[6] K. A. Brakke, J. M. Mantock, and K. Fukunaga, "Systematic feature extraction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-4, pp. 291-297, May 1982.

[7] E. M. Rounds, "A combined nonparametric approach to feature selection and binary decision tree design," *Pattern Recognition*, vol. 12, pp. 312-317, 1980.

[8] Y. X. Gu, Q. R. Wang, and C. Y. Suen, "Application of a multilayer decision tree in computer recognition of Chinese characters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-5, pp. 83-89, Jan. 1983.

[9] M. T. Y. Lai and C. Y. Suen, "Automatic recognition of characters by Fourier descriptors and boundary line encodings," *Pattern Recognition*, vol. 14, pp. 383-393, 1981.

[10] J. F. Fleming and R. F. Hemmings, "A method of recognition for handwritten block capitals," *Pattern Recognition Lett.*, vol. 1, pp. 457-464, 1983.

[11] M. Nadler, "Structural codes for omni-font and handprinted characters: II," *Proc. 5th Int. Joint Conf. Pattern Recognition*, Dec. 1980, pp. 848-852.

[12] B. Duerr, W. Haettich, H. Tropf, and G. Winkler, "A combination of statistical and syntactical pattern recognition applied to classification of unconstrained handwritten numerals," *Pattern Recognition*, vol. 12, pp. 189-199, 1980.

[13] K. Yamamoto and S. Mori, "Recognition of handprinted characters by an outermost point method," *Pattern Recognition*, vol. 12, pp. 229-236, 1980.

[14] C. Y. Suen, M. Berthod, and S. Mori, "Automatic recognition of handprinted characters—The state of the art," *Proc. IEEE*, vol. 68, pp. 469-487, Apr. 1980.

[15] C. Y. Suen, "Distinctive features in automatic recognition of handprinted characters," *Signal Processing*, vol. 4, pp. 193-207, Apr. 1982.

[16] S. Mori, K. Yamamoto, and M. Yasuda, "Research on machine recognition of handprinted characters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, pp. 386-405, July 1984.

[17] B. Chandrasekaran, "Independence of measurements and the mean recognition accuracy," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 452-456, July 1971.

[18] H. Van Steenis, "The IBM 1275 recognition system and its development," presented at 4th Cong. Deutsche Gesellschaft für Kybernetik, Berlin Tech. Univ., Berlin, West Germany, Apr. 6-9, 1970

[19] L. A. Kamentsky and C. N. Liu, "Computer-automated design of multifont print recognition logic," *IBM J. Res. Devel.*, vol. 7, pp. 2-13, Jan. 1963.

[20] —, "A theoretical and experimental study of a model for pattern recognition," in *Computer and Information Sciences*. Washington, DC: Spartan, 1964, pp. 194-218.

[21] C. N. Liu, "A programmed algorithm for designing multifont character recognition logics," *IEEE Trans. Comput.*, vol. C-13, pp. 586-593, Oct. 1964.

[22] R. Bakis, N. M. Herbst, and G. Nagy, "An experimental study of machine recognition of hand-printed numerals," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-4, pp. 119-132, July 1968.

[23] W. W. Bledsoe and I. Browning, "Pattern recognition and reading by machine," in *Proc. Eastern Joint Comput. Conf.*, pp. 225-232, 1969.

[24] L. Uhr and V. Vossler, "A pattern recognition program that generates, evaluates and adjusts its own operators," *Proc. WJCC*, pp. 555-561, 1961.

[25] J. Ullman, "Experiments with the *n*-tuple method of pattern recognition," *IEEE Trans. Comput.*, vol. C-18, pp, 1135-1137, Dec. 1969.

[26] I. Flores and L. Grey, "Optimization of reference signals for character recognition systems," *IRE Trans. Comput.*, vol. EC-9, pp. 54-61, Mar. 1960.

[27] G. Nagy, "State of the art in pattern recognition," *Proc. IEEE*, vol. 56, pp. 836-848, May 1968.

[28] "Character recognition 1971," Brit. Comput. Soc., London, England, 1971.

