Perceptron Convergence Theorem

(Rosenblatt early form, without learning rate)
嚴謹證明與幾何詮釋（含連續版本）

導讀（Zh-TW）. 本文以早期文獻的表述重建「感知機收斂定理」的嚴謹證明：假設資料在有限維歐幾里得空間中可由一個閾值為 $\theta > 0$ 的超平面嚴格分開，則使用「錯誤修正法」更新權重，更新次數必為有限，因而演算法在有限步後停止（收斂）。我們先給出離散版核心證明，接著給出連續對應（作為直覺類比），最後以凸錐（cone）與對偶錐（dual cone）給出幾何詮釋。全文證明部分以 English 撰寫，避免邏輯歧義；中文僅作註釋與說明。全程**不使用學習率**，完全對應早期敘述。

1 Setting and Notation

Let $w_1, \ldots, w_N \in \mathbb{R}^m$ be a finite set of nonzero vectors. Assume there exists a vector $y \in \mathbb{R}^m$ and a threshold $\theta > 0$ such that

$$\langle w_i, y \rangle > \theta \qquad \text{for all } i = 1, \ldots, N. \tag{1}$$

Consider a (possibly infinite) training sequence in which each w_i occurs infinitely often. Let $v_0 \in \mathbb{R}^m$ be arbitrary. The perceptron with *error-correction* update is: when the current sample is w and $\langle v, w \rangle \leq \theta$ (mistake or not confident enough), set $v \leftarrow v + w$; otherwise keep v unchanged.

As is classical (and done in Rosenblatt's derivation), it suffices to restrict attention to the subsequence of *actual updates*. Index these updates by $n = 1, 2, \ldots$, and denote the misclassified sample at step n by $w_n \in \{w_1, \ldots, w_N\}$. Then

$$v_n = v_{n-1} + w_n, \qquad \langle v_{n-1}, w_n \rangle \leq \theta \quad (n \geq 1). \tag{2}$$

Define $M := \max_i \|w_i\|^2$.

Remark 1 (中文註解). 我們只保留「真的有更新」的步驟，不會影響是否收斂的結論。以上兩式正是早期文獻在簡化後使用的核心不等式。

2 Discrete Perceptron Convergence Theorem

Theorem 1 (Discrete form)**.** *Suppose* (1) *holds for some y and $\theta > 0$. Then the update rule* (2) *can occur only finitely many times. Equivalently, there exists $m < \infty$ such that $v_n = v_m$ for all $n \geq m$.*

Proof. First, by (1) and (2),

$$\langle v_n, y\rangle = \langle v_{n-1}, y\rangle + \langle w_n, y\rangle > \langle v_{n-1}, y\rangle + \theta, \tag{3}$$

hence, inductively,

$$\langle v_n, y\rangle \geq \langle v_0, y\rangle + n\theta \qquad (n \geq 1). \tag{4}$$

By Cauchy–Schwarz, $\langle v_n, y\rangle^2 \leq \|v_n\|^2 \|y\|^2$, so (4) implies the *quadratic* lower bound

$$\|v_n\|^2 \ \geq \ \frac{(\langle v_0, y\rangle + n\theta)^2}{\|y\|^2}. \tag{5}$$

On the other hand, expanding the difference of squared norms and using (2),

$$\|v_n\|^2 - \|v_{n-1}\|^2 = 2\ \langle v_{n-1}, w_n\rangle + \|w_n\|^2 \leq 2\theta + M.$$

Summing from 1 to n yields the *linear* upper bound

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

Combining (5) and (6) we obtain, for every n at which an update occurs,

$$\frac{(\langle v_0, y\rangle + n\theta)^2}{\|y\|^2} \ \leq \ \|v_0\|^2 + (2\theta + M)\, n. \tag{7}$$

The left-hand side is quadratic in n with leading coefficient $\theta^2/\|y\|^2 > 0$, while the right-hand side is affine in n. Hence (7) cannot hold for all integers n. In particular, let T be the larger real root of the quadratic equality obtained from (7); then *no* update can occur for any integer $n > T$. Consequently the number of updates is finite, and the process must terminate. $\square$

Remark 2 (Explicit bound). Writing out the larger root gives an explicit (though notationally heavy) bound

$$T \ = \ \frac{\|y\|^2 (2\theta + M) - 2\theta\ \langle v_0, y\rangle + \sqrt{\big(\|y\|^2 (2\theta + M) - 2\theta\ \langle v_0, y\rangle\big)^2 - 4\theta^2\big(\langle v_0, y\rangle^2 - \|y\|^2 \|v_0\|^2\big)}}{2\theta^2},$$

hence the total number of updates is at most $\lceil T \rceil$.

