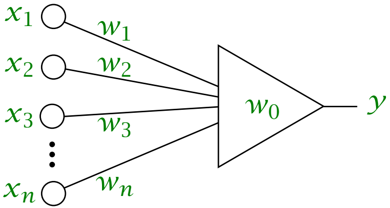
The Perceptron Convergence Theorem

Robert Snapp
snapp@cs.uvm.edu

Department of Computer Science
University of Vermont

The Peceptron Algorithm

Consider a linear threshold unit (LTU) with n inputs $x_1, \ldots, x_n \in \mathbb{R}$ and $n+1$ weights $w_0, w_1, \ldots, w_n$.



Letting $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$ and $\mathbf{w} = (w_1, w_2, \ldots, w_n)^T \in \mathbb{R}^n$, we have

$$y = \operatorname{sgn}(\mathbf{w}^T\mathbf{x} + w_0) = \begin{cases} 1, & \text{if } \mathbf{w}^T\mathbf{x} + w_0 > 0, \\ -1, & \text{if } \mathbf{w}^T\mathbf{x} + w_0 \leq 0. \end{cases}$$

Perceptron Algorithm (cont.)

Let $\mathcal{X}_m$ denote a *dichotomy* of m patterns, i.e., a set of m patterns (or *feature vectors*) such that each pattern is assigned to exactly one of two classes),

$$\mathcal{X}_m = \{(\mathbf{x}_1, \ell_1), \ldots (\mathbf{x}_m, \ell_m)\},$$

where for $i = 1, \ldots, m$,

- $\mathbf{x}_i \in \mathbb{R}^n$ represents the i-th *feature vector*, and
- $\ell_i \in \{-1, +1\}$ the corresponding *class label*.

Perceptron Algorithm (cont.)

Assume that $\mathcal{X}_m$ represents a *linearly separable dichotomy*, that is there exists a weight vector $\mathbf{w} \in \mathbb{R}^n$ and $w_0 \in \mathbb{R}$ such that, for $i = 1, \ldots, m$,

$$\operatorname{sgn}(\mathbf{w}^T \mathbf{x}_i + w_0) = \ell_i.$$

If this is the case, how can we find a $\mathbf{w}$ and w_0 that satisfies the above?

Perceptron Algorithm (cont.)

The Perceptron algorithm (Rosenblatt):

```
float w[n] = <n random floats>;
float w_0 = <a random float>;
bool errorDetected = true;
while(errorDetected) {
  errorDetected = false;
  for(int i = 1; i ≤ m; i++) {
     if (sgn(w^T x_i + w_0) ≠ ℓ_i) {
        errorDetected = true;
        w  += ℓ_i x_i;
        w_0 += ℓ_i;
     }
  }
}
return {w, w_0};
```

Perceptron Algorithm (cont.)

The fixed-increment Perceptron algorithm (with arbitrary increment $\eta > 0$);

```
float w[n] = <n random floats>;
float w_0 = <a random float>;
bool errorDetected = true;
float η = <positive_number>;
while(errorDetected) {
  errorDetected = false;
  for(int i = 1; i ≤ m; i++) {
     if (sgn(w^T x_i + w_0) ≠ ℓ_i) {
        errorDetected = true;
        w  += ηℓ_i x_i;
        w_0 += ηℓ_i;
     }
  }
}
return {w, w_0};
```

Perceptron Convergence Theorem

If $\mathcal{X}_m = \{(\mathbf{x}_1, \ell_1), \ldots, (\mathbf{x}_m, \ell_m)\}$ describes a linearly separable dichotomy, then the fixed-increment perceptron algorithm terminates after a finite number of weight updates.

A geometric picture of the perceptron algorithm emerges after transforming into *augmented* (or *homogeneous coordinates*). Let

$$\widehat{\mathbf{x}} = \begin{pmatrix} 1 \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^{n+1}; \qquad \text{and let,} \qquad \widehat{\mathbf{w}} = \begin{pmatrix} w_0 \\ \mathbf{w} \end{pmatrix} = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{pmatrix} \in \mathbb{R}^{n+1}.$$

In the following, *hatted* vectors will always represent augmented vectors. Similarly, let

$$\widehat{\mathcal{X}}_m = \left\{ (\widehat{\mathbf{x}}_i, \ell_i) = \left((1, \mathbf{x}_i^T)^T, \ell_i \right) \middle| i = 1, 2, \ldots, m \right\}$$

Perceptron Convergence Theorem (cont.)

