Deep Learning - Theory and Practice

IE 643
Lecture 4

August 16, 2022.

Recap: Training a Perceptron

Perceptron - Training

Perceptron Training Procedure

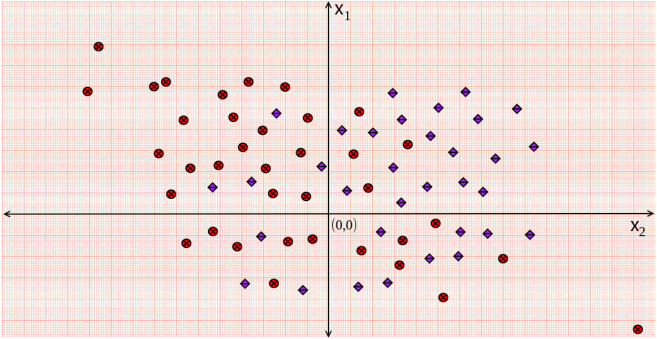
$w^1 = 0$
for $t \leftarrow 1, 2, 3, \ldots$ **do**
  receive (x^t, y^t), $x^t \in \mathbb{R}^d$, $y^t \in \{+1, -1\}$.
  $\widehat{y} = Perceptron(x^t; w^t)$
  if $\widehat{y} \neq y^t$ **then**
    $w^{t+1} = w^t + y^t x^t$
  else
    $w^{t+1} = w^t$

Convergence of Perceptron Training

Perceptron Convergence - Geometric Intuition

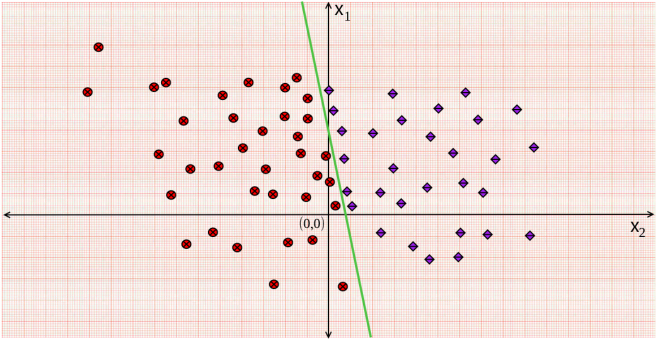
- What are some natural assumptions to expect the perceptron training to converge?
- Let us first motivate such assumptions through geometric intuition.

Perceptron Convergence - Geometric Intuition



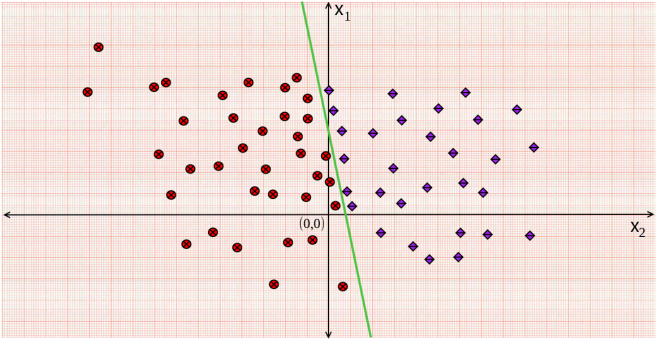
- Can the data be separated by a hyperplane?

Perceptron Convergence - Geometric Intuition



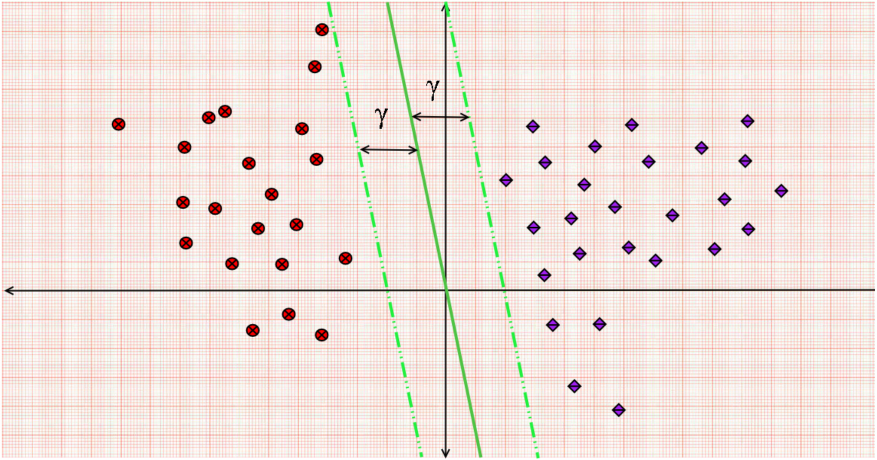
- **First assumption:** At least the data should be such that the samples with label 1 can to be separated by a hyperplane from samples with label -1.

Perceptron Convergence - Geometric Intuition

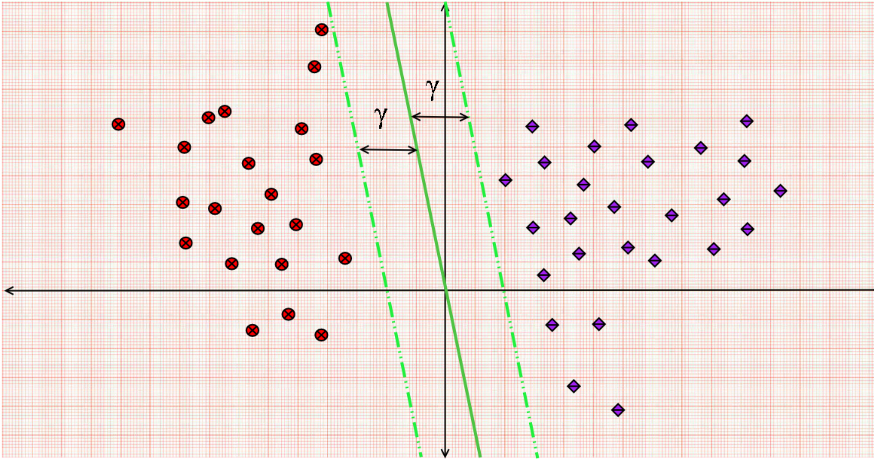


- **First assumption:** At least the data should be such that the samples with label 1 can to be separated by a hyperplane from samples with label -1.
- Is this assumption sufficient?

