A MATHEMATICAL THEORY OF PATTERN RECOGNITION[1]

By Arthur Albert

The Arcon Corporation, Lexington, Mass.

1. Summary. Let X_0 and X_1 be (unknown) disjoint subsets of a Hilbert space H, such that the convex hulls of X_0 and X_1 are a positive distance apart. Suppose that samples are drawn independently and at random from $X_0 \cup X_1$. After the nth sample, Z_n, it is required to guess whether Z_n came from X_0 or X_1. After each guess, we are told whether we were right or wrong.

In this paper, a decision procedure is exhibited, having the property that the probability of making an error on the nth trial converges to zero with increasing n. Furthermore, the guessing rule used on the $n + 1$st trial depends on the past data only through the rule used on the nth trial, the value of Z_n, and whether or not the guess about Z_n was correct.

The application to pattern recognition problems of a dichotomous sort is immediate when we identify X_0 and X_1 with two classes of patterns which are observed in temporal succession. The rules for membership in X_0 and X_1 are not known, but we (or a machine) are/is told to which class each pattern belongs, *after* making a guess about that pattern. As the "training period" increases, errors are made with ever decreasing frequency.

2. Mathematical formulation. Representative samples are to be drawn independently and at random from each of two classes (call the classes "class zero" and "class one"). Associated with each phenomenon to be classified is a measurable attribute. The set of attributes corresponding to members of "class zero" will be denoted by X_0, while those corresponding to members of "class one" will be denoted by X_1. Suppose for the moment that X_0 and X_1 are non-overlapping sets and that there is a real valued function defined over $X_0 \cup X_1$ having the property that, for some constant, c,

$$\inf_{x \varepsilon X_0} f(x) > c > \sup_{x \varepsilon X_1} f(x).$$

If we did not know the rule for membership in X_0 or X_1, but we did know f and c, we could still carry out the decision procedure viz: After each observation, apply f to the attribute x. If $f(x) > c$, decide $x \varepsilon X_0$ (and thus the phenomenon belongs to "class zero"). Otherwise, decide $x \varepsilon X_1$.

Suppose f and c are not known: If we are told after each decision whether the attribute really belonged in X_0 or X_1, it seems reasonable to hope that, under certain circumstances, this information can be utilized in some way to make estimates of f and c. If we are clever enough to construct estimates of f and c that converge, in an appropriate sense, to the true values as more data

Received January 22, 1962.

[1] This work was initiated during the summer of 1961, while the author was employed at Laboratory for Electronics, Brighton, Massachusetts.

accumulates, we could base our decision on these estimates and make errors occur with ever decreasing frequency.

The key to the whole problem then, centers upon finding an iterative scheme for estimating f and c. Of course, in order to do so, we must add some structure to the problem.

We do so by assuming that $X_0 \cup X_1$ is a subset of a finite dimensional Euclidean space. Let $H(X_0 \cup X_1)$ be the smallest subspace containing $X_0 \cup X_1$. Assume further, that the function f, which serves to separate X_0 and X_1, is a continuous linear functional over $H(X_0 \cup X_1)$.

It should be clear by now that the phenomena can be identified with their attributes, insofar as the classification problem is concerned, provided we have been smart enough (or lucky enough) to pick the proper attributes.

A brief word should be said about the nature of the functional f. Defined[2] over $H(X_0 \cup X_1) \otimes H(X_0 \cup X_1)$ is an inner product $(\cdot, \cdot)$ taking on real values, satisfying the following conditions:

(1) $(x_1, x_2) = (x_2, x_1)$,

(2) $(x, x) > 0$ unless $x = 0$,

(3) $(\alpha x_1 + \beta x_2, x_3) = \alpha(x_1, x_3) + \beta(x_2, x_3)$.

The inner product induces a norm (or distance function) over $H(X_0 \cup X_1)$, viz: $\|x_1 - x_2\| = (x_1 - x_2, x_1 - x_2)^{\frac{1}{2}}$.

There is a well-known result of analysis called the Reisz Representation Theorem. As applied here, we find that for every continuous linear functional f over $H(X_0 \cup X_1)$, there is a $\xi_f \varepsilon H(X_0 \cup X_1)$ such that for every $x \varepsilon H(X_0 \cup X_1)$, $f(x) = (\xi_f, x)$.

In short, every continuous linear functional can be represented as an inner product with respect to some fixed element of H.

For our purpose then, it suffices to try to find a $\xi \varepsilon H$ and a constant c, such that

$$\text{(a)} \qquad \inf_{x \varepsilon X_0} (\xi, x) > c > \sup_{x \varepsilon X_1} (\xi, x).$$

The set of points in H for which $(\xi, x) = c$, is a hyperplane. When ξ and c are such that (a) holds, we say that there is a hyperplane which strictly separates X_0 and X_1. When we formulate our iterative scheme, we will want to include an assumption strong enough to guarantee the existence of a hyperplane which will strictly separate X_0 and X_1.

3. The assumptions. The assumptions are divided into three groups. Assumptions A_1 and A_2 are there to guarantee the existence of a hyperplane which strictly separates X_0 and X_1. Assumption B merely eliminates some trivial cases from consideration while C_1 and C_2 guarantee that errors will occur with ever-decreasing frequency.

[2] We assume for simplicity that $H(X_0 \cup X_1)$ is real. The results we obtain are all true in the complex case.

ASSUMPTION A_1. $X_0 \cup X_1$ is a subset of a finite dimensional Euclidean space with an inner product defined over it.

ASSUMPTION A_2. The distance between the convex hulls of X_0 and X_1 is positive. (The convex hull of a set, X, is the smallest convex set containing X.)

DEFINITION. Let P_1 be the class of probability measures over the Borel subsets of $H(X_0 \cup X_1)$ satisfying

(a) $\Pr[X_0 \cup X_1] = 1$,

(b) $\Pr[X_0] \Pr[X_1] > 0$,

(c) $\Pr[B] > 0$ for every open subset, B, of $H(X_0 \cup X_1)$ whose intersection with $X_0 \cup X_1$ is not empty.

ASSUMPTION B. Observations are chosen independently and at random in such a way that the probability measure induced over the attribute space is a member of P_1.

This assumption (aside from the independence clause) excludes the trivial case where all observations come from the same class. It also prevents some needless confusion about certain sets of measure zero.

DEFINITION.

(a) Let $U = \{u \,\varepsilon\, H(X_0 \cup X_1): \|u\| = 1 \text{ and } \inf_{x_0 \varepsilon X_0, x_1 \varepsilon X_1} (u, x_0 - x_1) \geqq 0\}$.

(b) Let $Q_1(u) = \{x \,\varepsilon\, X_1 : \inf_{x_1 \varepsilon X_1} (u, x - x_1) \geqq 0\}$,

$Q_0(u) = \{x \,\varepsilon\, X_0 : \sup_{x_0 \varepsilon X_0} (u, x - x_0) \leqq 0\}$,

$Q(u) = Q_0(u) \cup Q_1(u)$.

(c) Let P_2 be that subset of P_1 for which $\Pr\{\bigcup_{u \varepsilon U} Q(u)\} = 0$.

