ON CONVERGENCE PROOFS FOR PERCEPTRONS

A. Novikoff
Stanford Research Institute
Menlo Park, California

One of the basic and most proved theorems of perceptron theory is the convergence, in a finite number of steps, of an error correction procedure for an α-perceptron to a classification or dichotomy of the stimulus world, providing such a dichotomy is within the combinatorial capacities of the perceptron. In other words, if a solution exists, error correction will find one in a finite time. A proof is presented which is substantially shorter and more transparent than those now available, and which isolates out the principle on which the theorem depends. We believe this principle will find further use in other theorems of a similar nature.

- I -

The purpose of this paper is to exhibit an extremely short, and more notably, transparent proof of a theorem concerning perceptrons. The theorem itself must now be considered one of the most basic theorems concerning perceptrons, and indeed, is among the first theorems proved by Rosenblatt and his collaborators. It also enjoys the peculiar distinction of being one of the most often re-proved results in the field (see "References" at the end of this paper). The succession of proofs now available progresses from somewhat cloudy statements (which at one time caused doubt among "reasonable men" that the theorem was, in fact, true), to comparatively crisp statements of a purely mathematical nature which nonetheless use more print than is strictly necessary. [1-6]

More to the point, latter-day proofs fail to enunciate a simple principle involved. This principle permits one to modify the hypotheses in a variety of ways and still secure the results; and it may well be useful in establishing genuinely new theorems of like character. I therefore present this proof in its entirety, in part to verify my claim that it is as short a line as can be drawn from hypotheses to conclusion, and also with the hope of terminating an already lengthy process of successive refinements.

Whereas previous proofs of the present theorem appealed to a structure, called by Rosenblatt and his co-workers an α-perceptron, the present proof and ensuing discussion apply without modification to a structure consisting of a single threshold element acting on a weighted set of inputs.

Presented at the *Symposium on Mathematical Theory of Automata*, Polytechnic Institute of Brooklyn, April 24, 25, 26, 1962

- II -

In the private language of perceptron workers, the theorem considers a given α-type perceptron and a set of incoming signals divided into two disjoint classes. A "satisfactory" assignment of output weights for the associator units is defined as an assignment resulting in a response +1 for signals of Class I and -1 for signals of Class II. The theorem under discussion then states that, no matter what assignment of weights we begin with, the process of recursively readjusting the weights by the method known as "error correction" will terminate after a finite number of corrections in a satisfactory assignment, providing any such satisfactory assignment exists. More briefly, a finite number of corrections will teach the perceptron to perform any given dichotomy of signals, if the dichotomy is within the capacity of the perceptron at all.

- III -

This result can be translated into universally acceptable mathematical language. Using the formulation of reference 1, the theorem reads as follows: We are given a set of vectors, $w_1, \ldots, w_N$, in some fixed finite dimensional Euclidean space. These are assumed to satisfy the single hypothesis that there exists a vector y such that

$$(w_i, y) > \Theta > 0, \qquad i = 1, \ldots, N \quad . \tag{1}$$

We then consider an infinite sequence

$$w_{i_1}, w_{i_2}, w_{i_3}, \ldots, \quad (1 \leq i_k \leq N \text{ for every } k) \ ,$$

such that each vector, $w_1, \ldots, w_N$, occurs infinitely often. We now construct a sequence of vectors, $v_0, v_1, \ldots, v_N, \ldots$, recursively as follows:

v_0 is arbitrary,

and

$$v_n = \begin{cases} v_{n-1} & \text{if } (v_{n-1}, w_{i_n}) > \Theta \\ v_{n-1} + w_{i_n} & \text{if } (v_{n-1}, w_{i_n}) \leq \Theta \end{cases} \tag{2}$$