Consequently, the expression used by the LTU, simplifies to just an inner product, as

$$\mathbf{w}^T\mathbf{x} + w_0 = \widehat{\mathbf{w}}^T\widehat{\mathbf{x}} = \|\widehat{\mathbf{w}}\|\,\|\widehat{\mathbf{x}}\| \cos\theta,$$

where θ represents the angle between $\widehat{\mathbf{w}}$ and $\widehat{\mathbf{x}}$, measured in their common plane. Note that the cosine of θ determines the sign of the inner product.

Perceptron Algorithm (homogeneous coordinates)

The fixed-increment Perceptron algorithm (with arbitrary increment $\eta > 0$);

```
float ŵ[n+1] = <(n+1) random floats>;
bool errorDetected = true;
float η = <positive_number>;
while(errorDetected) {
  errorDetected = false;
  for(int i = 1; i ≤ m; i++) {
     if (sgn(ŵ^T x̂_i) ≠ ℓ_i) {
        errorDetected = true;
        ŵ  += η ℓ_i x̂_i;
}}}
return ŵ;
```

Normalized Coordinates

For $i = 1, 2, \ldots, m$, let

$$\widehat{\mathbf{x}}'_i = \ell_i \widehat{\mathbf{x}}_i.$$

Then, the fixed-increment perceptron algorithm becomes

```
float ŵ[n+1] = <(n+1) random floats>;
bool errorDetected = true;
float η = <positive_number>;
while(errorDetected) {
  errorDetected = false;
  for(int i = 1; i ≤ m; i++) {
     if (ŵ^T x̂'_i < 0) {
        errorDetected = true;
        ŵ  += η x̂'_i;
}}}
return ŵ;
```

Perceptron Convergence Theorem: A Proof

Given a linearly separable dichotomy (in normalized, homogeneous form),

$$\widehat{\mathcal{X}}'_m \stackrel{\text{def}}{=} \left\{ \widehat{\mathbf{x}}'_i = \ell_i \widehat{\mathbf{x}}_i \middle| (\widehat{\mathbf{x}}_i, \ell_i) \in \widehat{\mathcal{X}}_m \right\}$$

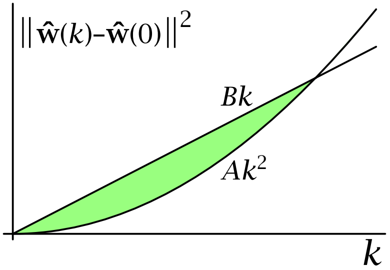
let $\widehat{\mathbf{w}}^\star \in \mathbb{R}^{n+1}$ denote a homogeneous weight vector that satisfies the given dichotomy, *i.e.*, $\widehat{\mathbf{w}}^{\star T} \widehat{\mathbf{x}}'_i > 0$, for $i = 1, 2, \ldots, m$.

Let $\widehat{\mathbf{w}}(k) \in \mathbb{R}^{n+1}$ denote the value of the perceptron's homogeneous weight vector after the k-th update.

Let $\widehat{\mathbf{x}}'(k) \in \widehat{\mathcal{X}}'_m$ denote the normalized, homogeneous feature vector that triggered the k-th update. Thus,

$$\begin{aligned}
\widehat{\mathbf{w}}(1) &= \widehat{\mathbf{w}}(0) \qquad + \eta \widehat{\mathbf{x}}'(1) \\
\widehat{\mathbf{w}}(2) &= \widehat{\mathbf{w}}(1) \qquad + \eta \widehat{\mathbf{x}}'(2) \\
&\vdots \\
\widehat{\mathbf{w}}(k) &= \widehat{\mathbf{w}}(k-1) + \eta \widehat{\mathbf{x}}'(k)
\end{aligned}$$

Perceptron Convergence Theorem: A Proof (cont.)



For a given dichotomy $\widehat{\mathcal{X}}'_m$, and parameter η, we will show that there exists constants *A* and *B* such that

$$Ak^2 \leq \left\|\widehat{\mathbf{w}}(k) - \widehat{\mathbf{w}}(0)\right\|^2 \leq Bk.$$

Thus the network must converge after no more than $k_{\max} = B/A$ updates.

Peceptron Convergence Theorem: A Proof (cont.)

Lemma (Cauchy-Schwartz Inequality): Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, then

$$\|\mathbf{a}\|^2 \|\mathbf{b}\|^2 \geq |\mathbf{a}^T \mathbf{b}|^2,$$

with equality, if and only if there exists a $c \in \mathbb{R}$, such that $c\mathbf{a} = \mathbf{b}$.