Perceptron Convergence - Geometric Intuition



Perceptron Convergence - Geometric Intuition



- **Refined assumption:** We not only want the data to be separated but the separation should be good enough!

Perceptron Convergence - Separability Assumption

Linear Separability Assumption

Let $D = \{(x^t, y^t)\}_{t=1}^{\infty}$ denote the training data where $x^t \in \mathbb{R}^d$, $y^t \in \{+1, -1\}$, $\forall t = 1, 2, \ldots$. Then there exist $\mathbb{R}^d \ni w^* \neq 0,\ \gamma > 0$, such that:

$$\langle w^*, x^t \rangle > \gamma \text{ where } y^t = 1,$$
$$\langle w^*, x^t \rangle < -\gamma \text{ where } y^t = -1.$$

Perceptron Convergence - Separability Assumption

Linear Separability Assumption

Let $D = \{(x^t, y^t)\}_{t=1}^{\infty}$ denote the training data where $x^t \in \mathbb{R}^d$, $y^t \in \{+1, -1\}$, $\forall t = 1, 2, \ldots$. Then there exist $\mathbb{R}^d \ni w^* \neq 0,\ \gamma > 0$, such that:

$$y^t \langle w^*, x^t \rangle > \gamma.$$

Perceptron Convergence - Mistake Bound

- We will try to derive useful bounds on the number of mistakes that a perceptron can commit during its training.
- **Assumption on data:** Linear Separability
- Assume that the T rounds of training have been completed in perceptron training. Assume T to be some large number.
- Assume that M mistakes are made by the perceptron in these T rounds. (Obviously, $M \leq T$.)
- We ask if the number of mistakes M can be bounded by some suitable quantity.

Perceptron Convergence - Mistake Bound

- We begin the analysis by considering an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.

Perceptron Convergence - Mistake Bound

- We begin the analysis by considering an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.

Perceptron Convergence - Mistake Bound

- We begin the analysis by considering an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.

Perceptron Convergence - Mistake Bound

- We begin the analysis by considering an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.
  - Perceptron update: $w^{t+1} = w^t + y^t x^t$.

Perceptron Convergence - Mistake Bound

- We begin the analysis by considering an arbitrary round $t \in \{1, 2, \dots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.
  - Perceptron update: $w^{t+1} = w^t + y^t x^t$.
- Now from linear separability assumption, we have $w^* \neq 0$ such that $y^t \langle w^*, x^t \rangle > \gamma$.

Perceptron Convergence - Mistake Bound

- We begin the analysis by considering an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.
  - Perceptron update: $w^{t+1} = w^t + y^t x^t$.
- Now from linear separability assumption, we have $w^* \neq 0$ such that $y^t \langle w^*, x^t \rangle > \gamma$.
- **First step:** To bound the difference $\langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle$.

Perceptron Convergence - Mistake Bound

- We begin the analysis by considering an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.
  - Perceptron update: $w^{t+1} = w^t + y^t x^t$.
- Now from linear separability assumption, we have $w^* \neq 0$ such that $y^t \langle w^*, x^t \rangle > \gamma$.
- Now we can write

$$\langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle = \langle w^*, w^t + y^t x^t \rangle - \langle w^*, w^t \rangle$$

Perceptron Convergence - Mistake Bound

- We begin the analysis by considering an arbitrary round $t \in \{1, 2, \dots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.
  - Perceptron update: $w^{t+1} = w^t + y^t x^t$.
- Now from linear separability assumption, we have $w^* \neq 0$ such that $y^t \langle w^*, x^t \rangle > \gamma$.
- Now we can write

$$\begin{aligned}
\langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle &= \langle w^*, w^t + y^t x^t \rangle - \langle w^*, w^t \rangle \\
&= \langle w^*, w^t \rangle + \langle w^*, y^t x^t \rangle - \langle w^*, w^t \rangle \text{ (how?)}
\end{aligned}$$

Perceptron Convergence - Mistake Bound

- We begin the analysis by considering an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.
  - Perceptron update: $w^{t+1} = w^t + y^t x^t$.
- Now from linear separability assumption, we have $w^* \neq 0$ such that $y^t \langle w^*, x^t \rangle > \gamma$.
- Now we can write

$$\begin{aligned}
\langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle &= \langle w^*, w^t + y^t x^t \rangle - \langle w^*, w^t \rangle \\
&= \langle w^*, w^t \rangle + \langle w^*, y^t x^t \rangle - \langle w^*, w^t \rangle \\
&= \langle w^*, y^t x^t \rangle
\end{aligned}$$

Perceptron Convergence - Mistake Bound

- We begin the analysis by considering an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.
  - Perceptron update: $w^{t+1} = w^t + y^t x^t$.
- Now from linear separability assumption, we have $w^* \neq 0$ such that $y^t \langle w^*, x^t \rangle > \gamma$.
- Now we can write

$$
\begin{aligned}
\langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle &= \langle w^*, w^t + y^t x^t \rangle - \langle w^*, w^t \rangle \\
&= \langle w^*, w^t \rangle + \langle w^*, y^t x^t \rangle - \langle w^*, w^t \rangle \\
&= \langle w^*, y^t x^t \rangle \\
&= y^t \langle w^*, x^t \rangle \ \text{(how?)}
\end{aligned}
$$

Perceptron Convergence - Mistake Bound

- We begin the analysis by considering an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.
  - Perceptron update: $w^{t+1} = w^t + y^t x^t$.
- Now from linear separability assumption, we have $w^* \neq 0$ such that $y^t \langle w^*, x^t \rangle > \gamma$.
- Now we can write

$$\begin{aligned}
\langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle &= \langle w^*, w^t + y^t x^t \rangle - \langle w^*, w^t \rangle \\
&= \langle w^*, w^t \rangle + \langle w^*, y^t x^t \rangle - \langle w^*, w^t \rangle \\
&= \langle w^*, y^t x^t \rangle \\
&= y^t \langle w^*, x^t \rangle > \gamma
\end{aligned}$$

Perceptron Convergence - Mistake Bound

- We begin the analysis by considering an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.
  - Perceptron update: $w^{t+1} = w^t + y^t x^t$.
- Now from linear separability assumption, we have $w^* \neq 0$ such that $y^t \langle w^*, x^t \rangle > \gamma$.
- Now we can write

$$\langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle > \gamma.$$

Perceptron Convergence - Mistake Bound

- Now when no mistake is made in round t, we have $w^{t+1} = w^t$.
- Hence $\langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle = 0$.