ASSUMPTION C_1. Observations are chosen independently and at random in such a way that the probability measure induced over the attribute space is a member of P_2.

ASSUMPTION C_2. $X_0 \cup X_1$ is bounded.

Assumption C_1 is a sort of absolute continuity restriction over the boundaries of X_0 and X_1. Since U is the set of all hyperplanes which separate X_0 and X_1, it is easy to see that $Q(u)$ is the set of points in the closure of $X_0 \cup X_1$ at which a supporting hyperplane with normal $u \,\varepsilon\, U$ can be constructed. $\bigcup_{u \varepsilon U} Q(u)$ can be thought of as a subset of the frontiers of X_0 and X_1; this set must have measure zero. This condition will be satisfied in particular if the probability measure on $X_0 \cup X_1$ is absolutely continuous with respect to Lebesgue measure over $H(X_0 \cup X_1)$.

Assumption C_2 may not be essential, but is included for aesthetic reasons.

With all this behind us, we are prepared to reveal the estimation procedure.

4. The procedure. A sequence of real world phenomena occur in temporal succesion in such a way that the associated sequence of measured attributes, $Z_0, Z_1, \cdots$ can be thought of as independent stochastic variables drawn from $X_0 \cup X_1$.

After observing Z_j, we are to guess whether the associated phenomenon came from class zero or class one, or equivalently, whether Z_j came from X_0 or X_1. After each decision, we are told whether we were right or wrong.

Let N be the smallest integer k for which the event $[Z_k \varepsilon X_0$ and $Z_{k+1} \varepsilon X_1$, or $Z_k \varepsilon X_1$ and $Z_{k+1} \varepsilon X_0]$ occurs.

The guessing rule.

(1) Decide $Z_0 \varepsilon X_0$.

(2) If $k \leqq N + 1$, decide that Z_k comes from the same set as Z_{k+1}. Thereafter, the decision rule depends upon two variables, ξ_n and w_n, to be specified shortly:

(3) For $j \geqq N + 1$, decide $Z_{j+1} \varepsilon X_0$ if $(\xi_j, Z_{j+1} - w_j) \geqq 0$, $Z_{j+1} \varepsilon X_1$ otherwise. (Since we will be concerned with what happens for $k > N$, we can assume without loss of generality, that $N = 0$.)

The idea behind this rule is simple enough, once ξ_j and w_j are specified: If Z_{j+1} lies above the hyperplane with normal ξ_j, passing through w_j, we decide Z_{j+1} comes from X_0. Otherwise, we decide Z_{j+1} comes from X_1.

The parameters ξ_j and w_j are defined recursively: If we guess correctly about Z_{j+1}, we take $\xi_{j+1} = \xi_j$ and $w_{j+1} = w_j$. The hyperplane is left unchanged in this case. If w_j and z_{j+1} are in different sets and we guess wrong, ξ_j is changed to ξ_{j+1} and the new plane passes through Z_{j+1}. If w_j and Z_{j+1} are in the same set and we guess wrong, the plane is translated till it passes through Z_{j+1}. This procedure can be stated formally as follows:

The estimation rule.

$$w_{j+1} = \begin{cases} w_j \text{ if } Z_{j+1} \varepsilon X_1 \text{ and } (\xi_j, Z_{j+1} - w_j) \leqq 0 \\ \quad \text{or } Z_{j+1} \varepsilon X_0 \text{ and } (\xi_j, Z_{j+1} - w_j) \geqq 0, \\ Z_{j+1} \quad \text{otherwise.} \end{cases}$$

$$w_1 = Z_1.$$

$$\xi_{j+1} = \begin{cases} \xi_j - 2(Z_{j+1} - w_j, \xi_j)[Z_{j+1} - w_j]/\|Z_{j+1} - w_j\|^2 \\ \quad \text{if } Z_{j+1} \varepsilon X_0, w_j \varepsilon X_1 \text{ and } (\xi_j, Z_{j+1} - w_j) < 0 \\ \quad \text{or } Z_{j+1} \varepsilon X_1, w_j \varepsilon X_0 \text{ and } (\xi_j, Z_{j+1} - w_j) > 0, \\ \xi_j \quad \text{otherwise.} \end{cases}$$

$$\xi_1 = \begin{cases} Z_1 - Z_0/\|Z_1 - Z_0\| & \text{if } Z_1 \varepsilon X_0, \\ Z_0 - Z_1/\|Z_1 - Z_0\| & \text{if } Z_1 \varepsilon X_1. \end{cases}$$

Denote by $\pi(\xi, w)$, the hyperplane with unit normal ξ, passing through w. In the case where $\xi_n = \xi_{n+1}$, and $w_{n+1} \neq w_n$, $\pi(\xi_{n+1}, w_{n+1})$ is obtained from $\pi(\xi_n, w_n)$ by a translation. If however, $\xi_{n+1} \neq \xi_n$, the estimation rule dictates that ξ_{n+1} be obtained by adding a multiple of $Z_{n+1} - w_n$ to ξ_n. The multiple is chosen so that $\|\xi_{n+1}\| = \|\xi_n\|$. In this case, $\pi(\xi_{n+1}, w_{n+1})$ is obtained from $\pi(\xi_n, w_n)$ by first rotating ξ_n, then translating $\pi(\xi_{n+1}, w_n)$ until it coincides with $\pi(\xi_{n+1}, w_{n+1})$. We will show that under the guessing rule and decision rule stated above, the probability of misclassifying the nth observation converges to zero as n grows large.

The problem of finding an iterative scheme for separating two convex sets has been treated by Rosenblatt, Block and Josephs in the Perceptron Papers [3] for the case where the convex sets in question are generated by a finite number of points. Since assumption C_1 appears to restrict us to infinite sets, the present method seems to be complementary.

The theorems proved in the following section depend, to various degrees, upon a succession of lemmas. For the sake of the reader who is not primarily interested in analytic acrobatics, these lemmas are proved in an appendix. The logical dependence of theorems upon lemmas and lemmas upon theorems is stated at the end of each proof. For example, if at the end of a proof we write L3, L4, T2, it means that the assertion in question depends upon Lemmas 3 and 4 along with Theorem 2. If things were written in the proper logical order, they would occur as: L1, L2, L3, L4, L5, L6, L7, T1, T2, T3, T4, T5.

5. The main results.

DEFINITION. Let $U = \{u: \|u\| = 1 \text{ and } \inf_{x_0 \varepsilon X_0, x_1 \varepsilon X_1} (u, x_0 - x_1) \geqq 0\}$. ($U$ is the collection of unit normals corresponding to hyperplanes which separate X_0 and X_1.)

THEOREM 1. *With probability* 1, $0 \leqq (u, \xi_n) \leqq (u, \xi_{n+1}) \leqq 1$ *for all* $u \varepsilon U$ *and all* $n \geqq 1$.