Proof:

Let,

$$\phi(t) \stackrel{\text{def}}{=} \|t\mathbf{a} + \mathbf{b}\|^2 = \|\mathbf{a}\|^2 t^2 + 2\mathbf{a}^T \mathbf{b} t + \|\mathbf{b}\|^2.$$

Note that $\phi(t)$ is a non-negative quadratic, of the form $\phi(t) = At^2 + Bt + C \geq 0$. Thus, the discriminant satisfies,

$$B^2 - 4AC \leq 0,$$

with equality if and only if there exists a value of $t \in \mathbb{R}$ for which $\phi(t) = 0$. (Letting $c = -t$ implies that $c\mathbf{a} = \mathbf{b}$.) More generally,

$$\begin{aligned} 4AC \geq B^2 \Rightarrow 4\|\mathbf{a}\|^2 \|\mathbf{b}\|^2 &\geq 4|\mathbf{a}^T \mathbf{b}|^2 \\ \Rightarrow \quad \|\mathbf{a}\|^2 \|\mathbf{b}\|^2 &\geq \quad |\mathbf{a}^T \mathbf{b}|^2 \qquad \blacksquare \end{aligned}$$

Convergence Proof: Lower Bound

Given a linearly separable dichotomy (in normalized, homogeneous form),

$$\widehat{\mathcal{X}}'_m \stackrel{\text{def}}{=} \left\{ \widehat{\mathbf{x}}'_i = \ell_i \widehat{\mathbf{x}}_i \middle| (\widehat{\mathbf{x}}_i, \ell_i) \in \widehat{\mathcal{X}}_m \right\}$$

let $\widehat{\mathbf{w}}^\star \in \mathbb{R}^{n+1}$ denote a homogeneous weight vector that satisfies the given dichotomy, *i.e.*, $\widehat{\mathbf{w}}^{\star T} \widehat{\mathbf{x}}'_i > 0$, for $i = 1, 2, \ldots, m$.

Let $\widehat{\mathbf{w}}(k) \in \mathbb{R}^{n+1}$ denote the value of the perceptron's homogeneous weight vector after the k-th update.

Let $\widehat{\mathbf{x}}'(k) \in \widehat{\mathcal{X}}'_m$ denote the normalized, homogeneous feature vector that triggered the k-th update. Thus,

$$\begin{aligned}
\widehat{\mathbf{w}}(1) &= \widehat{\mathbf{w}}(0) &&+ \eta \widehat{\mathbf{x}}'(1) \\
\widehat{\mathbf{w}}(2) &= \widehat{\mathbf{w}}(1) &&+ \eta \widehat{\mathbf{x}}'(2) \\
&\ \vdots \\
\widehat{\mathbf{w}}(k) &= \widehat{\mathbf{w}}(k-1) &&+ \eta \widehat{\mathbf{x}}'(k)
\end{aligned}$$

Convergence Proof: Lower Bound (cont.)

$$
\begin{aligned}
\widehat{\mathbf{w}}(1) &= \widehat{\mathbf{w}}(0) &&+ \eta\widehat{\mathbf{x}}'(1) \\
\widehat{\mathbf{w}}(2) &= \widehat{\mathbf{w}}(1) &&+ \eta\widehat{\mathbf{x}}'(2) \\
&\vdots \\
\widehat{\mathbf{w}}(k) &= \widehat{\mathbf{w}}(k-1) &&+ \eta\widehat{\mathbf{x}}'(k)
\end{aligned}
$$

Adding the above k equations, yields,

$$\widehat{\mathbf{w}}(k) = \widehat{\mathbf{w}}(0) + \eta\left(\widehat{\mathbf{x}}'(1) + \widehat{\mathbf{x}}'(2) + \cdots + \widehat{\mathbf{x}}'(k)\right).$$

Subtracting $\widehat{\mathbf{w}}(0)$ from both sides, and taking the inner-product with respect to the hypothetical solution $\widehat{\mathbf{w}}^\star$ yields,

$$\widehat{\mathbf{w}}^{\star T}(\widehat{\mathbf{w}}(k) - \widehat{\mathbf{w}}(0)) = \eta\widehat{\mathbf{w}}^{\star T}\left(\widehat{\mathbf{x}}'(1) + \widehat{\mathbf{x}}'(2) + \cdots + \widehat{\mathbf{x}}'(k)\right).$$