[29] W. G. L. Adaway, "The machine reading of travel documents," *Int. Conf. Electron. Image Processing, IEE Conf. Pub.*, no. 214, pp. 159-163, July 1982.

[30] "TO-5000 multi-font OCR," Totec Co. Ltd., Tokyo, Japan.

[31] "Workless station," Dest Corp., San Jose, CA.

Monotonicity of Linear Separability Under Translation

ALFRED M. BRUCKSTEIN AND THOMAS M. COVER

***Abstract*—A set of n pattern vectors are given in d-space and classified arbitrarily into two sets. The sets of patterns are said to be linearly separable if there exists a hyperplane that separates them. We ask whether translation of one of these sets in an arbitrary direction helps separability. Sometimes yes and sometimes no, but yes on the average. The average is taken over all classifications of the patterns into two sets. In fact, we prove that the probability of separability increases as the translation increases. Thus, we conclude that if points are drawn equiprobably from densities $f_0(x)$ and $f_1(x) = f_0(x + tw)$ then the probability of linear separability is minimum at $t = 0$ and increases with t for $t > 0$.**

***Index Terms*—Convex sets, linear separability, monotonicity, pattern classification.**

I. Introduction

Consider the standard statistical pattern classification problem in which the classifications $\theta_1, \theta_2, \cdots, \theta_n$ are independent identically distributed random variables with $P\{\theta_i = 0\} = P\{\theta_i = 1\} = 1/2$, and the corresponding vector-valued observations $x_1, x_2, \cdots, x_n \in R^d$ are conditionally independently drawn according to $f_{\theta i}(x)$, where $f_0(x)$ and $f_1(x)$ are known probability density functions. Thus, the probability density of the classified set $\{X_i, \theta_i\}_{i=1}^n$ is $2^{-n}\Pi_{i=1}^n f_{\theta i}(x_i)$. The realization $\{(x_i, \theta_i)\}_{i=1}^n$ is called linearly separable if there exists a vector v and a constant T such that

$$\begin{aligned} v^t x_i &> T \quad \text{for} \quad \theta_i = 1 \\ v^t x_i &< T \quad \text{for} \quad \theta_i = 0. \end{aligned} \tag{1}$$

The following result is well known (see, e.g., [1], [2]).

Theorem 1: If $f_0(x) = f_1(x)$ then

$$\Pr\{\{(X_i, \theta_i)\}_{i=1}^n \text{ is linearly separable}\} = \frac{1}{2^{n-1}} \sum_{i=0}^{d} \binom{n-1}{i}. \tag{2}$$

Note that the probability does not depend on the underlying density. The proof of this theorem is based on a purely geometric argument which provides the number of dichotomies that can be induced by hyperplanes on a set of points in R^d in general position. This also makes it clear why the result is distribution free.

We now consider densities that differ by a translation. Given w, an arbitrary unit vector in R^d, we prove the following.

Theorem 2: If $f_1(x) = f_0(x + tw)$ then

$$\Pr\{\{(X_i, \theta_i)\}_{i=1}^n \text{ is linearly separable}\} \tag{3}$$

Manuscript received April 9, 1984; revised January 15, 1985. Recommended for acceptance by Josef V. Kittler. This work was supported in part by the National Science Foundation under Grants ECS78-10003 and NSF-82-11568, and in part by the Joint Services Electronics Program under Contract DAAG29-81-K-0057.

A. M. Bruckstein is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305.

T. M. Cover is with the Department of Electrical Engineering and Statistics, Stanford University, Stanford, CA 94305.

0162-8828/85/0500-0355$01.00 © 1985 IEEE

is a monotonic nondecreasing function of $t > 0$.

II. Proof of Theorem 2

We shall first take the probability out of the problem by a simple observation and prove a counterpart of Theorem 2 for a strictly geometrical setup. The probabilistic result will easily follow.