PROOF. By construction, $\|\xi_n\| = \|\xi_{n+1}\|$ for every n. Since $\|\xi_1\| = 1$, we see that $(\xi_n, u) \leqq \|\xi_n\| \cdot \|u\| = 1$ for all $u \varepsilon U$. ξ_{n+1} is always of the form:

$$\xi_{n+1} = \xi_n - 2s_{n+1},$$

where

$$(s_{n+1}, u) \leqq 0 \quad \text{for every} \quad u \varepsilon U.$$

Thus,

$$(\xi_{n+1}, u) \geqq (\xi_n, u) \quad \text{for every } u \varepsilon U.$$

Since $(\xi_1, u) \geqq 0$, the conclusion follows.

From the monotonicity and boundedness of the sequence $\{(\xi_n, u)\}$, we deduce immediately that:

THEOREM 2.

(a) *For every* $u \varepsilon U$, $W(u) = \lim_{n\to\infty} (\xi_n, u)$ *exists with probability one.*

(b) $\sup_{u \varepsilon U} W(u) \leqq 1$. (T1)

COROLLARY.

(a) $\sup_{u\varepsilon U} \sup_{\delta>0} \lim_{n\to\infty} \Pr\{\sup_{k\geqq n} (\xi_{k+1} - \xi_k, u) > \delta\} = 0$.

(b) $\sup_{u\varepsilon U} \sup_{\delta>0} \limsup_{n\to\infty} \Pr\{(\xi_{n+1} - \xi_n, u) > \delta\} = 0$.

PROOF. The monotone convergence of (ξ_n, u) assures that $(\xi_{n+1} - \xi_n, u)$ converges to zero with probability one. This is the content of part (a). The second assertion follows since $[(\xi_{n+1} - \xi_n, u) > \delta] \subset [\sup_{k\geqq n} (\xi_{k+1} - \xi_k, u) > \delta]$.

In fact, ξ_n converges to some element in U with probability one:

THEOREM 3. (a) $\Pr[\sup_{u\varepsilon U} W(u) = 1] = 1$. (b) $\Pr\{\bigcup_{u\varepsilon U} [\xi_n \to u]\} = 1$.

PROOF. (a) Suppose not. Then for some $\sigma > 0$, and some $\epsilon > 0$,

$$\Pr[\sup_{u \varepsilon U} W(u) \leqq 1 - \sigma] \geqq \epsilon.$$

Thus,

$$\Pr[\xi_n \,\varepsilon\, V(\sigma)] \geqq \epsilon \quad \text{for all} \quad n,$$

where

$$V(\sigma) = \{\xi: \|\xi\| = 1 \quad \text{and} \quad \sup_{u\varepsilon U}(u, \xi) \leqq 1 - \sigma\}.$$

From Lemma 7, $\alpha > 0$ can be chosen so that for some μ,

(3.1)
$$\Pr[\langle Z_{n+1}, Z_{n+2}\rangle \epsilon S^*(\xi, \alpha) \mid \xi_n = \xi, w_n = w] = \Pr[\langle Z_{n+2}, Z_{n+1}\rangle \epsilon S^*(\xi, \alpha) \mid \xi_n = \xi, w_n = w] \geqq \mu > 0$$

for all $\xi \,\varepsilon\, V(\sigma)$ and all $w \,\varepsilon\, X_0 \cup X_1$ and all n. (μ does not depend upon n, w or ξ. Here,

$$S^*(\xi, \alpha) = (X_1 \otimes X_0) \cap \{\langle z, z'\rangle \,\varepsilon\, H(X_0 \cup X_1) \otimes H(X_0 \cup X_1) : (z - z', \xi)/\|z - z'\|^2 > \alpha\}.)$$

Lemma 2 guarantees that for some $u^* \,\varepsilon\, U$ and some $\delta > 0$

$$\inf_{z'\varepsilon X_0, z\varepsilon X_1} (u^*, z' - z) \geqq \delta.$$

Now we assert that if $\xi_n = \xi \,\varepsilon\, V(\sigma)$ and $w_n = w \,\varepsilon\, X_1$, then

$$[\langle Z_{n+1}, Z_{n+2}\rangle \,\varepsilon\, S^*(\xi, \alpha)] \subset [(\xi_{n+2} - \xi_{n+1}, u^*) > \gamma]$$

where

$$\gamma = 2\alpha\delta \inf_{x\varepsilon X_0, y\varepsilon X_1} \|x - y\|^2 / \sup_{x\varepsilon X_0, y\varepsilon X_1} \|x - y\|^2 > 0.$$

Case 1. $(Z_{n+1} - w, \xi) \leqq 0$. In this case, no error is made on trial $n + 1$, so $\xi_{n+1} = \xi_n$ and $w_{n+1} = w_n$. Since $\langle Z_{n+1}, Z_{n+2}\rangle \,\varepsilon\, S^*(\xi, \alpha)$, it follows that $(Z_{n+1} - Z_{n+2}, \xi_n) = (Z_{n+1} - w_n, \xi_n) - (Z_{n+2} - w_n, \xi_n) = (Z_{n+1} - w_n, \xi_n) - (Z_{n+2} - w_{n+1}, \xi_n) > 0$. Hence, $(Z_{n+2} - w_{n+1}, \xi_{n+1}) < 0$, so that a mistake is made on trial $n + 2$. Since Z_{n+2} and w_{n+1} are in different sets, we must take $\xi_{n+2} = \xi_{n+1} - 2(Z_{n+2} - w_n, \xi_n)[Z_{n+2} - w_n]/\|Z_{n+2} - w_n\|^2$. Thus,

$$(\xi_{n+2} - \xi_{n+1}, u^*) = \frac{2\|Z_{n+2} - Z_{n+1}\|^2}{\|Z_{n+2} - w_n\|^2} \frac{(Z_{n+2} - w_n, u^*)}{\|Z_{n+2} - Z_{n+1}\|^2} [(Z_{n+1} - Z_{n+2}, \xi_n) + (w_n - Z_{n+1}, \xi_n)].$$

Since,

$$(Z_{n+1} - w_n, \xi_n) \leqq 0,$$

$$(\xi_{n+2} - \xi_{n+1}, u^*) \geqq 2\rho\delta(Z_{n+1} - Z_{n+2}, \xi_n)/\|Z_{n+1} - Z_{n+2}\|^2 \geqq 2\rho\alpha\delta,$$

where

$$0 < \rho = \inf_{x \varepsilon X_0, y \varepsilon X_1} \|x - y\|^2 / \sup_{x \varepsilon X_0, y \varepsilon X_1} \|x - y\|^2 \leqq 1.$$

Case 2. $(Z_{n+1} - w, \xi) > 0$. In this case, an error is made on trial $n + 1$. Since Z_{n+1} and w_n are in the same set, we take $\xi_{n+1} = \xi_n$ and $w_{n+1} = Z_{n+1}$. On trial $n + 2$, $(Z_{n+2} - w_{n+1}, \xi_{n+1}) = (Z_{n+2} - Z_{n+1}, \xi) < 0$, since

$$\langle Z_{n+1}, Z_{n+2} \rangle \varepsilon S^*(\xi, \alpha).$$

Thus, another mistake is made. This time,

$$\xi_{n+2} = \xi_{n+1} - 2(Z_{n+2} - Z_{n+1}, \xi)[Z_{n+2} - Z_{n+1}]/\|Z_{n+2} - Z_{n+1}\|^2.$$