The assertion of the theorem is that the sequence $\{v_n\}$ is convergent, which in this case really means that for some index m we have

$$v_m = v_{m+1} = v_{m+2} = \ldots = \tilde{v} \quad .$$

In particular $(\tilde{v}, w_i) > \theta$ for $i = 1, \ldots, N$, since each w_i occurs arbitrarily far out in the sequence $\{w_{i_k}\}$. (It is only to obtain this consequence that we impose the restriction that each w_i occurs infinitely often in the training sequence.) It can be argued that, theoretically, there is no loss of generality in taking $\theta = 0$.[1] While this is true, it often has the effect of smuggling in numbers of large magnitude. For this reason, we retain the general θ, but in the concluding paragraph we do consider the relation between the general case and the case $\theta = 0$.

The precise correspondence between the theorem's original verbal description and the above purely mathematical assertion is provided in reference 1, where the definition of α-perceptron is also given. For those already familiar with these notions and anxious to interpret the following discussion in perceptron terms, let me give a brief glossary describing the correspondence here: the vectors $\{w_i\}$ represent the columns of a matrix (b_{ij}) which is fundamental to the description of an α-perceptron. The vector y represents the "satisfactory" assignment of associator outputs which we assume to exist: y has as many components as there are associators. The sequence $\{w_{i_n}\}$ represents the "training sequence," and the rule for defining $\{v_n\}$ describes the error-correction procedure. The positive number θ is a threshold which must be exceeded for the response of the perceptron to be correct: a vector v such that $(w_i, v) > \theta$ is an assignment of associator outputs which successfully dichotomizes the i-th signal. (If $(w_i, v) \leq \theta$, then the i-th signal has either been identified as in the incorrect class or the perceptron has refrained from a commitment, depending on whether or not we have strict inequality.)

- IV -

The proof of the theorem is greatly facilitated by considering only those indices n such that $v_{n+1} \neq v_n$. If we drop out of the training sequence all terms for which this is not the case, and which are clearly inessential, we are left with a training sequence for which correction takes place at every step. Readapting our notation, we may then assume without loss of generality that

$$v_n = v_{n-1} + w_{i_n}, \tag{3}$$

and for each n,

$$(v_{n-1}, w_{i_n}) \leq \theta \tag{4}$$

Now the assertion is that n, which in this notation counts the number of corrections up to the n-th step, can only range through a

finite set of integers. That is, Eqs. (1) and (4) cannot continue to hold simultaneously for all $n = 1, 2, 3, \ldots$.

Indeed, (1) alone implies that $v_n = v_0 + w_{i_1} + \ldots + w_{i_n}$ satisfies $(v_n, y) \geq (v_0, y) + n\theta$. Since $(v_n, y)^2 \leq \|v_n\|^2 \cdot \|y\|^2$, this implies that

$$\|v_n\|^2 \geq Cn^2 \tag{5}$$

for suitable choice of the positive constant C, providing n is sufficiently large.

On the other hand, (4) alone implies that the integer-argument function $\|v_n\|^2$ satisfies the difference inequality

$$\|v_n\|^2 - \|v_{n-1}\|^2 = 2(v_{n-1}, w_{i_n}) + (w_{i_n}, w_{i_n}) \leq 2\theta + M,$$

where $M = \max(w_{i_n}, w_{i_n})$. Adding these inequalities for $n = 1, 2, 3, \ldots$, we obtain

$$\|v_n\|^2 \leq \|v_0\|^2 + (2\theta + M)n. \tag{6}$$

Clearly, (5) and (6) are incompatible for n sufficiently large.

