling situation. The elements of D are presented in any order. A gambler starts with one dollar. The first pattern $x_1 \in D$ is exhibited. The gambler then announces amounts b_1 and b_0 that he bets on the true class being $f(x_1)=1$ and $f(x_1)=0$, respectively. Without loss of generality we can set $b_1+b_0=1$. The true value $f(x_1)$ is then announced, and the gambler loses the incorrect bet and is paid even money on the correct bet. Thus his new capital is

$$S_1=\begin{cases}2b_1, & f(x_1)=1\\ 2b_0, & f(x_1)=0.\end{cases}$$

Now a new pattern element $x_2 \in D$ is exhibited. Again the gambler announces proportions b_1 and b_0 of his current capital that he bets on $f(x_2)=1$ and $f(x_2)=0$, respectively. Without loss of generality let $b_0+b_1=1$. Thus the bet sizes are b_1S_1 and b_0S_1. Then $f(x)$ is announced and the gambler's new capital is

$$S_2=\begin{cases}2b_1S_1, & f(x_2)=1\\ 2b_0S_2, & f(x_2)=0.\end{cases} \tag{2.20}$$

Continuing in this fashion, we define

$$b_1^{(k)}\{x_k|\{x_1, f(x_1)\}, \ldots, \{x_{k-1}, f(x_{k-1})\}\}, x_k \in D$$

and $b_0^{(k)}=1-b_1^{(k)}$, $b_0^{(k)} \geqq 0$, $b_1^{(k)} \geqq 0$ as a gambling scheme that depends only on the previously observed properly classified (training) set.

The accrued capital is

$$S_k=\begin{cases}2b_1^{(k)}S_{k-1}, & f(x_k)=1\\ 2b_0^{(k)}S_{k-1}, & f(x_k)=0\end{cases} \tag{2.21}$$

for $k=1, 2, \ldots, n$ and $S_0=1$. Let

$$b=\{\{b_0^{(1)}, b_1^{(1)}\}, \{b_0^{(2)}, b_1^{(2)}\}, \ldots, \{b_0^{(n)}, b_1^{(n)}\}\}$$

denote a sequence of gambling functions.

Theorem 1. For any $F \subseteq D^{\{0,1\}}$, there exists a gambling scheme b^* achieving $S_n(f)=S^*=2^{n-\log|F|}$ units, for all f in F and for all orders of presentation of the elements $x \in D$. Moreover, there exists no b that dominates b^* $\forall f$; thus b^* is minimax. This gambling scheme is given by the expression

$$b_1^{(k)*}(x)=\frac{|\{g \in F: g(x_i)=f(x_i), i=1, 2, \ldots, k-1, \text{ and } g(x)=1\}|}{|\{g \in F: g(x_i)=f(x_i), i=1, 2, \ldots, k-1\}|}. \tag{2.22}$$

Remark. This gambling scheme simply asserts at time k, "Bet all of the current capital on the hypotheses $f(x_k)=1$ and $f(x_k)=0$ in proportion to the number of functions g in F that *agree* on the training set and assign the new pattern x_k to classes $g(x_k)=1$ and $g(x_k)=0$, respectively".

Example: Let D denote a set of n vectors in Euclidean d-space $\mathbb{R}^d$. Let us also assume that $\{x_1, x_2, \ldots, x_n\}=D$ is in *general position* in the sense that every d-element subset of D is linearly independent. Let F be the set of all linear threshold functions on D; i.e., $f \in F$ implies there exists $w \in \mathbb{R}^d$, such that

$$f(x)=\operatorname{sgn}(w^t x-T), \quad \forall x \in D,$$

where

$$\operatorname{sgn}(t)=\begin{cases}1, & t \geqq 0\\ 0, & t<0.\end{cases}$$

Then from COVER [2.42], we have $|F|=2\sum_{k=0}^{d}\binom{n-1}{k}$, $\forall d, n$. Using bounds derived from Stirling's approximation, it can be shown that

$$\log\left(2\sum_{k=0}^{d}\binom{n-1}{k}\right) \approx nH\left(\frac{d}{n}\right), \quad \text{for} \quad n \geqq 2d. \tag{2.23}$$

Thus we conclude, for $n \geqq 2d$, that an amount $S_n=2^{n[1-H(d/n)]}$ can be won if in fact the n patterns are linearly separable in $\mathbb{R}^d$. Note also that $H(d/n)$ is the Kolmogorov complexity of most of the linear threshold functions $f \in F$. Finally, we observe that S_n is not much greater than 1 until $n \geqq 2d$, at which point S_n grows exponentially. This is yet more evidence that $n=2d$ is a natural definition of the capacity of a linear threshold pattern recognition device with d variable weights [2.42]. This result is a special case of the following theorem:

Theorem 2: There exists a betting scheme b^* such that the total accumulated capital satisfies $S(f) \geqq 2^{n-K(f|D)-\log(n+1)}$.