Perceptron Convergence - Mistake Bound

Recall our assumptions:

- Assume that the T rounds of training have been completed in perceptron training. Assume T to be some large number.
- Assume that M mistakes are made by the perceptron in these T rounds. (Obviously, $M \leq T$.)

Perceptron Convergence - Mistake Bound

$$\sum_{t=1}^{T} \langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle = \sum_{\substack{t \in \{1,\dots,T\}, \\ t:\text{mistake is made} \\ \text{at round } t}} \langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle + \sum_{\substack{t \in \{1,\dots,T\}, \\ t:\text{no mistake is made} \\ \text{at round } t}} \langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle$$

Perceptron Convergence - Mistake Bound

$$\sum_{t=1}^{T} \langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle = \sum_{\substack{t \in \{1,\dots,T\}, \\ t:\text{mistake is made} \\ \text{at round } t}} \langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle +$$

$$\sum_{\substack{t \in \{1,\dots,T\}, \\ t:\text{no mistake is made} \\ \text{at round } t}} \langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle$$

$$= \sum_{\substack{t \in \{1,\dots,T\}, \\ t:\text{mistake is made} \\ \text{at round } t}} \langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle + 0 \text{ (how?)}$$

Perceptron Convergence - Mistake Bound

$$\sum_{t=1}^{T} \langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle = \sum_{\substack{t \in \{1,\ldots,T\},\\ t:\text{mistake is made}\\ \text{at round } t}} \langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle$$

$$> M\gamma \text{ (how?)}$$

Perceptron Convergence - Mistake Bound

Also note:

$$\sum_{t=1}^{T} \langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle = \langle w^*, w^{T+1} \rangle \text{ (homework!)}$$

Perceptron Convergence - Mistake Bound

Hence we have:

$$
\begin{aligned}
\sum_{t=1}^{T} \langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle &> M\gamma \\
\implies \langle w^*, w^{T+1} \rangle &> M\gamma
\end{aligned}
$$

Perceptron Mistake Bound - An upper bound

Perceptron Convergence - Mistake Bound

Now we will handle the inner product term:

$$\langle w^*, w^{T+1} \rangle > M\gamma$$

Perceptron Convergence - Mistake Bound

Now we will handle the inner product term:

$$\langle w^*, w^{T+1} \rangle > M\gamma$$

- From Cauchy-Schwarz inequality we have, $\langle w^*, w^{T+1} \rangle \leq \|w^*\|_2 \|w^{T+1}\|_2$. (Homework: Prove this inequality!)
- **Note:** $\|w^{T+1}\|_2$ denotes the Euclidean ℓ_2 norm of w^{T+1}.

Perceptron Convergence - Mistake Bound

Now we will handle the inner product term:

$$\langle w^*, w^{T+1} \rangle > M\gamma$$

- From Cauchy-Schwarz inequality we have, $\langle w^*, w^{T+1} \rangle \leq \|w^*\|_2 \|w^{T+1}\|_2$. (Homework: Prove this inequality!)
- **Note:** $\|w^{T+1}\|_2$ denotes the Euclidean ℓ_2 norm of w^{T+1}.
- We will now see how to bound $\|w^{T+1}\|_2$.

Perceptron Convergence - Mistake Bound

- Again we consider an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.

Perceptron Convergence - Mistake Bound

- Again we consider an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.

Perceptron Convergence - Mistake Bound

- Again we consider an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.

Perceptron Convergence - Mistake Bound

- Again we consider an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.
  - Perceptron update: $w^{t+1} = w^t + y^t x^t$.

Perceptron Convergence - Mistake Bound

- Again we consider an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.
  - Perceptron update: $w^{t+1} = w^t + y^t x^t$.

Now, we have

$$\|w^{t+1}\|_2^2 = \|w^t + y^t x^t\|_2^2$$

Perceptron Convergence - Mistake Bound

- Again we consider an arbitrary round $t \in \{1, 2, \dots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.
  - Perceptron update: $w^{t+1} = w^t + y^t x^t$.

  Now, we have

$$\begin{aligned} \|w^{t+1}\|_2^2 &= \|w^t + y^t x^t\|_2^2 \\ &= \|w^t\|_2^2 + \|y^t x^t\|_2^2 + 2\langle w^t, y^t x^t \rangle \end{aligned}$$

Perceptron Convergence - Mistake Bound

- Again we consider an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.
  - Perceptron update: $w^{t+1} = w^t + y^t x^t$.

Now, we have

$$\begin{aligned} \|w^{t+1}\|_2^2 &= \|w^t + y^t x^t\|_2^2 \\ &= \|w^t\|_2^2 + \|y^t x^t\|_2^2 + 2\langle w^t, y^t x^t \rangle \\ &= \|w^t\|_2^2 + \|x^t\|_2^2 + 2y^t \langle w^t, x^t \rangle \end{aligned}$$

Perceptron Convergence - Mistake Bound

- Again we consider an arbitrary round $t \in \{1, 2, \dots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.
  - Perceptron update: $w^{t+1} = w^t + y^t x^t$.

Now, we have

$$
\begin{aligned}
\|w^{t+1}\|_2^2 &= \|w^t + y^t x^t\|_2^2 \\
&= \|w^t\|_2^2 + \|y^t x^t\|_2^2 + 2\langle w^t, y^t x^t \rangle \\
&= \|w^t\|_2^2 + \|x^t\|_2^2 + 2y^t \langle w^t, x^t \rangle \\
\implies \|w^{t+1}\|_2^2 &\leq \|w^t\|_2^2 + \|x^t\|_2^2 \ \text{(How?)}
\end{aligned}
$$

Perceptron Convergence - Mistake Bound

- Again we consider an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.
  - Perceptron update: $w^{t+1} = w^t + y^t x^t$.