3 Continuous Analog (for intuition)

Consider a smooth curve $v : [0, b) \to \mathbb{R}^m$ such that for some fixed y and constants $c > 0$, $\theta \in \mathbb{R}$,

$$\langle \dot{v}(t), y\rangle \geq c \quad \text{for } 0 \leq t < b, \tag{8}$$

$$\frac{1}{2}\frac{d}{dt}\|v(t)\|^2 = \langle v(t), \dot{v}(t)\rangle \leq \theta \quad (0 \leq t < b). \tag{9}$$

Integrating (8) gives $\langle v(t), y\rangle \geq \langle v(0), y\rangle + ct$. Cauchy–Schwarz then yields $\|v(t)\|^2 \geq \{\langle v(0), y\rangle + ct\}^2/\|y\|^2$. Integrating (9) gives $\|v(t)\|^2 \leq 2\theta t + \|v(0)\|^2$. As in the discrete case, the resulting quadratic vs. linear growth are incompatible for large t; hence t is bounded above. This provides a faithful continuous analog of the discrete proof.

4 Geometric Interpretation via Cones

Define the (finitely generated) convex cone

$$C := \Big\{ \sum_{i=1}^{N} \lambda_i w_i \;\Big|\; \lambda_i \geq 0 \Big\},$$

and its dual cone

$$C^* := \{ v \in \mathbb{R}^m : \langle w_i,\, v \rangle \geq 0 \text{ for all } i \}.$$

Proposition 1. *The separability condition* (1) *holds for some y and $\theta > 0$ if and only if the dual cone C^* has nonempty interior (equivalently, C is a proper cone).*

Proof sketch. ($\Rightarrow$) If $\langle w_i,\, y \rangle > \theta > 0$ for all i, then in particular $\langle w_i,\, y \rangle > 0$; hence $y \in \mathrm{int}(C^*)$. ($\Leftarrow$) If $y \in \mathrm{int}(C^*)$, then $\min_{1 \leq i \leq N} \langle w_i,\, y \rangle =: m > 0$. For any $\theta \in (0, m)$, (1) holds. The equivalence to C being proper is standard: C is proper (pointed) iff C^* has nonempty interior. □

Remark 3 (Error-correction as a constructive path into C^*). The update sequence $v_n = \sum_{i=1}^{N} k_i w_i$ (with integers $k_i \geq 0$ counting updates on each w_i) can be viewed as a recursive construction steering v_n toward $\mathrm{int}(C^*)$. Prioritizing samples w that are "closest" to $\mathrm{int}(C^*)$ (and of larger norm) accelerates termination—this echoes margin-based heuristics in modern treatments.

5 Auxiliary Lemmas (Completeness)

Lemma 1 (Cauchy–Schwarz). *For all $u, v \in \mathbb{R}^m$, $|\langle u,\, v \rangle| \leq \|u\|\,\|v\|$.*

Lemma 2 (Quadratic vs. affine dominance). *Let $a > 0$, $b, c \in \mathbb{R}$. The inequality $an^2 + bn + c \leq \alpha n + \beta$ cannot hold for all integers n.*

Proof. Rearranging gives a quadratic with positive leading coefficient; such a polynomial tends to $+\infty$, contradicting the affine upper bound for large n. □

6 Concluding Notes

The proof of Theorem 1 requires neither step size parameters nor stochastic assumptions; it only uses: (i) strict separability (1), (ii) the update subsequence (2), (iii) Cauchy–Schwarz, and (iv) a simple telescoping bound on squared norms. Hence it matches the early perceptron convergence theorem in spirit and logic.

Keywords: Perceptron, error-correction, convergence, convex cone, dual cone, separability.