The process of generating the classified samples for translated underlying distributions is equivalent to choosing all the points according to a density $f_0(x)$, classifying them at random (i.e., by flipping a fair coin) and then translating the points of class 1 by the translation vector tw. Consider thus a set of n points in R^d, in general position, and all its 2^n subsets, $S_1, S_2, \cdots, S_{2^n} \subseteq \{x_1, x_2, \cdots, x_n\}$. Each subset defines a classification or dichotomy $\{S_k, S_k^c\}$ of the original set of points, where S^c denotes the complement of S with respect to $\{x_1, x_2, \cdots, x_n\}$. The set pair $\{S_k, S_k^c\}$ is said to be linearly separable if there exists a separating hyperplane, i.e., if there exists a vector v and a constant T such that

$$v^t x \geqslant T \quad \text{for} \quad x \in S$$
$$v^t x < T \quad \text{for} \quad x \in S^c. \tag{4}$$

We now evaluate the number of linearly separable dichotomies of points which result when translating the subsets S_k by tw to form the set $S_k + tw = \{x_i + tw \mid x_i \in S_k\}$. Associate to each S_k a translation separability indicator function, as follows

$$I_k(t) = \begin{cases} 1, & \text{if } \{S_k + tw, S_k^c\} \text{ is linearly separable} \\ 0, & \text{if } \{S_k + tw, S_k^c\} \text{ is not linearly separable.} \end{cases} \tag{5}$$

Definition: Let $\Pi(t)$ be the the number of linearly separable sets of points among $\{S_k + tw, S_k^c\}$, for $k = 1, 2, \cdots, 2^n$.

It follows from (5) that

$$(t) = \sum_{k=0}^{2^n} I_k(t). \tag{6}$$

Suppose that $\{S_k, S_k^c\}$ is not linearly separable. Then translation of the points of S_k can only make the dichotomy separable. If, however, we start with a separable dichotomy of points, translation may at some "critical" distance t_c produce a nonseparable dichotomy. This happens when the convex hull of the points in $S_k + tw$ intersects the convex hull of the "static" points S_k^c. As $t \to \infty$ the dichotomy will again become separable (see Fig. 1). In Fig. 1(a), we see that $S_k = \{x_1, x_4, x_5, x_6, x_9\}$ becomes separable from S_k^c at translation $t = t_1$, and remains separable thereafter. Fig. 1(b) shows $S_l = \{x_2, x_4, x_5, x_7\}$ becoming nonseparable at $t = t_c$ and regaining separability for $t > t_2$. Thus it is clear that the count function $\Pi(t)$ is piecewise constant. Also, if $\{x_1, x_2, \cdots, x_n\}$ is in general position, then,

$$\Pi(0) = 2 \sum_{i=0}^{d} \binom{n-1}{i} \quad \text{and} \quad \Pi(\infty) = 2^n. \tag{7}$$

We shall argue that the following result holds.

Geometric Theorem: If $\{x_1, x_2, \cdots, x_n\}$ is in general position, the right-continuous version of $\Pi(t)$ is a piecewise constant *nondecreasing* function of t, for $t > 0$.

Proof: It is enough to consider the behavior of $\Pi(t)$ for $n > d + 1$ because $\Pi(t) = \Pi(0) = 2^n$ for all t when $n \leqslant d + 1$. Showing that $\Pi(t)$ is nondecreasing is equivalent to proving that no "down-jumps" will occur. Since $\Pi(t) = \Sigma I_k(t)$ and the $I_k(t)$ are not necessarily monotonic, we wish to find for every $I_p(t)$ having a down-jump at t_c, another $I_q(t)$ having a cancelling up-jump at t_c. Suppose that $\{S_p + tw, S_q^c\}$ becomes nonseparable at $t = t_c$. This happens because (at least) one of the points of one subset crosses a face of the convex hull of the other [Fig. 2(a)]. Thus at the critical translation we will have a "collision hyperplane" on which there are points of both $S_p + t_c w$ and S_p^c. Now let S_q be the uniquely defined subset comprising points which either belonged to S_p and when translated by $t_c w$ moved to the collision plane or belonged to S_p^c and were not on the collision plane. It is easy to see that at the critical translation t_c, $\{S_q + tw, S_q^c\}$ will change from nonseparable to separable and thus $I_p(t) + I_q(t)$ will not have a down-jump at t_c [Fig. 2(b)]. Thus we have identified a set S_q such that when $I_p(t)$ has a down-jump, $I_q(t)$ has a counteracting up-jump. This matching is always possible by construction. This proves the theorem.