Now, we have

$$\begin{aligned} \|w^{t+1}\|_2^2 &= \|w^t + y^t x^t\|_2^2 \\ &= \|w^t\|_2^2 + \|y^t x^t\|_2^2 + 2\langle w^t, y^t x^t \rangle \\ &= \|w^t\|_2^2 + \|x^t\|_2^2 + 2y^t \langle w^t, x^t \rangle \\ \implies \|w^{t+1}\|_2^2 &\leq \|w^t\|_2^2 + \|x^t\|_2^2 \text{ (How?)} \end{aligned}$$

Thus $\|w^{t+1}\|_2^2 - \|w^t\|_2^2 \leq \|x^t\|_2^2$.

Perceptron Convergence - Mistake Bound

Assumption on boundedness of $\|x^t\|_2$

We shall assume further that $\forall t = 1, 2, \ldots,$ the ℓ_2 norm (or length) of x^t is bounded, which is denoted as:

$$\|x^t\|_2 \leq R \quad \forall t = 1, 2, \ldots$$

Perceptron Convergence - Mistake Bound

Assumption on boundedness of $\|x^t\|_2$

We shall assume further that $\forall t = 1, 2, \ldots,$ the ℓ_2 norm (or length) of x^t is bounded, which is denoted as:

$$\|x^t\|_2 \leq R \quad \forall t = 1, 2, \ldots$$

- This is yet another assumption to help our analysis.
- Bounded $\|x^t\|_2$ is not very unrealistic, however finding a suitable value for R might be difficult.

Perceptron Convergence - Mistake Bound

Assumption on boundedness of $\|x^t\|_2$

We shall assume further that $\forall t = 1, 2, \ldots,$ the ℓ_2 norm (or length) of x^t is bounded, which is denoted as:

$$\|x^t\|_2 \leq R \quad \forall t = 1, 2, \ldots$$

- This is yet another assumption to help our analysis.
- Bounded $\|x^t\|_2$ is not very unrealistic, however finding a suitable value for R might be difficult.
- This is where normalizing all x^t might help, so that $\|x^t\|_2 \leq 1$ can be assumed.
- **Note:** The set $\{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$ is called a unit ball in $\mathbb{R}^d$.

Perceptron Convergence - Mistake Bound

We thus have

$$\|w^{t+1}\|_2^2 - \|w^t\|_2^2 \le \|x^t\|_2^2 \implies \|w^{t+1}\|_2^2 - \|w^t\|_2^2 \le R^2.$$

Perceptron Convergence - Mistake Bound

We thus have

$$\|w^{t+1}\|_2^2 - \|w^t\|_2^2 \le \|x^t\|_2^2 \implies \|w^{t+1}\|_2^2 - \|w^t\|_2^2 \le R^2.$$

Again, summing $\|w^{t+1}\|_2^2 - \|w^t\|_2^2$ over $t = 1, \dots, T$ we get

Perceptron Convergence - Mistake Bound

We thus have

$$\|w^{t+1}\|_2^2 - \|w^t\|_2^2 \leq \|x^t\|_2^2 \implies \|w^{t+1}\|_2^2 - \|w^t\|_2^2 \leq R^2.$$

Again, summing $\|w^{t+1}\|_2^2 - \|w^t\|_2^2$ over $t = 1, \dots, T$ we get

$$\sum_{i=1}^{T} \|w^{t+1}\|_2^2 - \|w^t\|_2^2 = \sum_{\substack{t\in\{1,\dots,T\},\\ t:\text{mistake is made}\\ \text{at round } t}} \|w^{t+1}\|_2^2 - \|w^t\|_2^2 + \sum_{\substack{t\in\{1,\dots,T\},\\ t:\text{no mistake is made}\\ \text{at round } t}} \|w^{t+1}\|_2^2 - \|w^t\|_2^2$$

Perceptron Convergence - Mistake Bound

We thus have

$$\|w^{t+1}\|_2^2 - \|w^t\|_2^2 \leq \|x^t\|_2^2 \implies \|w^{t+1}\|_2^2 - \|w^t\|_2^2 \leq R^2.$$

Again, summing $\|w^{t+1}\|_2^2 - \|w^t\|_2^2$ over $t = 1, \ldots, T$ we get

$$\begin{aligned}\sum_{i=1}^{T} \|w^{t+1}\|_2^2 - \|w^t\|_2^2 &= \sum_{\substack{t \in \{1,\ldots,T\}, \\ t:\text{mistake is made} \\ \text{at round } t}} \|w^{t+1}\|_2^2 - \|w^t\|_2^2 + \\ &\quad \sum_{\substack{t \in \{1,\ldots,T\}, \\ t:\text{no mistake is made} \\ \text{at round } t}} \|w^{t+1}\|_2^2 - \|w^t\|_2^2 \\ &= \sum_{\substack{t \in \{1,\ldots,T\}, \\ t:\text{mistake is made} \\ \text{at round } t}} \|w^{t+1}\|_2^2 - \|w^t\|_2^2\end{aligned}$$

Perceptron Convergence - Mistake Bound

We thus have

$$\|w^{t+1}\|_2^2 - \|w^t\|_2^2 \leq \|x^t\|_2^2 \implies \|w^{t+1}\|_2^2 - \|w^t\|_2^2 \leq R^2.$$

Again, summing $\|w^{t+1}\|_2^2 - \|w^t\|_2^2$ over $t = 1, \dots, T$ we get

$$\begin{aligned}\sum_{i=1}^{T} \|w^{t+1}\|_2^2 - \|w^t\|_2^2 &= \sum_{\substack{t\in\{1,\dots,T\},\\ t:\text{mistake is made}\\ \text{at round } t}} \|w^{t+1}\|_2^2 - \|w^t\|_2^2 + \\ &\quad \sum_{\substack{t\in\{1,\dots,T\},\\ t:\text{no mistake is made}\\ \text{at round } t}} \|w^{t+1}\|_2^2 - \|w^t\|_2^2 \\ &= \sum_{\substack{t\in\{1,\dots,T\},\\ t:\text{mistake is made}\\ \text{at round } t}} \|w^{t+1}\|_2^2 - \|w^t\|_2^2 \leq MR^2 \text{ (How?)}\end{aligned}$$