Other aspects of complexity and inference can be found in COVER [2.117].

References

2.1 E. FIX, J. L. HODGES: "Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties", Rep. 4, Project no. 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas (1951)
2.2 T. COVER, P. HART: IEEE Trans. Information Theory IT-13, 21 (1967)
2.3 E. FIX, J. L. HODGES: "Discriminatory Analysis. Nonparametric Discrimination: Small Sample Performance", Rep. 11, Project no. 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas (1952)
2.4 M. G. KENDALL: Discrimination and Classification, in *Proc. Intern. Symp. Multivariate Analysis*, ed. by P. R. KRISHNAIAH (Academic Press, New York 1966)
2.5 C. P. QUESENBERRY, M. P. GESSAMAN: Ann. Math. Statist. 39, 664 (1968)
2.6 M. W. ANDERSON, R. D. BENNING: IEEE Trans. Information Theory IT-16, 541 (1970)
2.7 G. W. BEAKLEY, F. B. TUTEUR: IEEE Trans. Computers C-21, 1337 (1972)
2.8 E. G. HENRICHON, K. S. FU: IEEE Trans. Computers C-19, 362 (1970)
2.9 E. G. HENRICHON, K. S. FU: IEEE Trans. Computers C-18, 614 (1969)
2.10 T. W. ANDERSON: Some Nonparametric Multivariate Procedures Based on Statistically Equivalent Blocks, in *Proc. Intern. Symp. Multivariate Analysis*, ed. by P. R. KRISHNAIAH (Academic Press, New York 1966)
2.11 J. OWEN, D. B. BRICK, E. A. HENRICHON: Pattern Recognition 2, 227 (1970)
2.12 M. P. GESSAMAN, P. H. GESSAMAN: J. Am. Stat. Assoc. 67, 468 (1972)
2.13 J. VAN RYZIN: Ann. Math. Statist. 37, 976 (1966)
2.14 T. COVER: Learning in Pattern Recognition, in *Methodologies of Pattern Recognition*, ed. by S. WATANABE (Academic Press, New York 1969)
2.15 H. ROBBINS: Ann. Math. Statist. 35, 1 (1964)
2.16 J. VAN RYZIN: Sankhya Ser. A 28, 261 (1966)
2.17 S. C. SCHWARTZ: Convergence of Risk in Adaptive Pattern Recognition Procedures, in Proc. 5th Allerton Conf. Circuit and System Theory, (1967) pp. 800–807
2.18 N. GLICK: J. Am. Stat. Assoc. 67, 116 (1972)
2.19 E. HEWITT, J. L. SAVAGE: Trans. Am. Math. Soc. 80, 470 (1955)
2.20 L. BREIMAN: *Probability* (Addison-Wesley, Reading, Mass. 1968)
2.21 J. VAN-RYZIN: J. Math. Anal. Appl. 20, 359 (1967)
2.22 C. T. WOLVERTON, T. J. WAGNER: IEEE Trans. Information Theory IT-15, 258 (1969)
2.23 L. REJTÖ, P. RÉVÉSZ: Problems of Control and Information Theory 2, 67 (1973)
2.24 C. R. PELTO: Technometrics 11, 775 (1969)
2.25 T. J. WAGNER: IEEE Trans. Information Theory IT-17, 566 (1971)
2.26 J. FRITZ: IEEE Trans. Information Theory (Sept. 1975)
2.27 D. L. WILSON: IEEE Trans. Systems Man Cybernetic SMC-2, 408 (1972)
2.28 P. HART: IEEE Trans. Information Theory IT-14, 515 (1968)
2.29 G. W. GATES: IEEE Trans. Information Theory IT-18, 431 (1972)
2.30 J. R. ULLMAN: IEEE Trans. Information Theory IT-20, 541 (1974)
2.31 C. W. SWONGER: Sample Set Condensation for a Condensed Nearest Rule for Pattern Recognition, in *Frontiers of Pattern Recognition*, ed. by S. WATANABE (Academic Press, New York 1972)
2.32 C. L. CHANG: IEEE Trans. Computers C-23, 1179 (1974)
2.33 M. E. HELLMAN: IEEE Trans. Systems Science Cybernetic SSC-6, 179 (1970)
2.34 C. K. CHOW: IEEE Trans. Information Theory IT-16, 41 (1970)
2.35 C. K. CHOW: IRE Trans. Electronic Computers EC-6, 247 (1957)
2.36 A. DVORETZKY, J. KIEFER, J. WOLFOWITZ: Ann. Math. Statist 27, 642 (1956)
2.37 G. T. TOUSSAINT: IEEE Trans. Information Theory IT-20, 472 (1974)
2.38 L. KANAL: IEEE Trans. Information Theory IT-20, 697 (1974)
2.39 S. C. FRALICK, R. W. SCOTT: IEEE Trans. Information Theory IT-17, 440 (1971)
2.40 W. H. ROGERS, T. J. WAGNER: Ann. Statist (1974), submitted
2.41 J. KIEFER, J. WOLFOWITZ: Trans. Am. Math. Soc. 87, 173 (1958)
2.42 T. COVER: IEEE Trans. Electronic Computers IT-10, 618 (1965)
2.43 D. H. FOLEY: IEEE Trans. Information Theory IT-18. 618 (1972)
2.44 T. J. WAGNER: Ann. Statist 1, 359 (1973)
2.45 K. FUKUNAGA, D. KESSELL: IEEE Trans. Information Theory IT-19, 434 (1973)