Convergence Proof: Lower Bound (cont.)

$$\widehat{\mathbf{w}}^{\star T}(\widehat{\mathbf{w}}(k) - \widehat{\mathbf{w}}(0)) = \eta \widehat{\mathbf{w}}^{\star T}\left(\widehat{\mathbf{x}}'(1) + \widehat{\mathbf{x}}'(2) + \cdots + \widehat{\mathbf{x}}'(k)\right).$$

Now define,

$$a = \min_{\widehat{\mathbf{x}}' \in \widehat{\mathcal{X}}'_m} \widehat{\mathbf{w}}^{\star T}\widehat{\mathbf{x}}' > 0.$$

Thus,

$$\widehat{\mathbf{w}}^{\star T}(\widehat{\mathbf{w}}(k) - \widehat{\mathbf{w}}(0) \geq \eta a k > 0.$$

Squaring both sides, with the Cauchy-Schwartz inequality, yields

$$\|\widehat{\mathbf{w}}^{\star}\|^2 \|\widehat{\mathbf{w}}(k) - \widehat{\mathbf{w}}(0)\|^2 \geq |\widehat{\mathbf{w}}^{\star T}(\widehat{\mathbf{w}}(k) - \widehat{\mathbf{w}}(0))|^2 \geq (\eta a k)^2.$$

Thus,

$$\|\widehat{\mathbf{w}}(k) - \widehat{\mathbf{w}}(0)\|^2 \geq \left(\frac{\eta a}{\|\widehat{\mathbf{w}}^{\star}\|}\right)^2 k^2.$$

Convergence Proof: Upper Bound

To construct an upper bound of the growth of $\|\widehat{\mathbf{w}}(k) - \widehat{\mathbf{w}}(0)\|^2$, we begin with the sequence of weight values generated by the Perceptron algorithm:

$$
\begin{aligned}
\widehat{\mathbf{w}}(1) &= \widehat{\mathbf{w}}(0) \qquad + \eta\widehat{\mathbf{x}}'(1) \\
\widehat{\mathbf{w}}(2) &= \widehat{\mathbf{w}}(1) \qquad + \eta\widehat{\mathbf{x}}'(2) \\
&\vdots \\
\widehat{\mathbf{w}}(k) &= \widehat{\mathbf{w}}(k-1) + \eta\widehat{\mathbf{x}}'(k)
\end{aligned}
$$

Now subtract $\widehat{\mathbf{w}}(0)$ from both sides,

$$
\begin{aligned}
\widehat{\mathbf{w}}(1) - \widehat{\mathbf{w}}(0) &= \eta\widehat{\mathbf{x}}'(1) \\
\widehat{\mathbf{w}}(2) - \widehat{\mathbf{w}}(0) &= \big(\widehat{\mathbf{w}}(1) - \widehat{\mathbf{w}}(0)\big) + \eta\widehat{\mathbf{x}}'(2) \\
&\vdots \\
\widehat{\mathbf{w}}(k) - \widehat{\mathbf{w}}(0) &= \big(\widehat{\mathbf{w}}(k-1) - \widehat{\mathbf{w}}(0)\big) + \eta\widehat{\mathbf{x}}'(k)
\end{aligned}
$$

Convergence Proof: Upper Bound (cont.)

Squaring both sides yields,

$$
\begin{aligned}
\left\|\widehat{\mathbf{w}}(1)-\widehat{\mathbf{w}}(0)\right\|^2 &= \eta^2\left\|\widehat{\mathbf{x}}'(1)\right\|^2 \\
\left\|\widehat{\mathbf{w}}(2)-\widehat{\mathbf{w}}(0)\right\|^2 &= \left\|\widehat{\mathbf{w}}(1)-\widehat{\mathbf{w}}(0)\right\|^2+2\eta\left(\widehat{\mathbf{w}}(1)-\widehat{\mathbf{w}}(0)\right)^T\widehat{\mathbf{x}}'(2)+\eta^2\left\|\widehat{\mathbf{x}}'(2)\right\|^2 \\
&\vdots \\
\left\|\widehat{\mathbf{w}}(k)-\widehat{\mathbf{w}}(0)\right\|^2 &= \left\|\widehat{\mathbf{w}}(k-1)-\widehat{\mathbf{w}}(0)\right\|^2+2\eta\left(\widehat{\mathbf{w}}(k-1)-\widehat{\mathbf{w}}(0)\right)^T\widehat{\mathbf{x}}'(k)+\eta^2\left\|\widehat{\mathbf{x}}'(k)\right\|^2
\end{aligned}
$$

Note that since $\widehat{\mathbf{x}}'(1)$ triggers the first weight update, it must have been misclassified by the weight vector $\widehat{\mathbf{w}}(0)$. Thus $\widehat{\mathbf{w}}(0)^T\widehat{\mathbf{x}}'(1) < 0$.
Similarly,

$$\widehat{\mathbf{w}}(j-1)^T\widehat{\mathbf{x}}'(j) < 0 \quad \text{for} \quad j = 1, 2, \ldots k.$$

Convergence Proof: Upper Bound (cont.)