Perceptron Convergence - Mistake Bound

Thus we have

$$\sum_{i=1}^{T} \|w^{t+1}\|_2^2 - \|w^t\|_2^2 \leq MR^2.$$

Perceptron Convergence - Mistake Bound

Thus we have

$$\sum_{i=1}^{T} \|w^{t+1}\|_2^2 - \|w^t\|_2^2 \leq MR^2.$$

On the other hand we get:

$$\sum_{i=1}^{T} \|w^{t+1}\|_2^2 - \|w^t\|_2^2 = \|w^{T+1}\|_2^2. \text{ (Homework!)}$$

Perceptron Convergence - Mistake Bound

Thus we have

$$\sum_{i=1}^{T} \|w^{t+1}\|_2^2 - \|w^t\|_2^2 \leq MR^2.$$

On the other hand we get:

$$\sum_{i=1}^{T} \|w^{t+1}\|_2^2 - \|w^t\|_2^2 = \|w^{T+1}\|_2^2. \text{ (Homework!)}$$

Combining both, we get

$$\|w^{T+1}\|_2^2 \leq MR^2.$$

Perceptron Convergence - Mistake Bound

Thus we have

$$\sum_{i=1}^{T} \|w^{t+1}\|_2^2 - \|w^t\|_2^2 \leq MR^2.$$

On the other hand we get:

$$\sum_{i=1}^{T} \|w^{t+1}\|_2^2 - \|w^t\|_2^2 = \|w^{T+1}\|_2^2.$$ (Homework!)

Combining both, we get

$$\|w^{T+1}\|_2^2 \leq MR^2.$$

Thus we have bounded $\|w^{T+1}\|_2$.

Perceptron Convergence - Mistake Bound

Recall: We wanted to handle the inner product term:

$$\langle w^*, w^{T+1} \rangle > M\gamma$$

Perceptron Convergence - Mistake Bound

Recall: We wanted to handle the inner product term:

$$\langle w^*, w^{T+1} \rangle > M\gamma$$

Then using Cauchy-Schwarz inequality we had

$$M\gamma < \langle w^*, w^{T+1} \rangle \leq \|w^*\|_2 \|w^{T+1}\|_2$$

Perceptron Convergence - Mistake Bound

Recall: We wanted to handle the inner product term:

$$\langle w^*, w^{T+1} \rangle > M\gamma$$

Then using Cauchy-Schwarz inequality we had

$$\begin{aligned} M\gamma < \langle w^*, w^{T+1} \rangle &\leq \|w^*\|_2 \|w^{T+1}\|_2 \\ \implies M^2\gamma^2 &< \|w^*\|_2^2 \|w^{T+1}\|_2^2 \end{aligned}$$

Perceptron Convergence - Mistake Bound

Recall: We wanted to handle the inner product term:

$$\langle w^*, w^{T+1} \rangle > M\gamma$$

Then using Cauchy-Schwarz inequality we had

$$\begin{aligned} M\gamma < \langle w^*, w^{T+1} \rangle &\leq \|w^*\|_2 \|w^{T+1}\|_2 \\ \implies M^2\gamma^2 &< \|w^*\|_2^2 \|w^{T+1}\|_2^2 \end{aligned}$$

Using the bound $\|w^{T+1}\|_2^2 \leq MR^2$ we obtain:

Perceptron Convergence - Mistake Bound

Recall: We wanted to handle the inner product term:

$$\langle w^*, w^{T+1} \rangle > M\gamma$$

Then using Cauchy-Schwarz inequality we had

$$\begin{aligned} M\gamma < \langle w^*, w^{T+1} \rangle &\leq \|w^*\|_2 \|w^{T+1}\|_2 \\ \implies M^2\gamma^2 &< \|w^*\|_2^2 \|w^{T+1}\|_2^2 \end{aligned}$$

Using the bound $\|w^{T+1}\|_2^2 \leq MR^2$ we obtain:

$$M^2\gamma^2 < \|w^*\|_2^2 MR^2$$

Perceptron Convergence - Mistake Bound

Recall: We wanted to handle the inner product term:

$$\langle w^*, w^{T+1} \rangle > M\gamma$$

Then using Cauchy-Schwarz inequality we had

$$M\gamma < \langle w^*, w^{T+1} \rangle \leq \|w^*\|_2 \|w^{T+1}\|_2$$
$$\implies M^2\gamma^2 < \|w^*\|_2^2 \|w^{T+1}\|_2^2$$

Using the bound $\|w^{T+1}\|_2^2 \leq MR^2$ we obtain:

$$M^2\gamma^2 < \|w^*\|_2^2 MR^2$$
$$\implies M < \frac{\|w^*\|_2^2 R^2}{\gamma^2}$$

Perceptron Convergence - Mistake Bound

Recall: We wanted to handle the inner product term:

$$\langle w^*, w^{T+1} \rangle > M\gamma$$

Then using Cauchy-Schwarz inequality we had

$$\begin{aligned} M\gamma < \langle w^*, w^{T+1} \rangle &\leq \|w^*\|_2 \|w^{T+1}\|_2 \\ \implies M^2\gamma^2 &< \|w^*\|_2^2 \|w^{T+1}\|_2^2 \end{aligned}$$

Using the bound $\|w^{T+1}\|_2^2 \leq MR^2$ we obtain:

$$\begin{aligned} M^2\gamma^2 &< \|w^*\|_2^2 MR^2 \\ \implies M &< \frac{\|w^*\|_2^2 R^2}{\gamma^2} \end{aligned}$$

Thus, assuming that $\|w^*\|_2$ and R can be controlled, the number of mistakes M is inversely proportional to γ, which determines the closeness of the data points to the separating hyperplane.