Whence,

$$(\xi_{n+2} - \xi_{n+1}, u^*)$$
$$= 2(Z_{n+1} - Z_{n+2}, \xi)(Z_{n+2} - Z_{n+1}, u^*)/\|Z_{n+2} - Z_{n+1}\|^2 \geqq 2\alpha\delta \geqq 2\alpha\rho\delta.$$

In either case, the assertion is true. Similarly, if $\xi_n = \xi \varepsilon V(\sigma)$ and $w_n = w \varepsilon X_0$, $[\langle Z_{n+2}, Z_{n+1} \rangle \varepsilon S^*(\xi, \alpha)] \subset [(\xi_{n+2} - \xi_{n+1}, u^*) > \gamma]$. From this and (3.1), $\Pr[(\xi_{n+2} - \xi_{n+1}, u^*) > \gamma \mid \xi_n = \xi, w_n = w] \geqq \mu > 0$ independent of n for all $\xi \varepsilon V(\sigma)$ and all w. By assumption, $0 < \epsilon \leqq \Pr[\xi_n \varepsilon V(\sigma)]$, so that

$$\Pr[(\xi_{n+2} - \xi_{n+1}, u^*) > \gamma \mid \xi_n \varepsilon V(\sigma)] \geqq \mu' > 0$$

independent of n. Hence, $\Pr[(\xi_{n+2} - \xi_{n+1}, u^*) > \gamma] \geqq \mu'\epsilon > 0$ for all n, so that

$$\limsup_{n\to\infty} \Pr[(\xi_{n+2} - \xi_{n+1}, u^*) > \gamma] > 0,$$

in contradiction to part (b) of the corollary to Theorem 2.

(b) Since (ξ_n, u) is a monotone sequence,

$$\sup_{u \varepsilon U} \lim_{n\to\infty} (\xi_n, u) = \sup_{u \varepsilon U} \sup_{n\to\infty} (\xi_n, u).$$

Since ξ_n is a bounded sequence, Lemma 3 guarantees that $\sup_{n\to\infty} (\xi_n, u)$ is a continuous function of u, and hence achieves its maximum over the compact set U, say at $\tilde{u}$. Thus, $\lim_{n\to\infty}(\xi_n, \tilde{u}) = 1$ so that

$$\lim_{n\to\infty} \|\xi_n - \tilde{u}\|^2 = \lim_{n\to\infty} \|\xi_n\|^2 + \|\tilde{u}\|^2 - 2(\tilde{u}, \xi_n) = 0.$$

This means that the event $[\sup_{u \varepsilon U} W(u) = 1]$ is contained in the event

$$\bigcup_{u \varepsilon U} [\xi_n \to u] \qquad \text{q.e.d.} \qquad \text{(L2, L3, L7, T2)}.$$

Theorem 3 guarantees that the unit normal, ξ_n, ends up pointing in the right direction. Now, if we can prove that the distance between $\pi(\xi_n, w_n)$ and some separating hyperplane, $\pi(u, w)$, tends to zero, it will readily follow that the probability of error converges to zero.

DEFINITION. (a) Let $Q_0(u) = \{z \varepsilon X_0 : \sup_{z' \varepsilon X_0} (u, z - z') \leqq 0\}$

$$Q_1(u) = \{z \varepsilon X_1 : \inf_{z' \varepsilon X_1} (u, z - z') \geqq 0\}$$

$$Q(u) = Q_0(u) \cup Q_1(u).$$

Since there can only be one plane of support to X_i with a given unit normal u passing through $Q_i(u)$, it follows that $\pi(u, q)$ and $\pi(u, q')$ are identical if $u \varepsilon U$ and q and q' are in the same $Q_i(u)$. Hereafter, we will denote by $\pi_i(u)$ the plane of support to X_i with normal u: $\pi_i(u) = \pi(u, q)$ where $q \varepsilon Q_i(u)$.

We will also have occasion to deal with the random variable

$$D_{ni}(u) = \sup_{z \varepsilon \pi(\xi_n, w_n) \cap X_i} \inf_{z' \varepsilon \pi_i(u)} \|z - z'\|, \qquad (i = 0, 1),$$

where we adopt the convention that $D_{ni}(u) = 0$ if $\pi(\xi_n, w_n) \cap X_i = \phi$. $D_{ni}(u)$ is an upper bound on the distance between $\pi_i(u)$ and points in X_i which lie on the "wrong side" of $\pi(\xi_n, w_n)$. In fact, an alternative representation for $D_{ni}(u)$ is:

$$D_{ni}(u) = \sup\{|(u, z - q)| : z \varepsilon X_i \& (\xi_n, z - w_n)(u, z - q) \leqq 0\}$$

where $q \varepsilon Q_i(u)$. Let us denote $\max\{D_{n0}(u), D_{n1}(u)\}$ by $D_n(u)$.

The next theorem shows that for some $u \varepsilon U$, $D_n(u) \to 0$ as $n \to \infty$.

THEOREM 4.

$$\Pr\{\bigcup_{u \varepsilon U} [D_n(u) \to 0]\} = 1.$$

PROOF. Let $d_{ni}(u) = 1 - (\xi_n, u) + |(w_n, \xi_n) - (u, q)|$, where $q \varepsilon Q_i(u)$, $(i = 0, 1)$. $d_{ni}(u)$ can be interpreted as a measure of the "distance" between $\pi(\xi_n, w_n)$ and $\pi_i(u)$. The first term measures the "angle" between ξ_n and u, while the second measures the difference between distances from the origin of the two planes. If we denote $\min\{d_{n0}(u), d_{n1}(u)\}$ by $d_n(u)$, it is easy to see that $D_n(u) \to 0$ if and only if $d_n(u) \to 0$, so we will show that

$$\Pr\{\bigcup_{u \varepsilon U} [d_n(u) \to 0]\} = 1.$$

The method of proof can be broken into two cases. In Case 1, $\xi_n = u$ for some n and some $u \varepsilon U$. In Case 2, $\xi_n \neq u$ for every $u \varepsilon U$ and every n, but $(\xi_n, u_0) \to 1$ for some $u_0 \varepsilon U$.

Case 1. Let $A = [\xi_n = u$ for some n and some $u \varepsilon U]$. Then

$$A = \bigcup_{n=1}^{\infty} (A_{n0} \cup A_{n1})$$

where

$$A_{ni} = \bigcup_{u \varepsilon U} \bigcup_{w \varepsilon X_i} A_n(u, w) \qquad (i = 0, 1),$$

and

$$A_n(u, w) = [\xi_{n-1} \not\in U, \xi_n = u, w_n = w].$$

Notice that (1) for each i, $A_{ni} \cap A_{mi} = \phi$

(2) for each n, $A_{n0} \cap A_{n1} = \phi$

and for each $u \varepsilon U$ and $w \varepsilon X_i$,

(3) $A_n(u, w) \cap A_m(u, w) = \phi$.

If $\xi_n = u \varepsilon U$ and $w_n = w \varepsilon X_0$, then under the estimation rule, $w_k \varepsilon X_0$ for every $k \geqq n$ and $d_{k0}(u) = |(w_k - q_0, u)|$ (where $q_0 \varepsilon Q_0(u)$) is non-increasing for $k \geqq n$.