- V -

The above argument is the discrete analog of the following theorem, which may seem more intuitive. Let v(t) be a curve in m-space described by a smooth vector function of the continuous variable t, such that

$$\left(\frac{dv}{dt}, y\right) \geq c > 0 \quad \text{for some fixed vector } y \tag{1}'$$

(i. e., the tangent vector to the curve lies on one side of a hyperplane); and

$$\frac{1}{2}\frac{d}{dt}\|v(t)\|^2 = \left[v(t), \frac{dv}{dt}\right] \leq \theta, \quad 0 \leq t < b \tag{2}'$$

(that is, $\|v\|^2$ grows at a bounded rate). Then an upper bound can be given for b (and in particular (1)' and (2)' are only compatible over a finite domain on the t-axis). The proof is virtually identical with the discrete case: integrating (1)' between 0 and t we obtain

$$[v(t), y] \geq [v(0), y] + ct, \tag{7}$$

while integrating the differential inequality (2)' we obtain

$$\left|\left|v(t)\right|\right|^2 \leq 2\theta t + \left|\left|v(0)\right|\right|^2 . \tag{8}$$

Since

$$\left|\left|v(t)\right|\right|^2 \geq \frac{\left[v(t), y\right]^2}{\left|\left|y\right|\right|^2} \geq \frac{\left\{\left[v(0), y\right] + ct\right\}^2}{\left|\left|y\right|\right|^2} \tag{7}'$$

by Schwartz's inequality and then (7), we see that t cannot exceed the larger root of the quadratic

$$\left\{\left[v(0), y\right] + ct\right\}^2 = \left|\left|y\right|\right|^2 \cdot \left[2\theta t + \left|\left|v(0)\right|\right|^2\right] .$$

(It may be of interest to compare the above argument for the continuous case with the extremely familiar phenomenon that

$$\left[v(t), \frac{dv}{dt}\right] = 0 \tag{9}$$

implies

$$\left|\left|v(t)\right|\right|^2 = \text{constant} ; \tag{8}'$$

that is, a curve whose tangent vector is always perpendicular to its position vector is constrained to lie on a sphere. Replacing the orthogonality condition (9) with the inequality (2)' results in an inequality (8) on the rate of growth of the function $f(t) = \left|\left|v(t)\right|\right|^2$, a weakening of the condition (8)' that f(t) be constant.)

The principle involved here is this: The condition of (2)', that the tangent vector have bounded scalar product with the position vector, clearly results in an upper bound for the instantaneous position of the curve as a function of time. On the other hand, if the tangent vectors dv/dt of the curve remain sufficiently large and do not depart too much from colinearity (as prescribed, for example, by (1)'), then a lower bound on the cumulative growth results, as in (7)'. This is intuitively clear: if dv/dt does not get too small, the total arc-length will increase with at least a certain rate. If, in addition, dv/dt is sufficiently "nearly colinear" then the serpentine path swept out by v(t) cannot reverse its direction enough to prevent its overall migration away from its starting point. The opposition of these two influences implies the termination of the process. We will not dwell on the point of how assorted variations on this theme will continue to produce theorems asserting that t, or in the discrete case n, must remain bounded. Whether each of these deserves to be dignified with the name "theorem" is a moot point.

- VI -

We conclude with a few words about the geometrical interpretation of the assumption (1). These are stated without proof, and assume familiarity with the theory of convex sets in Euclidean vector spaces. The condition that there exists a vector y satisfying (1) is precisely equivalent to the fact that the polyhedral cone C with generators $w_1, \ldots, w_N$, defined as all vectors of the form $\lambda_1 w_1 + \ldots + \lambda_N w_N$, $\lambda_1 \geq 0, \ldots, \lambda_N \geq 0$, is a proper cone (i.e., that C never contains both v and $-v$ except for $v = 0$, or equivalently, C, apart from its vertex, lies in the *interior* of a half space). It should be remarked that the existence of a vector y satisfying (1) with $\Theta > 0$ is neither stronger nor weaker than the requirement of the existence of a vector $\tilde{y}$ satisfying (1) with $\Theta = 0$, i.e.,

$$(w_i, \tilde{y}) > 0 \quad . \tag{1}''$$

In fact, y itself can serve for $\tilde{y}$ and conversely, given $\tilde{y}$ satisfying (1)" any sufficiently large positive multiple $y = \lambda\tilde{y}$, for