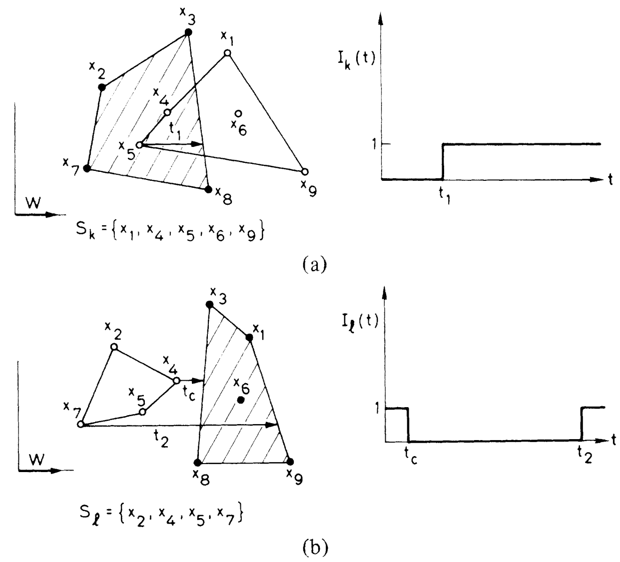

Fig. 1. Separability indicator functions.

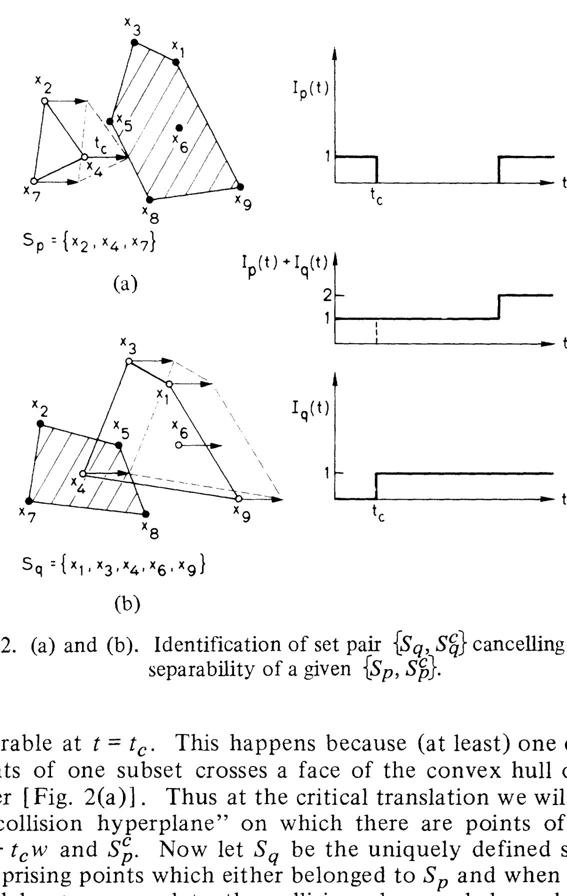

Fig. 2. (a) and (b). Identification of set pair $\{S_q, S_q^c\}$ cancelling loss of separability of a given $\{S_p, S_p^c\}$.