2.46 K. FUKUNAGA, L. HOSTETLER: IEEE Trans. Information Theory, IT-21, 285 (1975)
2.47 C. T. WOLVERTON: IEEE Trans. Information Theory IT-18, 119 (1972)
2.48 E. J. WEGMAN: Technometrics 14, 533 (1972)
2.49 E. J. WEGMAN: J. Statist. Comp. Simulation 1, 225 (1972)
2.50 T. COVER: A Hierarchy of Probability Density Function Estimates, in *Frontiers of Pattern Recognition*, ed. by S. WATANABE (Academic Press, New York 1972)
2.51 R. R. RAO: Ann. Math. Statist. 33, 659 (1962)
2.52 N. GLICK: Utilitas Math. 6, 61 (1974)
2.53 H. SCHEFFÉ: Ann. Math. Statist. 18, 434 (1947)
2.54 E. PARZEN: Ann. Math. Statist. 33, 1065 (1947)
2.55 T. CACOULLOS: Ann. Instit. Statist. Math. 18, 178 (1966)
2.56 E. A. NADARAYA: Theory Prob. Appl. 10, 186 (1965)
2.57 J. VAN RYZIN: Ann. Math. Statist 40, 1765 (1969)
2.58 D. O. LOFTSGAARDEN, C. P. QUESENBERRY: Ann. Math. Statist. 38, 1261 (1965)
2.59 T. J. WAGNER: IEEE Trans. Systems Man Cybernetic SMC-3, 289 (1973)
2.60 T. J. WAGNER: IEEE Trans. Information Theory IT-21, 438 (1974)
2.61 L. BONEVA, D. KENDALL, I. STEFANOV: J. Roy. Statist. Soc. 33, 1 (1971)
2.62 G. WAHBA: "Interpolating Spline Methods for Density Estimation II. Variable Knots", Tech. Rep. # 337, Dept. of Statistics, University of Wisconsin, Madison, Wisconsin (1973)
2.63 G. WAHBA: Ann. Statist. 3, 15 (1975)
2.64 G. WAHBA: Ann. Statist. 3, 30 (1975)
2.65 N. ABRAMSON, D. BRAVERMAN: IRE Trans, Information Theory IT-8, 58 (1962)
2.66 S. C. FRALICK: "The Synthesis of Machines Which Learn Without a Teacher", Techn. Rep. no. 6103-8, Stanford Electronics Labs., Stanford, California (1964)
2.67 H. RAIFFA, R. SCHLAIFER: *Applied Statistical Decision Theory* (Harvard University Press, Boston, Mass. 1961)
2.68 J. SPRAGINS: IEEE Trans. Information Theory IT-12, 223 (1966)
2.69 T. M. COVER: Ann. Math. Statist. 40, 828 (1969)
2.70 J. KOPLOWITZ: IEEE Trans. Information Theory IT-21, 44 (1975)
2.71 T. M. COVER, P. HIRSCHLER: Ann. Statist 3, 939 (1975)
2.72 C. T. MULLIS, R. A. ROBERTS: IEEE Trans. Information Theory IT-20, 440 (1974)
2.73 T. J. WAGNER: IEEE Trans. Information Theory IT-18, 523 (1972)
2.74 R. W. MUISE, R. R. BOORSTYN: "Detection with Time-Varying Finite-Memory Receivers", Abstracts of papers, 1972 IEEE Intern. Symp. Information Theory, (1972)
2.75 R. A. ROBERTS, J. R. TOOLEY: IEEE Trans. Information Theory IT-16, 685 (1970)
2.76 J. KOPLOWITZ, R. ROBERTS: IEEE Trans. Information Theory IT-19, 631 (1973)
2.77 J. R. TOOLEY, R. ROBERTS: IEEE Trans. Systems Man Cybernetic SMC-3, 294 (1973)
2.78 E. G. BAXA, L. W. NOLTE: IEEE Trans. Systems Man Cybernetics SMC-2, 42 (1972)
2.79 T. COVER, M. FREEDMAN, M. HELLMAN: Information and Control (to be published)
2.80 H. ROBBINS: Proc. Nat. Acad. Sci. 42, 920 (1956)
2.81 S. M. SAMUELS: Ann. Math. Statist. 39, 2103 (1968)
2.82 T. M. COVER: Information and Control 12, 371 (1968)
2.83 K. TANAKA: Bull. Math. Statist. 14, 31 (1970)
2.84 K. TANAKA: Bull. Math. Statist. 14, 61 (1970)
2.85 K. TANAKA: Mathematics 24, 249 (1970)
2.86 K. TANAKA: Bull. Math. Statist. 14, 13 (1971)
2.87 K. TANAKA, E. ISOGAI: Tamkang J. Math.
2.88 K. TANAKA, K. INADA, S. IWASE: Tamkang J. Math. (to be published)
2.89 K. TANAKA, K. INADA: "Some Extension of the Two-Armed Bandit Problem with Finite Memory", Sci. Rep. Niigata Univ., Series A, 10, 5 (1973)
2.90 K. TANAKA, K. INADA: "Some Statistical Method with Finite Memory", Sci. Rep. Niigata Univ., Series A, 10, 27 (1973)
2.91 K. TANAKA, E. ISOGAI, S. IWASE: "On a Sequential Procedure with Finite Memory for Testing Statistical Hypotheses", Sci. Rep. Niigata University, Series A, 11, 31 (1974)
2.92 T. TARUMI: "Estimation of the Direction of a Bivariate Normal Mean with Finite Memory", Memoirs of the Faculty of Science, Kyushu University, Series A, Mathematics 26, 351 (1972)
2.93 T. COVER, M. HELLMAN: Ann. Math. Statist. 41, 765 (1970)