Perceptron Convergence - Mistake Bound

References:

- **H.D. Block**: The perceptron: A model for brain functioning. *Reviews of Modern Physics* 34, 123-135 (1962).
- **A.B.J. Novikoff**: On convergence proofs on perceptrons. In: *Proceedings of the Symposium on the Mathematical Theory of Automata*, vol. XII, pp. 615-622 (1962).

Perceptron Convergence - Mistake Bound

Two questions remain:

Perceptron Convergence - Mistake Bound

Two questions remain:

- How do we compute w^* and γ in the linear separability assumption ?

Perceptron Convergence - Mistake Bound

Two questions remain:

- How do we compute w^* and γ in the linear separability assumption ?
- What is the intuition behind the Perceptron update rule?

Perceptron Convergence - Mistake Bound

First question: How do we compute w^* and γ in the linear separability assumption ?

Perceptron Convergence - Mistake Bound

First question: How do we compute w^* and γ in the linear separability assumption ?

One possible way is to solve the following problem:

$$w^*, \gamma = \underset{u, \mu > 0}{\operatorname{argmin}}\, 0$$
$$\text{s.t. } y^t \langle u, x^t \rangle > \mu, \ \forall\ t = 1, 2, \ldots.$$

Perceptron Convergence - Mistake Bound

First question: How do we compute w^* and γ in the linear separability assumption ?
One possible way is to solve the following problem:

$$w^*, \gamma = \underset{u, \mu > 0}{\operatorname{argmin}}\, 0$$
$$\text{s.t. } y^t \langle u, x^t \rangle > \mu, \ \forall\ t = 1, 2, \ldots.$$

This optimization problem is a **linear program** and is called a **Feasibility problem**.

Perceptron Convergence - Mistake Bound

First question: How do we compute w^* and γ in the linear separability assumption ?
One possible way is to solve the following problem:

$$w^*, \gamma = \underset{u, \mu > 0}{\operatorname{argmin}} \, 0$$
$$\text{s.t. } y^t \langle u, x^t \rangle > \mu, \ \forall \ t = 1, 2, \ldots .$$

This optimization problem is a **linear program** and is called a **Feasibility problem**.
Caveat: Leads to infinitely many constraints.
Thus, we need a finite data set of training samples.

Perceptron Convergence - Mistake Bound

Question: How do we adapt the perceptron training to finite data sets?

Perceptron Convergence - Mistake Bound

Question: How do we adapt the perceptron training to finite data sets?

Perceptron Training Procedure For Finite Data

Input: $D = \{(x^i, y^i)\}_{i=1}^N$, $x^i \in \mathbb{R}^d$, $y^i \in \{+1, -1\}$.
$w^1 = 0$, $t = 1$.
while True **do**
  for $i \leftarrow 1, 2, 3, \ldots, N$ **do**
    receive (x^i, y^i) from D.
    $(x^t, y^t) = (x^i, y^i)$.
    $\widehat{y} = Perceptron(x^t; w^t)$
    if $\widehat{y} \neq y^t$ **then**
      $w^{t+1} = w^t + y^t x^t$
    else
      $w^{t+1} = w^t$
    $t = t + 1$

Perceptron Convergence - Mistake Bound

Second question:

- What is the intuition behind the Perceptron update rule?

Will see later!

Perceptron - Caveat

- Not suitable when **linear separability assumption** fails
- Example: Classical XOR problem

Perceptron - Caveat

- Not suitable when **linear separability assumption** fails
- Example: Classical XOR problem

Heavily criticized by M. Minsky and S. Papert in their book: **Perceptrons**, *MIT Press*, 1969.

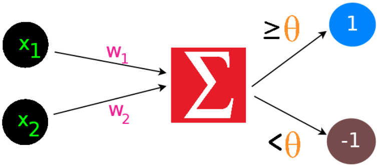
Perceptron - Caveat

- Not suitable when **linear separability assumption** fails
- Example: Classical XOR problem

x_1	x_2	$y = x_1 \oplus x_2$
0	0	-1
0	1	1
1	0	1
1	1	-1

Perceptron - Caveat

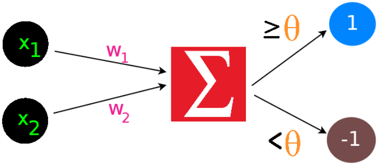
- Not suitable when **linear separability assumption** fails
- Example: Classical XOR problem



x_1	x_2	$y = x_1 \oplus x_2$	$\hat{y} = \text{sign}(w_1x_1 + w_2x_2 - \theta)$
0	0	-1	$\text{sign}(-\theta)$
0	1	1	$\text{sign}(w_2 - \theta)$
1	0	1	$\text{sign}(w_1 - \theta)$
1	1	-1	$\text{sign}(w_1 + w_2 - \theta)$

Perceptron - Caveat

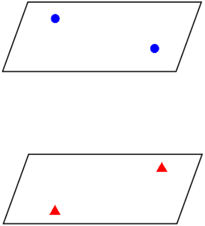
- Not suitable when **linear separability assumption** fails
- Example: Classical XOR problem



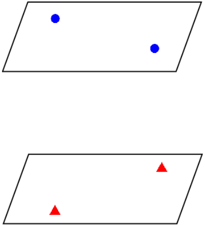
$$\text{sign}(-\theta) = -1 \implies \theta > 0$$
$$\text{sign}(w_2 - \theta) = 1 \implies w_2 - \theta \geq 0$$
$$\text{sign}(w_1 - \theta) = 1 \implies w_1 - \theta \geq 0$$
$$\text{sign}(w_1 + w_2 - \theta) = -1 \implies -w_1 - w_2 + \theta > 0$$

Note: This system is inconsistent. (Homework!)

Moving away from perceptron - Dealing with XOR problem

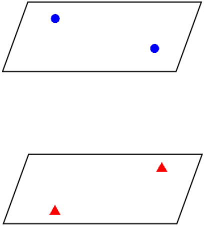


Moving away from perceptron - Dealing with XOR problem



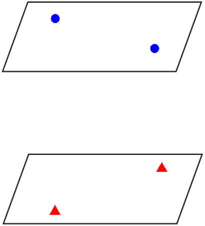
- Assume that the sample features $x \in \mathbb{R}^d$.