Since the observations Z_k are statistically independent, the monotonicity of the sequence $d_{k0}(u)$ leads us at once to the conclusion that

$$\Pr[d_{k0}(u) \nrightarrow 0 \,|A_n(u, w)] \leqq \Pr\{\bigcup_{\epsilon>0} \bigcap_{k \geqq n+1} [Z_k \not\in F_\epsilon(u)\}$$

$$\leqq \sum_{j=1}^{\infty} \Pr\{\bigcap_{k>n} [Z_k \not\in F_{1/j}(u)]\},$$

where $F_\epsilon(u) = \{z \varepsilon X_0 : |(z - q_0, u)| < \epsilon\}$. By assumption B, $\Pr[F_\epsilon(u)] > 0$ for every $\epsilon > 0$, and since the Z_k's are independent and identically distributed, $\Pr\{\bigcap_{k>n} [Z_k \not\in F_{1/j}(u)]\} = 0$ for every j and every n. Thus,

$$\Pr\{\bigcup_{u \varepsilon U} [d_{k0}(u) \rightarrow 0] \mid A_n(u, w)\} = 1$$

for every $w \varepsilon X_0$ and every $u \varepsilon U$. Consequently, $\Pr\{\bigcup_{u\varepsilon U} [d_{k0}(u) \rightarrow 0] \mid A_{n,0}\} = 1$ for every n. In the same fashion, $\Pr\{\bigcup_{u\varepsilon U} d_{k,1}(u) \rightarrow 0] \mid A_{n,1}\} = 1$ for every n.

From these results it is a routine matter to verify that

$$\Pr\{\bigcup_{u \varepsilon U} [d_n(u) \rightarrow 0] \cap A\} = \Pr\{A\}.$$

We leave Case 1 for the present and now examine

Case 2. Let $B(u)$ be the event $[(\xi_n, u) \rightarrow 1 \;\&\; (\xi_n, u) < 1$ for every $n]$. If $B(u)$ occurs, then given ϵ, $(0 < \epsilon < 1)$, $1 - \epsilon < (\xi_n, u) < 1$ for some n. Furthermore, for some finite integer s, $(\xi_{n+s,u}) > (\xi_n, u)$. We will show that $d_k(u) \leqq (8r + 1)\epsilon^{\frac{1}{2}}$ for all $k > n + s$ (where $r = \sup_{x\varepsilon X_0 x \mathbf{U}_1} \|x\|$) and so, $B(u) \subset [d_n(u) \rightarrow 0]$.

Let m be the largest integer for which $(\xi_m, u) < (\xi_{n+s}, u)$. Since (ξ_n, u) is monotone, $n \leqq m < n + s$. Furthermore, if $Z_{n+s} \varepsilon X_1$, then $w_m \varepsilon X_0$ and if $Z_{n+s} \varepsilon X_0$, then $w_m \varepsilon X_1$. In either case, $(Z_{n+s} - w_m, \xi_m)\,(Z_{n+s} - w_m, u) \leqq 0$. Since

$$|(Z_{n+s} - w_m, u - \xi_m)| \leqq 2r\,\|u - \xi_m\| \leqq (8r^2\epsilon)^{\frac{1}{2}},$$

the relation

$$(Z_{n+s} - w_m, u) = (Z_{n+s} - w_m, \xi_m) + (Z_{n+s} - w_m, u - \xi_m),$$

implies that both $|(Z_{n+s} - w_m, \xi_m)|$ and $|(Z_{n+s} - w_m, u)|$ are no greater than $(8r^2\epsilon)^{\frac{1}{2}}$. In particular, if $Z_{n+s} \varepsilon X_0$,

$$0 < (q_1 - w_{n+s}, u) = (q_1 - Z_{n+s}, u) \leqq (w_m - Z_{n+s}, u) \leqq (8r^2\epsilon)^{\frac{1}{2}}$$

for every $q_1 \varepsilon Q_1(u)$; similarly, if $Z_{n+s} \varepsilon X_1$, $0 \geqq (q_0 - w_{n+s}, u) \geqq -(8r^2\epsilon)^{\frac{1}{2}}$ for every $q_0 \varepsilon Q_0(u)$. Since

$$\begin{aligned} d_{n+s,i}(u) &= 1 - (\xi_{n+s,u}) + |(\xi_{n+s}, w_{n+s}) - (u, q_i)| \\ &\leqq \epsilon + |(\xi_{n+s} - u, w_{n+s})| + |(u, w_{n+s} - q_i)| \\ &\leqq \epsilon + r\|\xi_{n+s} - u\| + (8r^2\epsilon)^{\frac{1}{2}} \\ &\leqq \epsilon + (2r^2\epsilon)^{\frac{1}{2}} + (8r^2\epsilon)^{\frac{1}{2}} \leqq (6r+1)\epsilon^{\frac{1}{2}} \end{aligned}$$

when $Z_{n+s} \varepsilon X_i$, we conclude that $d_{n+s}(u) \leqq (6r+1)\epsilon^{\frac{1}{2}}$. Let $N_1 < N_2 < \cdots$ be the times at which mistakes are made after the $(n+s)$th observation. Let $N_{k_1} < N_{k_2} < \cdots$ be the times at which the normal $\xi_{N_k} \neq \xi_{N_{k-1}}$. Then, if a mistake is made at time N_ν, $(N_{k_j} < N_\nu < N_{k_{j+1}})$, ξ_{N_ν} is left unchanged. In particular, $\xi_{N_\nu} = \xi_{N_\sigma}$ where $\sigma = k_j$ and $Z_{N_\nu} = w_{N_\nu}$, so that:

$$(w_{N_\nu} - w_{N_\sigma}, \xi_{N_\sigma}) \begin{cases} \geqq 0 & \text{if} \quad w_{N_\sigma} \varepsilon X_1 \\ \leqq 0 & \text{if} \quad w_{N_\sigma} \varepsilon X_0 . \end{cases}$$

(This is proved by an easy induction argument; w_{N_ν} always lies on the "wrong side" of $\pi(\xi_{N_{\nu-1}}, w_{N_{\nu-1}})$, and $w_{N_{\nu-1}}$ lies in the same set as w_{N_σ} for $\sigma < \nu < k_{j+1}$.)

It is easy to verify that $d_{N_\sigma}(u) \leqq (6r+1)\epsilon^{\frac{1}{2}}$ and that

$$0 \leqq (w_{N_\sigma} - q_0, u) \leqq (8r^2\epsilon)^{\frac{1}{2}} \quad \text{if} \quad Z_{N_\sigma} \varepsilon X_0,$$

$$0 \geqq (w_{N_\sigma} - q_1, u) \geqq -(8r^2\epsilon)^{\frac{1}{2}} \quad \text{if} \quad Z_{N_\sigma} \varepsilon X_1 \text{ (where } \sigma = k_j).$$

(Use the same argument that was used for $d_{n+s}(u)$.)