2.94 T. COVER, M. HELLMAN: Ann. Math. Statist. **42**, 1075 (1971)
2.95 M. E. HELLMAN: IEEE Trans. Information Theory IT-**18**, 499 (1972)
2.96 R. A. FLOWER, M. E. HELLMAN: IEEE Trans. Information Theory IT-**18**, 429 (1972)
2.97 F. SAMANIEGO: IEEE Trans. Information Theory IT-**20**, 387 (1974)
2.98 P. F. LYNN, R. BOORSTYN: Bounds on Finite Detectors, presented at 1972 IEEE Intern. Symp. Information Theory, Asilomar, California (1972)
2.99 B. SHUBERT, C. ANDERSON: IEEE Trans. Information Theory IT-**19**, 644 (1973)
2.100 B. SHUBERT: IEEE Trans. Information Theory IT-**2**, 384 (1974)
2.101 F. SAMANIEGO: IEEE Trans. Information Theory IT-**19**, 636 (1973)
2.102 M. E. HELLMAN: IEEE Trans. Information Theory IT-**20**, 382 (1974)
2.103 B. SHUBERT: "Some Remarks on the Finite-Memory *K*-Hypotheses Problems", Techn. Rep. NPS55Sy74101, Naval Postgraduate School, Monterey, California (1974)
2.104 M. L. TSETLIN: Avtomat. i Telemeh. **22**, 1345 (1961) (Available in English translation)
2.105 V. Y. KRYLOV: Avtomat. i Telemeh. **24**, 1226 (1963) (Available in English translation)
2.106 V. I. VARSHAVSKII, I. P. VORONTSOVA: Avtomat. i Telemeh. **24**, 327 (1963) (Available in English translation)
2.107 K. S. FU, T. J. LI: "On the Behavior of Learning Automato and its Applications", Purdue University Techn. Rep. no. TR-EE 68-20 (1968)
2.108 B. CHANDRASEKARAN, D. W. S. SHEN: IEEE Trans. Systems Science Cybernetic SSC-**4** (1968)
2.109 T. COVER, M. HELLMAN: IEEE Trans. Information Theory IT-**16**, 185 (1970)
2.110 P. F. LYNN: Finite Memeory Detectors; Ph. D. Thesis, Dept. of Electr. Engg., Polytechnic Inst. of Brooklyn (1971)
2.111 C. T. MULLIS: A Class of Finite Memory Decision Processes; M. S. Thesis, Dept. Electr. Engg., Univ. of Colorado (1968)
2.112 A. N. KOLMOGOROV: Problemy Peredaci Informaccii **1**, 3 (1965)
2.113 R. J. SOLOMONOFF: Information and Control **7**, 1, 224 (1964)
2.114 G. J. CHAITIN: J. Assoc. Comput. Mach. **13**, 547 (1966)
2.115 M. MINSKY, S. PAPERT: *Perceptrons* (MIT Press, Cambridge, Mass 1969)
2.116 T. COVER: "Generalization on Patterns Using Kolmogorov Complexity", Proc. 1st Intern. Joint Conf. Pattern Recognition, Washington, C. D. (1973)
2.117 T. COVER: Ann. Statist (1974), submitted