Moving away from perceptron - Dealing with XOR problem



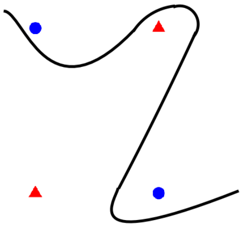
- Assume that the sample features $x \in \mathbb{R}^d$.
- **Idea:** Use a transformation $\psi : \mathbb{R}^d \to \mathbb{R}^q$, where $q \gg d$, to lift the data samples $x \in \mathbb{R}^d$ into $\psi(x) \in \mathbb{R}^q$ hoping to see a separating hyperplane in the transformed space.

Moving away from perceptron - Dealing with XOR problem

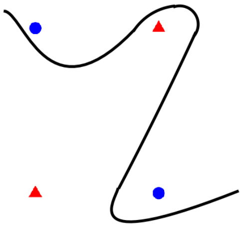


- Assume that the sample features $x \in \mathbb{R}^d$.
- **Idea:** Use a transformation $\psi : \mathbb{R}^d \to \mathbb{R}^q$, where $q \gg d$, to lift the data samples $x \in \mathbb{R}^d$ into $\psi(x) \in \mathbb{R}^q$ hoping to see a separating hyperplane in the transformed space.
- Forms the core idea behind kernel methods. (Will not be pursued in this course!)

Moving away from perceptron - Dealing with XOR problem

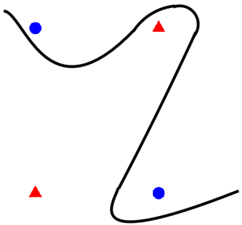


Moving away from perceptron - Dealing with XOR problem



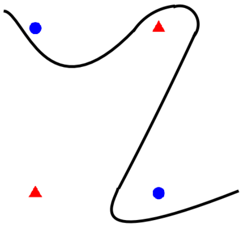
- **Idea:** The separating surface need not be linear and can be assumed to take some non-linear form.

Moving away from perceptron - Dealing with XOR problem



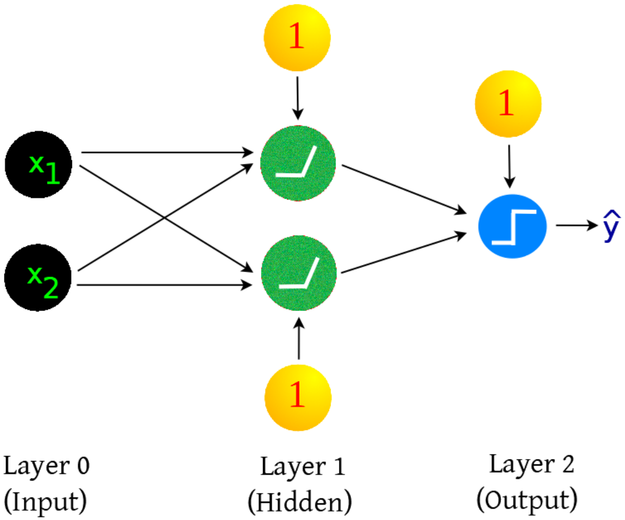
- **Idea:** The separating surface need not be linear and can be assumed to take some non-linear form.
- Hence for an input space $\mathcal{X}$ and output space $\mathcal{Y}$, the learned map $h : \mathcal{X} \to \mathcal{Y}$ can take some non-linear form.

Moving away from perceptron - Dealing with XOR problem

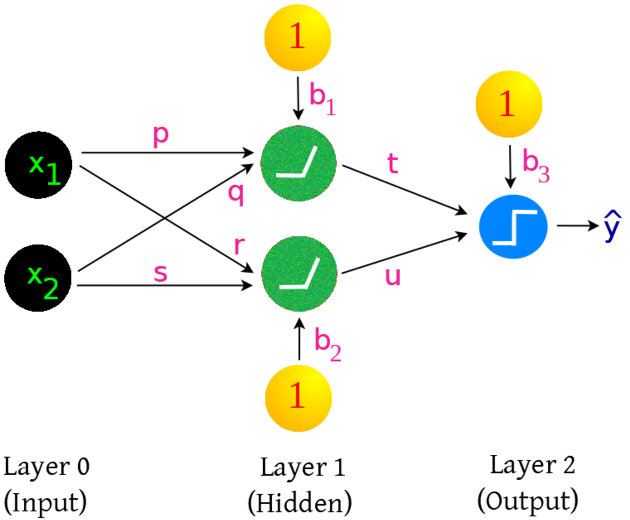


- **Idea:** The separating surface need not be linear and can be assumed to take some non-linear form.
- Hence for an input space $\mathcal{X}$ and output space $\mathcal{Y}$, the learned map $h : \mathcal{X} \to \mathcal{Y}$ can take some non-linear form.
- Forms the idea behind multi-layer perceptrons!

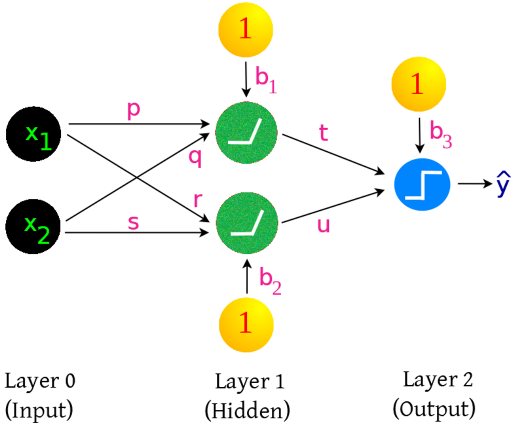
Moving away from perceptron - Dealing with XOR problem



Moving away from perceptron - Dealing with XOR problem



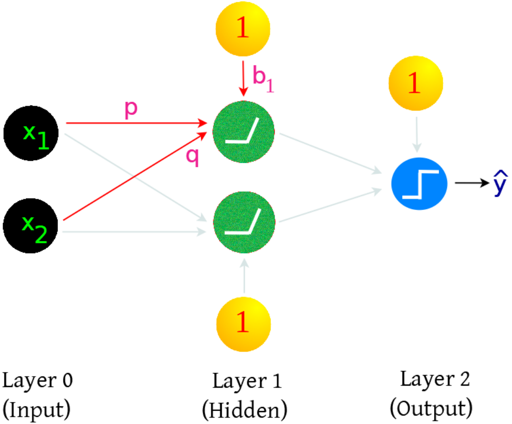
Moving away from perceptron - Dealing with XOR problem



Some notations

- n_k^ℓ denotes k-th neuron at layer ℓ.
- a_k^ℓ denotes the activation of the neuron n_k^ℓ.