Now suppose $Z_{N_\sigma} \varepsilon X_0$: then, $0 \leqq (w_{N_\nu} - q_0, u) = (w_{N_\nu} - w_{N_\sigma}, u - \xi_{N_\sigma}) + (w_{N_\nu} - w_{N_\sigma}, \xi_{N_\sigma}) + (w_{N_\sigma} - q_0, u)$. The first term is no greater than $(8r^2\epsilon)^{\frac{1}{2}}$, the second is negative and the third is no greater than $(8r^2\epsilon)^{\frac{1}{2}}$. Thus

$$|(w_{N_\nu} - q_0, u)| \leqq (32r^2\epsilon)^{\frac{1}{2}} \text{ if } Z_{N_\sigma} \epsilon X_0 .$$

Similarly $|(w_{N_\nu} - q_1, u)| \leqq (32r^2\epsilon)^{\frac{1}{2}}$ if $Z_{N_\sigma} \varepsilon X_1$. Thus,

$$\begin{aligned} d_{N_\nu,i}(u) &\leqq \epsilon + |(\xi_{N_\nu} - u, w_{N_\nu})| + |(u, w_{N_\nu} - q_i)| \\ &\leqq \epsilon + (2^{\frac{1}{2}} + 32^{\frac{1}{2}})(r^2\epsilon)^{\frac{1}{2}} \leqq (8r+1)\epsilon^{\frac{1}{2}} \end{aligned}$$

if $Z_{N_\nu} \varepsilon X_i$. $(i = 0, 1)$. In any event, $d_{N_\nu}(u) \leqq (8r+1)\epsilon^{\frac{1}{2}}$. Thus, whenever a mistake is made, $d_{N_\nu}(u) \leqq (8r+1)\epsilon^{\frac{1}{2}}$. Since $d_m(u)$ does not change unless a mistake is made, $d_m(u) \leqq (8r+1)\epsilon^{\frac{1}{2}}$ for all $m \geqq n+s$. Thus

$$B(u) \subset [\lim_{n\to\infty} d_n(u) = 0],$$

so that

$$B = \bigcup_{u \varepsilon U} B(u) \subset \bigcup_{u \varepsilon U} [d_n(u) \to 0].$$

Since $\Pr[A \cup B] = \Pr\{\bigcup_{u\varepsilon U} [(\xi_n, u) \to 1]\} = 1$, (Theorem 3), it then follows that $\Pr[D] = \Pr[D \cap (A \cup B)] = \Pr[D \cap A] + \Pr[D \cap B]$, where

$$D = \bigcup_{u \varepsilon U} [d_n(u) \to 0].$$

But we have shown in Cases 1 and 2, that: $B \subset D$ and $\Pr[A \cap D] = \Pr[A]$, so that

$$\Pr[D] = \Pr[A \cup B] = 1. \qquad \text{q.e.d. (T3)}$$

Thus we see that the "empirical" hyperplane, $\pi(\xi_n, w_n)$, "converges," in some sense, to one that separates X_0 and X_1. What does this tell us about the asymptotic behavior of the error probability? Recall, that Z_{n+1} is misclassified if and only if the event

$$E_{n+1} = [Z_{n+1} \varepsilon X_0 \text{ and } (\xi_n, Z_{n+1} - w_n) < 0] \cup [Z_{n+1} \varepsilon X_1 \text{ and } (\xi_n, Z_{n+1} - w_n) \geqq 0]$$

occurs. We state as our main result

THEOREM 5.

$$\lim_{n\to\infty} \Pr[E_n] = 0.$$

PROOF. If E_{n+1} occurs, then $(\xi_n, Z_{n+1} - w_n)\ (u, Z_{n+1} - q) \leqq 0$ for every $q \varepsilon Q(u)$ and every $u \varepsilon U$. Hence for every $u \varepsilon U$, $|(u, Z_{n+1} - q_i)| \leqq D_n(u)$ if $Z_{n+1} \varepsilon X_i$ and $q_i \varepsilon Q_i(u)$, $(i = 0, 1)$. Thus $E_{n+1} \subset \bigcap_{u\varepsilon U} (F_{n0}(u) \cup F_{n1}(u))$ where $F_{ni}(u) = [Z_{n+1} \varepsilon X_i] \cap [|(u, Z_{n+1} - q_i)| < D_n(u)]$, $(i = 0, 1)$.

Since the Z_{n+1} are identically distributed and Z_{n+1} is independent of $D_n(u)$,

$$\Pr[E_{n+1}] \leqq \Pr\{\bigcap_{u \varepsilon U} (F^*_{n0}(u) \cup F^*_{n1}(u))\}$$

where $F^*_{ni}(u) = [Z \varepsilon X_i] \cap [|(u, Z - q_i)| < D_n(u)]$ and Z is a r.v. with the same distribution as Z_{n+1}, and is independent of $D_n(u)$. Let

$$D = \bigcup_{u \varepsilon U} [D_n(u) \to 0].$$

$\Pr[D] = 1$ by Theorem 4, so

$$\Pr[E_{n+1}] \leqq \Pr\{\bigcap_{u \varepsilon U} F^*_{n0}(u) \cap D\} + \Pr\{\bigcap_{u \varepsilon U} F^*_{n1}(u) \cap D\}.$$

Hence

$$\limsup_{n\to\infty} \Pr[E_{n+1}] \leqq \limsup_{n\to\infty} \Pr\{\bigcap_{u \varepsilon U} F^*_{n0}(u) \cap D\} + \limsup_{n\to\infty} \Pr\{\bigcap_{u \varepsilon U} F^*_{n1}(u) \cap D\}$$

$$\leqq \Pr\{\limsup_{n} \bigcap_{u \varepsilon U} F^*_{n0}(u) \cap D\} + \Pr\{\limsup_{n} \bigcap_{u \varepsilon U} F^*_{n1}(u) \cap D\}$$

where

$$\limsup_n A_n = \bigcap_n \bigcup_{m \geqq n} A_m$$

is the event: "A_n occurs for infinitely many n". Suppose the event

$$\limsup_n \bigcap_{u \varepsilon U} F^*_{ni}(u) \cap D$$

occurs. Then for some u_0, the event $[D_m(u_0) \to 0]$ and $\limsup_n F^*_{ni}(u_0)$ occurs. But this implies that the event: "$D_m(u_0) \to 0$ and $|(u_0, Z - q_i)| < D_n(u_0)$ for infinitely many n" occurs, which implies the event $[Z \varepsilon Q_i(u_0)]$. $(i = 0, 1)$. Thus,

$$\limsup_n \bigcap_{u \varepsilon U} F^*_{ni}(u) \cap D \subset [Z \varepsilon \bigcup_{u \varepsilon U} Q_i(u)],$$

so that

$$\limsup_{n \to \infty} \Pr[E_n] \leqq \Pr[Z \varepsilon \bigcup_{u \varepsilon U} Q(u)] = 0 \qquad \text{q.e.d. (T4)}$$

We wish to point out that the restriction to Euclidean spaces can be weakened slightly. The results hold as long as $X_0 \cup X_1$ is a subset of a finite dimensional Hilbert space.

The extension of these results to infinite dimensional Hilbert spaces is unquestionably a worthy venture. Since U and V (σ) are no longer compact in any useful topology, (specifically the weak topology, the norm topology or any topology between the two), and since their compactness appears to be crucial to the present method of proof, a different method of proof will have to be found.

