

Convergence of Perceptrons

Nonlinearly Separable Case

Using a slightly different activation function, the perceptron learning rule converges to a modified L_1 loss.

We can show that the distance between \mathbf{w} and the optimal \mathbf{w}^* decreases when \mathbf{w} makes a mistake.

Let ramp be the activation function:

$$\text{ramp}(u) = \begin{cases} -1 & \text{if } u \leq -1 \\ u & \text{if } -1 < u < 1 \\ 1 & \text{if } u \geq 1 \end{cases}$$

Let the learning rule be:

$$\mathbf{if } d \neq \text{ramp}(u) \mathbf{ then } \mathbf{w} \leftarrow \mathbf{w} + d\eta\mathbf{x}$$

Let the error function be:

$$E(u, d) = \begin{cases} 0 & \text{if } d = \text{ramp}(u) \\ |d - u| & \text{otherwise} \end{cases}$$

Let $u^* = \mathbf{w}^* \cdot \mathbf{x}$.

Distance to Optimal Weights

Measure the distance between \mathbf{w} and \mathbf{w}^* by:

$$\|\mathbf{w} - \mathbf{w}^*\|^2 = \sum_i (w_i - w_i^*)^2$$

When $\text{ramp}(u) \neq d$, then \mathbf{w} is updated to $\mathbf{w}' = \mathbf{w} + d\eta\mathbf{x}$.

We want to know when \mathbf{w}' is closer to \mathbf{w}^* , so we analyze

$$\|\mathbf{w} - \mathbf{w}^*\|^2 - \|\mathbf{w}' - \mathbf{w}^*\|^2$$

and determine when it is positive.

$$\begin{aligned} & \|\mathbf{w} - \mathbf{w}^*\|^2 - \|\mathbf{w}' - \mathbf{w}^*\|^2 \\ &= \sum_i \left((w_i - w_i^*)^2 - (w'_i - w_i^*)^2 \right) \\ &= \sum_i \left((w_i - w_i^*)^2 - (w_i + d\eta x_i - w_i^*)^2 \right) \\ &= \sum_i \left(2d\eta x_i (w_i^* - w_i) - (d\eta x_i)^2 \right) \\ &= 2d\eta \sum_i w_i^* x_i - 2d\eta \sum_i w_i x_i - d^2 \eta^2 \sum_i x_i^2 \\ &= 2d\eta u^* - 2d\eta u - \eta^2 \|\mathbf{x}\|^2 \\ &\geq 2\eta (E(d, u) - E(d, u^*)) - \eta^2 \|\mathbf{x}\|^2 \end{aligned}$$

The last line in the derivation is justified by:

If $\text{ramp}(u) \neq d$, then, because $d \in \{-1, 1\}$,

$$E(d, u) = |d - u| = d(d - u)$$

If $\text{ramp}(u^*) \neq d$, then $E(d, u^*) = d(d - u^*)$.

If $\text{ramp}(u^*) = d$, then

$$E(d, u^*) = 0 \geq d(d - u^*)$$

So, $E(d, u) - E(d, u^*) \leq d(d - u) - d(d - u^*)$,
which simplifies to $d(u^* - u)$.

Interpretation of Result

$$\begin{aligned} & \|\mathbf{w} - \mathbf{w}^*\|^2 - \|\mathbf{w}' - \mathbf{w}^*\|^2 \\ & \geq 2\eta \left(E(d, u) - \left(E(d, u^*) + \eta \|\mathbf{x}\|^2 / 2 \right) \right) \end{aligned}$$