A typical sample function $\Pi(t)$ is given in Fig. 3. Note the finite set of departures from monotonicity occurring at critical values t_c, mentioned in the proof. These points are those at which general position is temporarily lost. If $X_1, X_2, \cdots, X_n$ are independently drawn according to some probability density,

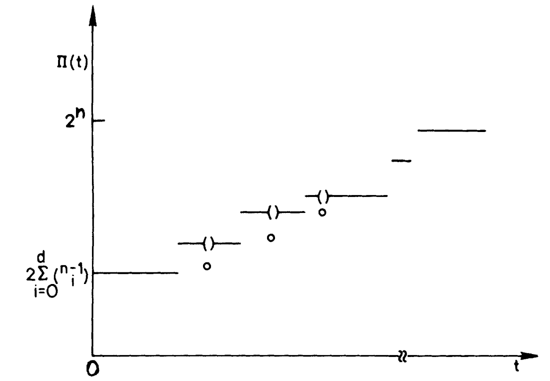

Fig. 3. A typical separability count function. The right continuous version is monotonic nondecreasing.

then, with probability one, any given t will not be such an exceptional point. This will be used in the proof of Theorem 2.

The geometric result immediately implies that the probability of linear separability when the underlying distributions are shifted versions of one another is a nondecreasing function of the shift parameter. This probability is given by

$$P_t(n, d) = \int \frac{\Pi(t; x_1, x_2, \cdots, x_n)}{2^n} \prod_{i=1}^{n} f(x_i) dx_i \tag{8}$$

where $\Pi(t; x_1, x_2, \cdots, x_n)$ is the count function $\Pi(t)$ corresponding to $\{x_1, x_2, \cdots, x_n\}$ in the geometric theorem. Monotonicity can be shown by examining

$$P_t(n, d) - P_{t-\delta}(n, d) = \int \frac{\Pi(t; x_1, x_2, \cdots, x_n) - \Pi(t-\delta; x_1, x_2, \cdots, x_n)}{2^n} \prod_{i=1}^{n} f(x_i) dx_i \tag{9}$$

The integrand $\Pi(t; x_1, x_2, \cdots, x_n) - \Pi(t-\delta; x_1, x_2, \cdots, x_n)$ is nonnegative almost everywhere with respect to the measure $\Pi_{i=1}^{n} f(x_i) dx_i$ as argued above. Thus $P_t(n, d) - P_{t-\delta}(n, d) \geq 0$. This completes the proof of Theorem 2.

A related question we can resolve in a similar fashion is the following. If separability is defined as the probability that there exists a hyperplane passing through a fixed point **O** in R^d, how does shifting the underlying distribution of one of the classes influence it? The answer is the same as in the previous case: separability increases monotonically with the shift parameter. The proof of this result proceeds as follows: separability of an initially separable sample is lost if either the point **O** penetrates the convex hull of the moving class points or the convex hull of the moving points intersects the polyhedral convex cone generated by the stationary points (see Fig. 4). But, in both these cases we can find corresponding, uniquely defined dichotomies which become separable at exactly the critical translations.

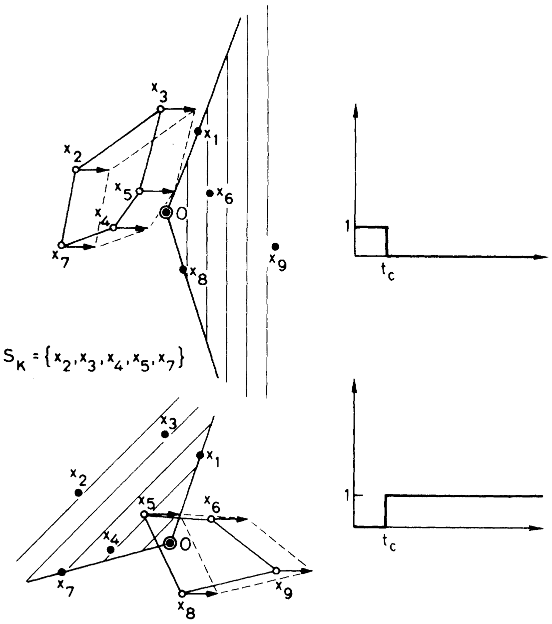

Fig. 4. Cancelling loss of separability with hyperplanes passing through a fixed point **O**.

