Proposition 1. *If the sets P and N are finite and linearly separable, the perceptron learning algorithm halts on a solution vector $\mathbf{w}_{t+1}$.*

Proof. Without loss of generality, consider the set $P' = P \cup N^-$ where $N^- = \{ -\mathbf{n} \mid \mathbf{n} \in N \}$. We can do this because a plane that separates P and N separates $\varnothing$ and P'.

Fix a solution vector $\mathbf{w}^*$. After $t+1$ steps a new weight vector $\mathbf{w}_{t+1} = \mathbf{w}_t + \mathbf{p}_{i(t)}$ has been computed ($i(t)$ is the index of the data vector $\mathbf{p}$ picked aon step t).

Recall that the cosine of the angle ρ between $\mathbf{w}^*$ and $\mathbf{w}_{t+1}$ is:

$$\cos\rho = \frac{\mathbf{w}^* \cdot \mathbf{w}_{t+1}}{\|\mathbf{w}^*\|\|\mathbf{w}_{t+1}\|} \tag{1}$$

Then in the numerator we have:

$$\begin{aligned} \mathbf{w}^* \cdot \mathbf{w}_{t+1} &= \mathbf{w}^* \cdot (\mathbf{w}_t + \mathbf{p}_{i(t)}) \\ &= \mathbf{w}^* \cdot \mathbf{w}_t + \mathbf{w}^* \cdot \mathbf{p}_{i(t)} \\ &\geq \mathbf{w}^* \cdot \mathbf{w}_t + \delta \text{ where } \delta = \min\{ \mathbf{w}^* \cdot \mathbf{p} \mid \forall \mathbf{p} \in P' \} \end{aligned}$$

Clearly $\mathbf{w}^* \cdot \mathbf{p}_{i(t)} \geq \delta$, $\forall t$. So by induction we have

$$\mathbf{w}^* \cdot \mathbf{w}_{t+1} \geq \mathbf{w}^* \cdot \mathbf{w}_0 + (t+1)\delta. \tag{2}$$

To see this we can expand the first few terms. Notice it's the same as in the normalized case:

$$\begin{aligned} \mathbf{w}^* \cdot \mathbf{w}_1 = \mathbf{w}^* \cdot (\mathbf{w}_0 + \mathbf{p}_{i(0)}) &= \mathbf{w}^* \cdot \mathbf{w}_0 + \mathbf{w}^* \cdot \mathbf{p}_{i(0)} \\ &\geq \mathbf{w}^* \cdot \mathbf{w}_0 + \delta \\ \mathbf{w}^* \cdot \mathbf{w}_2 = \mathbf{w}^* \cdot (\mathbf{w}_1 + \mathbf{p}_{i(1)}) &= \mathbf{w}^* \cdot \mathbf{w}_1 + \mathbf{w}^* \cdot \mathbf{p}_{i(1)} \\ &\geq \mathbf{w}^* \cdot \mathbf{w}_0 + \delta + \mathbf{w}^* \cdot \mathbf{p}_{i(1)} \\ &\geq \mathbf{w}^* \cdot \mathbf{w}_0 + (2)\delta \\ \mathbf{w}^* \cdot \mathbf{w}_3 = \mathbf{w}^* \cdot (\mathbf{w}_2 + \mathbf{p}_{i(2)}) &= \mathbf{w}^* \cdot \mathbf{w}_2 + \mathbf{w}^* \cdot \mathbf{p}_{i(2)} \\ &\geq \mathbf{w}^* \cdot \mathbf{w}_0 + (2)\delta + \mathbf{w}^* \cdot \mathbf{p}_{i(2)} \\ &\geq \mathbf{w}^* \cdot \mathbf{w}_0 + (3)\delta \\ &\vdots \\ \mathbf{w}^* \cdot \mathbf{w}_{t+1} = \mathbf{w}^* \cdot (\mathbf{w}_t + \mathbf{p}_{i(t)}) &= \mathbf{w}^* \cdot \mathbf{w}_t + \sum_{j=0}^{t} \mathbf{w}^* \cdot \mathbf{p}_{i(j)} \\ &\geq \mathbf{w}^* \cdot \mathbf{w}_0 + (t+1)\delta \end{aligned}$$

For the denominator $\|\mathbf{w}^*\|\|\mathbf{w}_{t+1} = k\|\mathbf{w}_{t+1}\|$, consider:

$$\begin{aligned}\|\mathbf{w}_{t+1}\|^2 &= (\mathbf{w}_t + \mathbf{p}_{i(t)}) \cdot (\mathbf{w}_t + \mathbf{p}_{i(t)}) \\ &= \|\mathbf{w}_t\|^2 + 2\mathbf{w}_t \cdot \mathbf{p}_{i(t)} + \|\mathbf{p}_{i(t)}\|^2\end{aligned}$$

Here is a difference from the normalized case. $\|\mathbf{p}_{i(t)}\|^2$ is not necessarily equal to 1, so let $\epsilon = \max\{\, \|\mathbf{p}_{i(t)}\|^2 \mid \forall \mathbf{p} \in P \,\}$. Then $\|\mathbf{p}_{i(t)}\|^2 \leq \epsilon$, $\forall t$. As in the normalized case, $2\mathbf{w}_t \cdot \mathbf{p}_{i(t)} \leq 0$ since prior to correction $\mathbf{p}_{i(t)}$ either lies on or "behind" the hyperplane.
Therefore, $\|\mathbf{w}_{t+1}\|^2 \leq \|\mathbf{w}_t\|^2 + \epsilon$ and by induction,

$$\|\mathbf{w}_{t+1}\|^2 \leq \|\mathbf{w}_0\| + (t+1)\epsilon. \tag{3}$$

Then from (1), (2), (3), we get

$$\cos\rho \geq \frac{\mathbf{w}^* \cdot \mathbf{w}_0 + (t+1)\delta}{k\sqrt{\|\mathbf{w}_0\|^2 + (t+1)\epsilon}}$$

Since $\sqrt{t}$ is monotonically increasing and unbounded, and $\delta > 0$ (because we are looking for an *absolute* linear separation), the RHS can become arbitrarily large. However, the LHS is bound by 1:

$$1 \geq \cos\rho \geq \frac{\mathbf{w}^* \cdot \mathbf{w}_0 + (t+1)\delta}{\|\mathbf{w}^*\|\sqrt{\|\mathbf{w}_0\|^2 + (t+1)\epsilon}} \propto \frac{t}{\sqrt{t}} = \sqrt{t}$$

So t must have some maximum value, thus the algorithm halts. Since the algorithm only halts on a solution vector, the algorithm finds a solution. □

Having $\delta > 0$ is important. If we allowed non-absolute linear separability, then the fraction could asymptotically approach 0 and in those cases the algorithm would never halt!

Proof. The PLA works because all solutions have an associated neighborhood of solutions:

$$\begin{aligned}&(\forall \mathbf{w}_j^* \,\exists\, \theta_j \,\forall \mathbf{w}_j (\mathbf{w}_j^* \in \text{solutions}(P, N) \,\wedge\, |\mathbf{w}_j^* \angle \mathbf{w}_j| < \theta_j) \implies \mathbf{w}_j \in \text{solutions}(P, N)) \\ &\wedge ((t \to \infty \implies \rho \to 0) \implies |\rho_{t+1}| < |\theta^*|) \\ &\therefore \mathbf{w}_{t+1} \in \text{solutions}(P, N)\end{aligned}$$

□