Convergence of Perceptron Learning Algorithm

Hsuan-Tien Lin, September 26, 2020

Page 14 of lecture 2 proves that

$$\mathbf{w}_f^T \mathbf{w}_{t+1} \geq \mathbf{w}_f^T \mathbf{w}_t + \underbrace{\min_n y_n \mathbf{w}_f^T \mathbf{x}_n}_{\|\mathbf{w}_f\| \cdot \rho},$$

where we defined the last term to be a constant $\|\mathbf{w}_f\| \cdot \rho$ on page 16 of lecture 2. Assume that $\mathbf{w}_0 = \mathbf{0}$, as set on page 15 of lecture 2. Then, we have

$$\begin{array}{rcl}
\mathbf{w}_f^T \mathbf{w}_0 & = & \square \\
\mathbf{w}_f^T \mathbf{w}_1 & \geq & \mathbf{w}_f^T \mathbf{w}_0 + \|\mathbf{w}_f\| \cdot \rho \\
\mathbf{w}_f^T \mathbf{w}_2 & \geq & \square + \|\mathbf{w}_f\| \cdot \rho \\
\square & \geq & \square + \|\mathbf{w}_f\| \cdot \rho \\
\cdots & & \cdots \\
\square & \geq & \mathbf{w}_f^T \mathbf{w}_{T-1} + \|\mathbf{w}_f\| \cdot \rho
\end{array}$$

Summing all inequalities above, we get

$$\mathbf{w}_f^T \mathbf{w}_T \geq \square \cdot \|\mathbf{w}_f\| \cdot \rho \tag{1}$$

This says the inner product grows by at least $\square \cdot \|\mathbf{w}_f\| \cdot \rho$ after T updates. Note that for linear separable data with both classes of examples, a separating $\mathbf{w}_f$ means $\|\mathbf{w}_f\| > 0$ and $\rho > 0$.

Now let's look at the results on page 15 of lecture 2. It proves that

$$\|\mathbf{w}_{t+1}\|^2 \leq \|\mathbf{w}_t\|^2 + \underbrace{\max_n \|\mathbf{x}_n\|^2}_{R^2},$$

where we defined the last term to be a constant R^2 on page 16 of lecture 2. Assume that $\mathbf{w}_0 = \mathbf{0}$, as set on page 15 of lecture 2. Then, we have

$$\begin{array}{rcl}
\|\mathbf{w}_0\|^2 & = & \square \\
\|\mathbf{w}_1\|^2 & \leq & \|\mathbf{w}_0\|^2 + R^2 \\
\|\mathbf{w}_2\|^2 & \leq & \square + R^2 \\
\square & \leq & \square + R^2 \\
\cdots & & \cdots \\
\square & \leq & \|\mathbf{w}_{T-1}\|^2 + R^2
\end{array}$$

Summing all inequalities above, we get

$$\|\mathbf{w}_T\|^2 \leq \square \cdot R^2 \tag{2}$$

This says the squared length grows by at most $\square R^2$ after T updates.

If $\|\mathbf{w}_T\|^2 = 0$, this means $\mathbf{w}_f^T \mathbf{w}_T = 0$ as well, which contradicts (1) for $T \geq 1$ because $\|\mathbf{w}_f\| > \square$, $\|\rho\| > \square$, and $\mathbf{w}_f^T \mathbf{w}_0 = \square$. So $\|\mathbf{w}_T\|^2$ must be strictly positive for $T \geq 1$. Then, we can "divide" (1) by the "square root" of (2) to get

$$\frac{\mathbf{w}_f^T \mathbf{w}_T}{\|\mathbf{w}_T\|} \geq \square \cdot \frac{\rho \|\mathbf{w}_f\|}{R}. \tag{3}$$

That is, for linear separable data with both classes of examples, and for $T \geq 1$, let θ_T denote the angle between $\mathbf{w}_f$ and $\mathbf{w}_T$, we have

$$\square \geq \cos\theta_T = \frac{\mathbf{w}_f^T \mathbf{w}_T}{\square \cdot \|\mathbf{w}_T\|} \geq \square \cdot \frac{\rho}{R}$$

This proves that T, the number of updates of PLA, cannot be more than

$$\frac{R^2}{\rho^2},$$

which is the conclusion on page 16 of lecture 2. That is, PLA will converge with no more than $\frac{R^2}{\rho^2}$ updates.