Finally, a word about Assumption C_1: It seems a pity that the results of Theorem 5 rely upon the fact that $\Pr\{\bigcup_{u \varepsilon U} Q(u)\} = 0$, but this is indeed an essential assumption if the simple geometric procedure proposed here is to work. To see that this is so, consider the following example where points are chosen independently and at random from certain subsets of the plane: let $x_1 = \langle 0, 0 \rangle$, $x_2 = \langle 0, 1 \rangle$, $x_3 = \langle 0, 2 \rangle$ and $x_4 = \langle 1, 1 \rangle$. Let $X_0 = \{x_1, x_2\}$ and $X_1 = \{x_3, x_4\}$, and let $\Pr\{x_j\} = \frac{1}{4}$ for $j = 1, 2, 3, 4$.

It is easy to verify that all assumptions except C_1 are satisfied and that the sequence of observations $Z_0 = x_3$, $Z_1 = x_1$, $Z_2 = x_2$ will occur with positive probability. Thereafter, an error will occur whenever x_4 occurs. Apparently, no simple modification of either the guessing rule or the estimation rule will permit us to deal with situations of this type unless more detailed information about the structure of X_0 and X_1 are known *a-priori*.

6. Acknowledgments. I would like to thank Prof. Herman Chernoff for reading and commenting upon an early version of this manuscript. His suggestions lead to a substantially more direct proof of Theorem 4 than was originally comtemplated.

APPENDIX—Lemmas

This section is devoted to the proof of lemmas required by the theorems in the text. It should be pointed out that the lemmas are purely geometric or probabalistic in nature and do not depend upon the procedure set forth in Section 4.

DEFINITION.

(a) For any subset, A, of a real, finite dimensional Euclidean space H, let $H(A)$ be the smallest subspace of H which contains A.

(b) Let $K(A)$ be the convex hull of A. ($K(A)$ is the smallest convex set containing A.)

LEMMA 1. *$a \varepsilon K(A)$ if and only if there are non-negative constants $\lambda_1, \lambda_2, \cdots \lambda_k$ such that $\sum_{i=1}^{k} \lambda_i = 1$, and points $a_1, a_2, \cdots a_k$ in A such that $a = \sum_{i=1}^{k} \lambda_i a_i$.*

PROOF. See page 31 of [1].

DEFINITION. Let X and Y be subsets of H. Define

$$Y - X = \{z: z = y - x, y \varepsilon Y, x \varepsilon X\}.$$

LEMMA 2. $\inf_{z \varepsilon Y-X} (u, z) > 0$ *for some u in the closure of $K(Y - X)$ if and only if*

$$\inf_{z \varepsilon K(Y)-K(X)} \|z\| > 0.$$

PROOF. Necessity: Let $c_2 = \inf_{y \varepsilon Y} (u, y)$ and $c_1 = \sup_{x \varepsilon X} (u, x)$. By hypothesis, $c_2 > c_1$. From Lemma 1, $\inf_{y \varepsilon K(Y)} (u, y) = c_2$ and $\sup_{x \varepsilon K(X)} (u, x) = c_1$. Thus, $\inf_{z \varepsilon K(Y)-K(X)} (u, z) > 0$. The result follows from the Schwartz inequality.

Sufficiency: Notice that $K(Y - X) \subset K(Y) - K(X)$, so that

$$\inf_{z \varepsilon \mathbf{K}} \|z\| > 0$$

where $\mathbf{K}$ is the closure of $K(Y - X)$ in the norm topology. Let $\delta = \inf_{z \varepsilon \mathbf{K}} \|z\|$, and let z_n be a bounded sequence in $\mathbf{K}$ for which $\lim_{n \to \infty} \|z_n\| = \delta$.

We can pick a point $u \varepsilon \mathbf{K}$ and a convergent subsequence of $\{z_n\}$ so that the subsequence converges to u; whence $\|u\| = \delta$.

Let s be any point with $(u, s) < \delta^2$, and let $f(\lambda) = \|\lambda s + (1 - \lambda)u\|^2$. A routine computation shows that

$$f'(0) = 2(u, s - u) < 0.$$

Since $f(0) = \delta^2$, there must be a value of μ, $(0 < \mu < 1)$, such that $\|s_0\| = \|\mu s + (1 - \mu)u\| < \delta$. Thus, s_0 is not a member of $\mathbf{K}$. Since that set is convex and contains u, it cannot contain s either. Hence, $(u, z) \geqq \delta^2$ for every $z \varepsilon \mathbf{K} \supset Y - X$. (L1)

LEMMA 3. *Let A be any bounded subset of H. The functional $g(\xi) = \sup_{a \varepsilon A} (a, \xi)$ is continuous in the norm topology.*

PROOF.

$$g(\xi_n) - g(\xi) = \sup_{a \varepsilon A} (\xi_n - \xi + \xi, a) - g(\xi) \leqq \sup_{a \varepsilon A} (\xi_n - \xi, a)$$
$$\leqq \|\xi_n - \xi\| \alpha, \text{ where } \alpha = \sup_{a \varepsilon A} \|a\|. \text{ Similarly, } g(\xi) - g(\xi_n) \leqq \|\xi_n - \xi\| \alpha.$$

Hence,

$$|g(\xi) - g(\xi_n)| \to 0$$

when $\|\xi_n - \xi\| \to 0$.

LEMMA 4. *Let X and Y be subsets of H and let B_1 and B_2 be open subsets of $H(X \cup Y)$ such that $X \subset B_1$ and $Y \subset B_2$. If B is an open*[3] *subset of $B_1 \otimes B_2$ and $B \cap (X \otimes Y)$ is non-empty, then there are open subsets B_X and B_Y of $H(X \cup Y)$ such that*

$$B_X \otimes B_Y \subset B,\ B_X \cap X \neq \phi \text{ and } B_Y \cap Y \neq \phi.$$

PROOF. Pick $\langle x_0, y_0\rangle \in B \cap (X \otimes Y)$ and construct an open sphere about $\langle x_0, y_0\rangle$ so that it is a subset of B. Let δ be the radius of that sphere and let

$$B_X = \{x \in B_1 : \|x - x_0\| < \delta/2^{\frac{1}{2}}\}$$

$$B_Y = \{y \in B_2 : \|y - y_0\| < \delta/2^{\frac{1}{2}}\}.$$

q.e.d.

DEFINITION.

(a) Let $U = \{u \in H(X_0 \cup X_1) : \|u\| = 1 \text{ and } \sup_{z \in X_1 - X_0} (u, z) \leqq 0\}$.

(b) For every $\sigma > 0$, let

$$V(\sigma) = \{\xi \in H(X_0 \cup X_1) : \|\xi\| = 1 \text{ and } \sup_{u \in U} (\xi, u) \leqq 1 - \sigma\}.$$

(c) $S(\xi, \alpha) = \{\langle x, y\rangle \in H(X_1 \cup X_0) \otimes H(X_1 \cup X_0) : (x - y, \xi)/\|x - y\|^2 > \alpha\}$.