Thus,

$$\begin{aligned}
\left\|\widehat{\mathbf{w}}(1)-\widehat{\mathbf{w}}(0)\right\|^2 &= \eta^2\left\|\widehat{\mathbf{x}}'(1)\right\|^2 \\
\left\|\widehat{\mathbf{w}}(2)-\widehat{\mathbf{w}}(0)\right\|^2 &\leq \left\|\widehat{\mathbf{w}}(1)-\widehat{\mathbf{w}}(0)\right\|^2 - 2\eta\widehat{\mathbf{w}}(0)^T\widehat{\mathbf{x}}'(2)+\eta^2\left\|\widehat{\mathbf{x}}'(2)\right\|^2 \\
&\vdots \\
\left\|\widehat{\mathbf{w}}(k)-\widehat{\mathbf{w}}(0)\right\|^2 &\leq \left\|\widehat{\mathbf{w}}(k-1)-\widehat{\mathbf{w}}(0)\right\|^2 - 2\eta\widehat{\mathbf{w}}(0)^T\widehat{\mathbf{x}}'(k)+\eta^2\left\|\widehat{\mathbf{x}}'(k)\right\|^2
\end{aligned}$$

Summing the k inequalities above, yields

$$\begin{aligned}
\left\|\widehat{\mathbf{w}}(k)-\widehat{\mathbf{w}}(0)\right\|^2 \leq \eta^2&\left(\left\|\widehat{\mathbf{x}}'(1)\right\|^2+\left\|\widehat{\mathbf{x}}'(2)\right\|^2+\cdots+\left\|\widehat{\mathbf{x}}'(k)\right\|^2\right) \\
&-2\eta\widehat{\mathbf{w}}(0)^T\left(\widehat{\mathbf{x}}'(2)+\cdots+\widehat{\mathbf{x}}'(k)\right)
\end{aligned}$$

Convergence Proof: Upper Bound (cont.)

Now define

$$M = \max_{\widehat{\mathbf{x}}' \in \widehat{\mathcal{X}}'_m} \left\| \widehat{\mathbf{x}}' \right\|^2,$$

and

$$\mu = 2 \min_{\widehat{\mathbf{x}}' \in \widehat{\mathcal{X}}'_m} \widehat{\mathbf{w}}(0)^T \widehat{\mathbf{x}}'.$$

(Note that $\mu < 0$, unless $\widehat{\mathbf{w}}(0)$ solves the dichotomy.)
Whence,

$$\begin{aligned} \left\| \widehat{\mathbf{w}}(k) - \widehat{\mathbf{w}}(0) \right\|^2 \leq \eta^2 \left(\left\| \widehat{\mathbf{x}}'(1) \right\|^2 + \left\| \widehat{\mathbf{x}}'(2) \right\|^2 + \cdots + \left\| \widehat{\mathbf{x}}'(k) \right\|^2 \right) \\ - 2\eta \widehat{\mathbf{w}}(0)^T \left(\widehat{\mathbf{x}}'(2) + \cdots + \widehat{\mathbf{x}}'(k) \right) \end{aligned}$$

becomes

$$\left\| \widehat{\mathbf{w}}(k) - \widehat{\mathbf{w}}(0) \right\|^2 \leq (\eta^2 M - \eta\mu)k.$$

Convergence Proof: Summary

Thus we have shown

$$Ak^2 \leq \left\| \widehat{\mathbf{w}}(k) - \widehat{\mathbf{w}}(0) \right\|^2 \leq Bk.$$

with

$$A = \left(\frac{\eta a}{\|\widehat{\mathbf{w}}^\star\|} \right)^2, \quad \text{and,} \quad B = \eta(\eta M - \mu).$$

Thus,

$$k_{\max} = \frac{\eta M - \mu}{\eta a^2} \|\widehat{\mathbf{w}}^\star\|^2.$$