Convergence Proof for the Perceptron Algorithm

Michael Collins

Figure 1 shows the perceptron learning algorithm, as described in lecture. In this note we give a convergence proof for the algorithm (also covered in lecture).

The convergence theorem is as follows:

**Theorem 1** Assume that there exists some parameter vector \( \theta^* \) such that \( ||\theta^*|| = 1 \), and some \( \gamma > 0 \) such that for all \( t = 1 \ldots n \),

\[
y_t (x_t \cdot \theta^*) \geq \gamma
\]

Assume in addition that for all \( t = 1 \ldots n \), \( ||x_t|| \leq R \).

Then the perceptron algorithm makes at most \( \frac{R^2}{\gamma^2} \) errors. (The definition of an error is as follows: an error occurs whenever we have \( y_t' \neq y_t \) for some \( (j,t) \) pair in the algorithm.)

Note that for any vector \( x \), we use \( ||x|| \) to refer to the Euclidean norm of \( x \), i.e.,

\[
||x|| = \sqrt{\sum_i x_i^2}.
\]

**Proof:** First, define \( \theta^k \) to be the parameter vector when the algorithm makes its \( k \)'th error. Note that we have

\[
\theta^1 = \theta = \theta^*
\]

Next, assuming the \( k \)'th error is made on example \( t \), we have

\[
\theta^{k+1} \cdot \theta^* = (\theta^k + y_t x_t) \cdot \theta^*
\]

\[
= \theta^k \cdot \theta^* + y_t x_t \cdot \theta^* \tag{1}
\]

\[
\geq \theta^k \cdot \theta^* + \gamma \tag{2}
\]

Eq. 1 follows by the definition of the perceptron updates. Eq. 3 follows because by the assumptions of the theorem, we have

\[
y_t x_t \cdot \theta^* \geq \gamma
\]
**Definition:** \( \text{sign}(z) = 1 \) if \( z \geq 0 \), \(-1\) otherwise.

**Inputs:** number of iterations, \( T \); training examples \((x_t, y_t)\) for \( t \in \{1 \ldots n\} \) where \( x \in \mathbb{R}^d \) is an input, and \( y_t \in \{-1, +1\} \) is a label.

**Initialization:** \( \theta = 0 \) (i.e., all parameters are set to 0)

**Algorithm:**
- For \( j = 1 \ldots T 
  - For \( t = 1 \ldots n \)
    1. \( y' = \text{sign}(x_t \cdot \theta) \)
    2. If \( y' \neq y_t \) Then \( \theta = \theta + y_t x_t \), Else leave \( \theta \) unchanged

**Output:** parameters \( \theta \)

![Figure 1: The perceptron learning algorithm.](image)

It follows by induction on \( k \) (recall that \( \|\theta^1\| = 0 \)), that

\[
\theta^{k+1} \cdot \theta^* \geq k \gamma
\]

In addition, because \( \|\theta^{k+1}\| \times \|\theta^*\| \geq \theta^{k+1} \cdot \theta^* \), and \( \|\theta^*\| = 1 \), we have

\[
\|\theta^{k+1}\| \geq k \gamma \tag{4}
\]

In the second part of the proof, we will derive an upper bound on \( \|\theta^{k+1}\| \). We have

\[
\|\theta^{k+1}\|^2 = \|\theta^k + y_t x_t\|^2 \tag{5}
\]

\[
= \|\theta^k\|^2 + y_t^2 \|x_t\|^2 + 2 y_t x_t \cdot \theta^k \tag{6}
\]

\[
\leq \|\theta^k\|^2 + R^2 \tag{7}
\]

The equality in Eq. 5 follows by the definition of the perceptron updates. Eq. 7 follows because we have: 1) \( y_t^2 \|x_t\|^2 = \|x_t\|^2 \leq R^2 \) by the assumptions of the theorem, and because \( y_t^2 = 1 \); 2) \( y_t x_t \cdot \theta^k \leq 0 \) because we know that the parameter vector \( \theta^k \) gave an error on the \( t \)th example.

It follows by induction on \( k \) (recall that \( \|\theta^1\|^2 = 0 \)), that

\[
\|\theta^{k+1}\|^2 \leq k R^2 \tag{8}
\]
Combining the bounds in Eqs. 4 and 8 gives

\[ k^2 \gamma^2 \leq \|g^{k+1}\|^2 \leq kR^2 \]

describing which it follows that

\[ k \leq \frac{R^2}{\gamma^2} \]

\[ \square \]