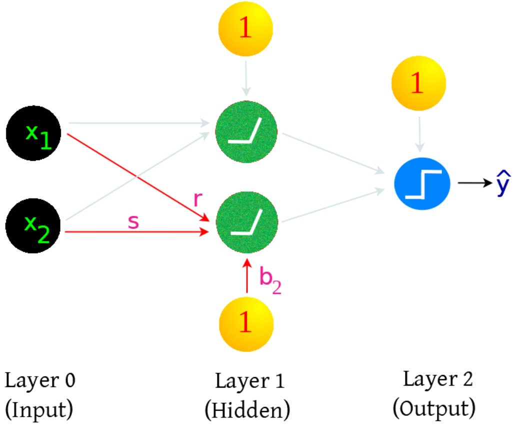
Moving away from perceptron - Dealing with XOR problem



- Activation at neuron n_1^1:

$$a_1^1 = \max\{px_1 + qx_2 + b_1, 0\}.$$

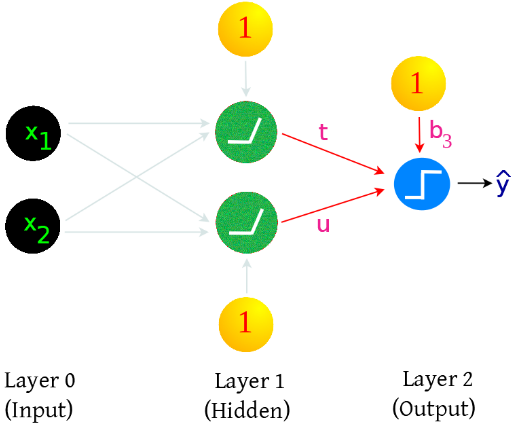
Moving away from perceptron - Dealing with XOR problem



- Activation at neuron n_2^1:

$$a_2^1 = \max\{rx_1 + sx_2 + b_2, 0\}.$$

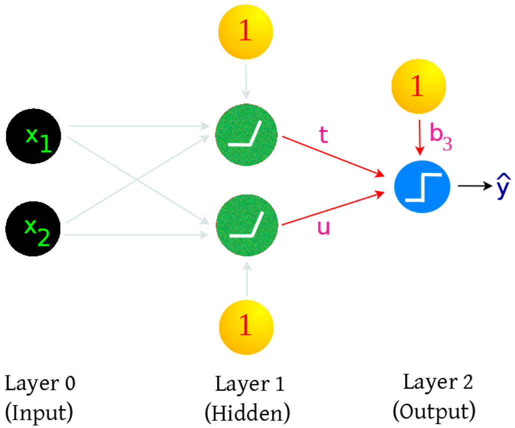
Moving away from perceptron - Dealing with XOR problem



- Activation at neuron n_1^2:

$$a_1^2 = \text{sign}(ta_1^1 + ua_2^1 + b_3).$$

Moving away from perceptron - Dealing with XOR problem

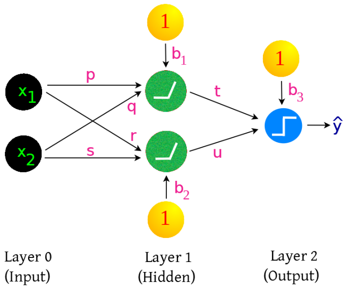


- Activation at neuron n_1^2:

$$a_1^2 = \text{sign}(ta_1^1 + ua_2^1 + b_3).$$

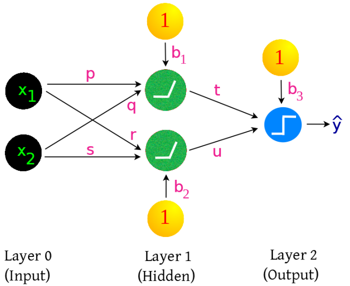
- **Note:** The activation a_1^2 is the output of the network denoted by $\hat{y}$.

Moving away from perceptron - Dealing with XOR problem



x_1	x_2	a_1^1	a_2^1	$\hat{y}$	y
0	0	$\max\{b_1, 0\}$	$\max\{b_2, 0\}$	$\text{sign}(ta_1^1 + ua_2^1 + b_3)$	-1
0	1	$\max\{q + b_1, 0\}$	$\max\{s + b_2, 0\}$	$\text{sign}(ta_1^1 + ua_2^1 + b_3)$	+1
1	0	$\max\{p + b_1, 0\}$	$\max\{r + b_2, 0\}$	$\text{sign}(ta_1^1 + ua_2^1 + b_3)$	+1
1	1	$\max\{p + q + b_1, 0\}$	$\max\{r + s + b_2, 0\}$	$\text{sign}(ta_1^1 + ua_2^1 + b_3)$	-1

Moving away from perceptron - Dealing with XOR problem



x_1	x_2	a_1^1	a_2^1	$\hat{y}$	y
0	0	$\max\{b_1, 0\}$	$\max\{b_2, 0\}$	$\text{sign}(ta_1^1 + ua_2^1 + b_3)$	-1
0	1	$\max\{q + b_1, 0\}$	$\max\{s + b_2, 0\}$	$\text{sign}(ta_1^1 + ua_2^1 + b_3)$	+1
1	0	$\max\{p + b_1, 0\}$	$\max\{r + b_2, 0\}$	$\text{sign}(ta_1^1 + ua_2^1 + b_3)$	+1
1	1	$\max\{p + q + b_1, 0\}$	$\max\{r + s + b_2, 0\}$	$\text{sign}(ta_1^1 + ua_2^1 + b_3)$	-1

Homework: Find weights $p, q, r, s, t, u, b_1, b_2, b_3$ such that the MLP solves the XOR problem.