$$\lambda > \frac{\Theta}{\min\limits_{i=1,\ldots,N} (w_i, \tilde{y})}$$

will satisfy (1)'. Condition (1)" is the customary way of specifying that C be proper. For the continuous analog of (1)', the requirement that infinitely many vectors $v'(t)$ satisfy (1)' is actually stronger than requiring

$$\left[v(t), y\right] > 0, \quad 0 \leq t < b \, . \tag{1}*$$

In fact, the left-hand side, though positive for each t, need not be bounded away from zero. If (1)* is assumed to hold for $0 \leq t \leq b$ (equality permitted at b) and $v'(t)$ is assumed continuous, then, as in the finite case, it is again true that (1)* and (1)' are precisely equivalent.

The cone C^* of all vectors v such that $(w, v) \geq 0$ for all w in C or equivalently such that

$$(w_i, v) \geq 0 \, , \quad i = 1, \ldots, N \tag{10}$$

is called the *dual* cone to C; its interior consists of all v for which every inequality in (10) is strict. The bigger C is, the smaller C^* is, and vice versa, but in general, neither need include the other. For example, when C is a half space, C^* is a half line; when C is proper, C^* has an interior, and indeed the $\tilde{y}$ of (1)'' is in the

interior of C^*. If C^* has an interior, C and the interior of C^* overlap. As already observed, if y is in the interior of C^*, then a suitable positive multiple of y will satisfy (1). Let the set of vectors satisfying (1) be denoted by D, which is a subset of the interior of C^*. The relation between (1) and the dual cone defined by (10) may be summarized as follows: (1) has a solution (that is, D is non-empty) if and only if C^* has an interior.

The error correction procedure is a *recursive construction* of a vector in D of the form

$$k_1 w_1 + \ldots + k_N w_N$$

where k_i is the number of times w_i occurred in the irredundant training sequence before the termination of the correction process. The mere existence of such a vector (without a construction algorithm) is assured by the fact that C and D overlap.

In general, if (1) is fulfilled, so that $w_1, \ldots, w_N$ generate a proper cone, it will happen that some subset will generate the same cone (which then has the same dual C^*), and it suffices to restrict attention to this subset in constructing the training sequence. To accelerate the termination of the correction process, one should use for correction those w's which are themselves nearest the interior of C^*, and which are as long as possible. If, for example, $w_3 = w_1 + w_2$, the single addition of w_3 will accomplish as much as the successive additions of w_1 and w_2. The question of whether a dichotomy is within the combinatorial capacity of an α-type perceptron reduces to whether or not C^* has an interior, or equivalently, whether or not C is proper. This question is discussed in a somewhat different context by Joseph and Hay[7] and Keller.[8]

REFERENCES

1. H.D. Block, "The Perceptron: A Model for Brain Functioning. I," *Rev. Modern Phys.*, Vol. 34, No. 1 (January 1962).
2. F. Rosenblatt, Report VG-1196-G-4, Cornell Aeronautical Laboratory (February 1960).
3. R.D. Joseph, Tech. Memo 12, Project PARA, Cornell Aeronautical Laboratory (May 1960).
4. R.D. Joseph, Tech. Memo 13, Project PARA, Cornell Aeronautical Laboratory (July 1960).
5. R.A. Stafford, "Learning by Threshold Elements in Cascade," Tech. Note 20081, Aeronutronics (8 March 1961).
6. R.C. Singleton, "A Test for Linear Separability as Applied to Self-Organizing Machines," paper presented at Conference on Self-Organizing Systems, May 22-24, 1962, Chicago, Illinois (*Proceedings* in press, Spartan Books).

7. T.F. Joseph and L. Hay, Tech. Memo 8, Project PARA, Cornell Aeronautical Laboratory (1960).

8. H.B. Keller, "Finite Automata, Pattern Recognition and Perceptron," *J. Assoc. Comp. Mach.*, Vol. 8, No. 1 (January 1961).