(d) $S^*(\xi, \alpha) = (X_1 \otimes X_0) \cap S(\xi, \alpha)$.

LEMMA 5.

(a) *U and $V(\sigma)$ are compact.*

(b) *If $\sigma > 0$ is given, there is a positive value of α for which $S^*(\xi, \alpha)$ is non-empty for every $\xi \in V(\sigma)$.*

PROOF.

(a) By Lemma 3, U and $V(\sigma)$ are always closed subsets of the unit sphere (L3-b), and so are compact.

(b) If not, then for some $\sigma > 0$, every sequence of positive constants α_n which converges to zero admits the choice of a corresponding sequence $\xi_n \in V(\sigma)$ for which $\bigcup_n S^*(\xi_n, \alpha_n) = \phi$. It is clear that

$$h(\xi) = \sup_{z \in X_1 - X_0} (z, \xi)/\|z\|^2$$

is continuous and hence achieves its minimum (at ξ_0, say) in the compact set $V(\sigma)$. If $h(\xi_0) \leqq 0$, then we have $\xi_0 \in U$ which is not possible since $V(\sigma) \cap U = \phi$. Thus $h(\xi_0) > 0$.

However, if $\bigcup_n S^*(\xi_n, \alpha_n) = \phi$, this implies that

$$\lim_{n\to\infty} \inf h(\xi_n) \leqq \lim_{n\to\infty} \alpha_n = 0.$$

[3] Open in the product norm topology: The open sphere of radius δ $\langle x_0, y_0\rangle$ is given by $\{\langle x, y\rangle : \|x + x_0\|^2 + \|y - y_0\|^2 < \delta^2\}$.

Since ξ_0 is minimal, $h(\xi_0) \leqq 0$ which is a contradiction.

(L3).

Lemma 2 and Assumption A_2 insure the existence of a $u^* \varepsilon U$ and constants $c_2 > c_1$ such that

$$\inf_{y \varepsilon X_0} (u^*, y) > c_2 > c_1 > \sup_{x \varepsilon X_1} (u^*, x).$$

Let

$$B_1 = \{z \varepsilon H(X_0 \cup X_1): (u^*, z) < c_1\}$$

$$B_0 = \{z \varepsilon H(X_0 \cup X_1): (u^*, z) > c_2\}.$$

Clearly, $X_1 \subset B_1$, $X_0 \subset B_0$ and both B_1 and B_0 are open subsets of $H(X_0 \cup X_1)$. Thus, we obtain:

Lemma 6. *For every $\xi \varepsilon H(X_0 \cup X_1)$ and every positive α, $S(\xi, \alpha) \cap (B_1 \otimes B_0)$ is an open subset of $B_1 \otimes B_0$ in the product norm topology.*

Proof. Since

$$\inf_{\langle x, y\rangle \varepsilon B_1 \otimes B_0} \|x - y\| > 0,$$

the function $h_\xi(x, y) = (x - y, \xi)/\|x - y\|^2$ is continuous over $B_1 \otimes B_0$ for each ξ, so that $S(\xi, \alpha) \cap (B_1 \otimes B_0)$ is open in the product norm topology.

Lemma 7. *If $\sigma > 0$ is given and two random variables Z and Z' are chosen independently from $X_0 \cup X_1$, then*

$$\sup_{\alpha>0} \inf_{\xi \varepsilon V(\sigma)} \Pr\{\langle Z, Z'\rangle \varepsilon S^*(\xi, \alpha)\} > 0.$$

Proof. Let $G(\xi, \alpha) = \Pr\{\langle Z, Z'\rangle \varepsilon S^*(\xi, \alpha)\}$. We will show that for each $\alpha > 0$, $G(\xi, \alpha)$ is lower semicontinuous and that for some $\alpha > 0$, $G(\xi, \alpha)$ is positive for every ξ in $V(\sigma)$. Since $V(\sigma)$ is compact, the two together furnish the desired result:

Suppose ξ_n converges to ξ. Then $(x, \xi_n) \to (x, \xi)$ for every x. In particular, if D is any subset of $H(X_0 \cup X_1)$ whose elements are bounded away from zero in norm, we have $(x, \xi_n - \xi)/\|x\|^2 \to 0$ for all $x \varepsilon D$. Now, if $\langle x_1, x_2\rangle \varepsilon S^*(\xi, \alpha)$, then

$$(x_1 - x_2, \xi)/\|x_1 - x_2\|^2 > \alpha,$$

and for some n,

$$(x_1 - x_2, \xi_m)/\|x_1 - x_2\|^2 > \alpha$$

for all $m \geqq n$. Thus,

$$S^*(\xi, \alpha) \subset \bigcup_{n \geqq 1} \bigcap_{m \geqq n} S^*(\xi_m, \alpha),$$

so that

$$G(\xi, \alpha) = \Pr\{\langle Z, Z'\rangle \varepsilon S^*(\xi, \alpha)\} \leqq \Pr\{\langle Z, Z'\rangle \in \bigcup_{n \geqq 1} \bigcap_{m \geqq n} S^*(\xi_m, \alpha)\}$$

$$\leqq \liminf_{n \to \infty} \Pr\{\langle Z, Z'\rangle \in S^*(\xi_n, \alpha)\} = \liminf_{n \to \infty} G(\xi_n, \alpha). \qquad ([3], \text{p } 150.)$$

Thus, $G(\xi, \alpha)$ is, for each $\alpha > 0$, lower-semi continuous.

Now Lemma 5 guarantees that given $\sigma > 0$, $(B_1 \otimes B_0) \cap S^*(\xi, \alpha_\sigma)$ is non-empty for some $\alpha_\sigma > 0$, whenever $\xi \varepsilon V(\sigma)$. Since $(B_1 \otimes B_0) \cap S(\xi, \alpha_\sigma)$ is open, we apply Lemma 4 to assert the existence of open (in the norm topology) sets O_1 and O_0 in $H(X_0 \cup X_1)$ for which

$$O_1 \otimes O_0 \subset B_1 \otimes B_0 \cap S(\xi, \alpha_\sigma), \qquad O_0 \cap X_0 \neq \phi \text{ and } O_1 \cap X_1 \neq \phi.$$

From Assumption B, $\Pr[O_0] > 0$ and $\Pr[O_1] > 0$. Since

$$B_0 \cap X_1 = B_1 \cap X_0 = \phi, \Pr\{\langle Z, Z' \rangle \varepsilon S(\xi, \alpha) \cap B_1 \otimes B_0\} = \Pr\{\langle Z, Z' \rangle \varepsilon S^*(\xi, \alpha)\},$$

so that $G(\xi, \alpha_\sigma) \geqq \Pr\{\langle Z, Z' \rangle \varepsilon O_1 \otimes O_0\} = \Pr[O_0]\Pr[O_1] > 0.$ (L4, L5).

REFERENCES

[1] BLACKWELL, DAVID and GIRSHICK, M. A. (1954). *Theory of Games and Statistical Decisions.* Wiley, New York.

[2] LOÈVE, MICHEL (1955). *Probability Theory.* Van Nostrand, New York.

[3] ROSENBLATT, FRANK (1962). *Principles of Neurodynamics.* Spartan Books, Washington, D. C.