III. An Open Question on Separability

We conjecture that the probability that there exists a separating hyperplane is always higher than the probability $P_0(n, d)$ of separating n points in d-space drawn from identical densities if the underlying distributions are simply different, rather than shifted. For $d = 1$ we can prove this assertion since one can obtain an explicit expression for the probability of separability in terms of the underlying distribution densities $f_0(x)$ and $f_1(x)$. The result is

$$P(n, 1) = \frac{n}{2^n} \int_{-\infty}^{+\infty} \{[1 + \psi(s)]^{n-1} + [1 - \psi(s)]^{n-1}\} f_i(s) ds \tag{10}$$

for $i = 0$ or 1, where

$$\psi(s) = \int_{-\infty}^{s} [f_1(\xi) - f_0(\xi)] d\xi. \tag{11}$$

The proof of this fact is as follows.

The probability that n points drawn at random from two sets having distribution densities $f_0(x)$ and $f_1(x)$ are separable about a *fixed* threshold s, on the line, is easily obtained as

$$P(n, 1; s) = \frac{1}{2^n} \{[1 + \psi(s)]^n + [1 - \psi(s)]^n\}$$
$$= \sum_k \frac{\binom{n}{k}}{2^n} \left\{ \left(1 - \int_0^s f_1(\xi) d\xi\right)^k \left(\int_0^s f_2(\xi) d\xi\right)^{n-k} + \left(\int_0^s f_1(\xi) d\xi\right)^k \left(1 - \int_0^s f_2(\xi) d\xi\right)^{n-k} \right\}. \tag{12}$$

Now, note that we have a separable realization if and only if one of the points, say x_i, belongs to either class $\theta_i = 1$ or $\theta_i = -1$ and the others form a separable realization with x_i as threshold. This proves (10), since separability is implied by the occurrence of one of n *disjoint* events with probabilities given by either p_+ or p_-, where

$$p_{\pm} = \Pr[\theta_i = \pm 1] \int_{-\infty}^{+\infty} P(n-1, 1; s) \Pr[X_i \in (s, s + ds) \mid \theta_i = \pm 1]. \tag{13}$$

Although (10) seems to be asymmetric with respect to the class distributions, it is not difficult to recognize that the result is the same if we interchange $f_1(x)$ and $f_2(x)$.

Using expression (10), we readily prove the stated conjecture for $d = 1$. Indeed, if $f_1(x) \neq f_0(x)$, we always have $P(n, 1) >$

$P_0(n, 1) = n/2^{n-1}$, since the strict inequality $[1 + \psi(s)]^{n-1} + [1 - \psi(s)]^{n-1} > 2$ must hold for some values of s.

REFERENCES

[1] T. M. Cover, "Geometrical and statistical properties of systems of linear equations with applications in pattern recognition," *IEEE Trans. Electron. Comput.*, pp. 326-334, June 1965.

[2] J. T. Tou and R. C. Gonzales, *Pattern Recognition Principles.* London, England: Addison-Wesley, 1979.

On the Effect of Noise on the Moore-Penrose Generalized Inverse Associative Memory

G. S. STILES AND DONG-LIH DENQ

***Abstract*—Monte Carlo simulations of the continuous Moore-Penrose generalized inverse associative memory (Kohonen [1]) have shown that the noise-to-signal ratio is improved on recall in the autoassociative case as long as the number of vector pairs stored is less than the number of components per vector. In the heteroassociative case, however, the noise-to-signal ratio may actually be greatly increased upon recall, particularly as the number of vector pairs stored approaches the number of components per vector. The increase in output noise-to-signal ratio in the heteroassociative case is found to be due to the fact that the inverse of the product of the key vector matrix with its transpose may increase without bound in spite of the fact that the key vectors are linearly independent.**

***Index Terms*—Associative memory, associative recall, correlation matrix memory.**

I. INTRODUCTION

Associative memories are systems that allow the recall of data by the (possibly partial) specification of a key related to the data item sought. If the key is identical to the data item the recall operation is termed autoassociative; if the key differs from the data item the recall is heteroassociative (Kohonen [1]). Associative memories (AM) of the type we shall be discussing may be further classified as continuous or discrete (Murakami and Aibara [3]), depending upon whether the items stored are composed of elements that may take on a continuous range of values or only values from a finite set.