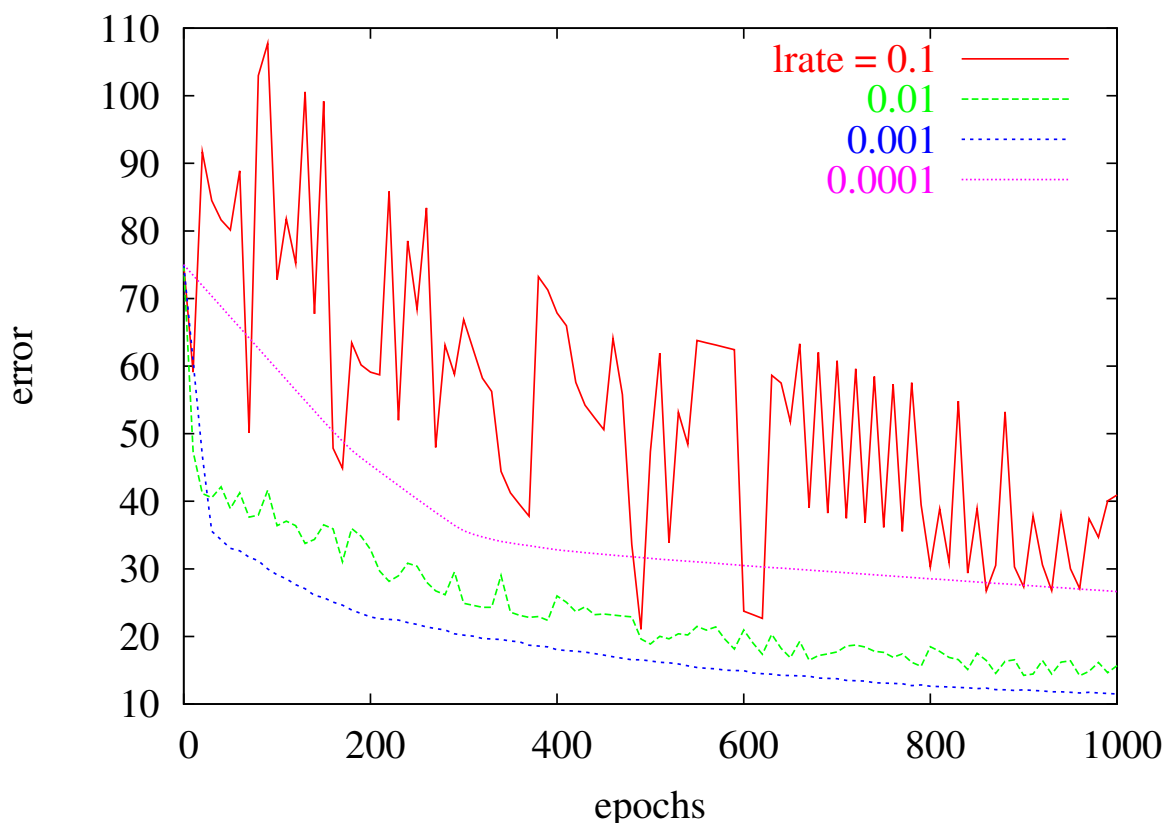
This shows that the distance from \mathbf{w} to \mathbf{w}^* decreases if \mathbf{w} 's error is higher than \mathbf{w}^* 's error (plus an extra term).

We need to choose η to make $\eta \|\mathbf{x}\|^2$ small. η should be no more than $1/X^2$ where X is the maximum length of an input vector.

Let W be the length of \mathbf{w}^* . Over many examples, the algorithm's error will be comparable to the optimal error plus $W^2 X^2$ plus a small error per example.

Below, I plot how the error decreases for different learning rates on iris2d.data using only the third output. The largest input vector is $(6.9, 2.3)$. Including the extra 1 for the bias, then $X \approx 7.28$ and $1/X^2 \approx 0.0189$.

$\eta = 0.001$ is the best over 1000 epochs.



In a second experiment, I modified each (x_1, x_2) in iris2d.data to:

$$\left(x_1/10, x_2/5, (x_1/10)^2, (x_2/5)^2, x_1x_2/50\right)$$

That is, I performed scaling of the inputs and added second-order terms. In this case, $X \approx 1.44$ and $1/X^2 \approx 0.485$.

Below, is the result of learning the second class. The best learning rate is 0.1.