Kohonen [1] and Kohonen and Ruohonen [2] have proposed the Moore-Penrose generalized inverse as a mechanism for implementing an associative memory. In this model the kth item to be stored consists of two parts, the key vector x and the data vector y (in the autoassociative case the paired x's and y's are identical). Each vector consists of nc components (we have assumed for convenience that the key and data vectors are of the same length; this assumption is not necessary). The number of vector pairs stored is nv. The nv key and data vectors are represented as column vectors in the key matrix X and the data matrix Y. X_{ik} thus represents the ith component of the kth key vector.

A recall matrix W is computed by the formula

$$W = YX^+ \tag{1}$$

where X^+ is the Moore-Penrose generalized inverse (Ben-Israel and Greville [6]) of the matrix X. X^+ can in general be found from the singular-value decomposition of X (Noble and Daniel [7]). In the case that the column vectors of X are linearly independent, X^+ may be found from

$$X^+ = (X^T X)^{-1} X^T. \tag{2}$$

In the simulations discussed below we have used an algorithm described by Rust *et al.* [4] to find the generalized inverse.

The recall operation is performed by multiplying an input vector x' (which may be a corrupted version of one of the key vectors) by the recall matrix W. The product (output) vector $z = Wx'$ should then be "close" to the data vector y paired with the key vector x that is "closest" to the input vector x'. In the case that the key vectors are linearly independent and the input vector x' is identical to one of the keys x, the output vector z will be identical to the data vector y paired with the key vector x. If the keys are not linearly independent then the output z will be the closest approximation (in the least squares sense) to the desired vector y.

The least squares measure of closeness is equivalent to minimizing the square of the Euclidean norm of the difference of the two vectors in question. The square of the norm is simply the dot product of the difference with itself, $(z - y) \cdot (z - y)$, and can be interpreted as the noise power in z when compared to y. Ideally, the associative recall operation will result in an output noise-to-signal ratio that is smaller than the input noise-to-signal ratio.

Kohonen [1] analyzed rigorously the improvement in noise-to-signal ratio for the autoassociative AM. Using the fact that the autoassociative recall matrix is the orthogonal projection operator on the space spanned by the key vectors, he proved that the output noise-to-signal ratio should be nv/nc times the input noise-to-signal ratio.

In the autoassociative case the output vector is a linear combination of the original data vectors (or of the key vectors, since the key and data vectors are identical), with the data vector closest to the input expected to be the dominant term in the sum. In the heteroassociative case the output can also be expressed as a linear combination of the original data vectors and the coefficients in this combination are identical to those which would have been obtained if the operation had been autoassociative. Based on this observation, Kohonen [1] reasoned that the heteroassociative operation should show the same improvement in noise-to-signal ratio as the autoassociative. We have found, however, that this is not always the case.

We first came across this problem in a study of several discrete AM schemes (Stiles and Denq [5]). Using Monte Carlo methods we evaluated the recall accuracy of the discrete generalized inverse AM operating with vectors of components restricted to the values -1 and $+1$. The recall operation was performed as described above with the addition that the components of the output z were quantized to -1 or $+1$ depending upon whether they were respectively less than, greater than, or equal to zero. In the autoassociative case we found that the noise-to-signal ratios followed the result derived by Kohonen; i.e., the output noise-to-signal ratio was nv/nc times the input ratio.

In the heteroassociative case, however, we found a great discrepency in the vicinity of $nv = nc$. In this region the output noise-to-signal ratio could become many times that of the in-

Manuscript received November 18, 1983; revised November 19, 1984. This work was supported in part by a grant from the Office of the Vice-President for Research, Utah State University, Logan, UT 84322.

G. S. Stiles is with the Department of Electrical Engineering, Utah State University, Logan, UT 84322, on leave at the Department of Electrical and Computer Engineering, Syracuse University, Syracuse, NY 13210.

D.-L. Denq is with the Department of Electrical Engineering, Utah State University, Logan, UT 84322.

0162-8828/85/0500-0358$01.00 © 1985 IEEE