Réseaux de Neurones Artificiels

M1 IDD 2019–2020 *Représentation des connaissances et raisonnement*

Stéphane Airiau

LAMSADE

DAUPHINE
UNIVERSITÉ PARIS

Apprentissage Automatique

- Apprentissage supervisé :
  - on a des exemples (beaucoup !) et une étiquette pour chaque exemple
  - on veut utiliser un algorithme pour faire ensuite de la prédiction pour l'étiquette : on donne un exemple, il faut devinier l'étiquette
- Apprentissage non supervisé :
  - on a des exemples, mais pas d'étiquette
  - clustering ⇨ découvrir une structure cachée dans les données
- apprentissage par renforcement :
  - un agent interagit avec un environnement
  - il reçoit une observation et un signal qui dépendent de l'action exécutée par l'agent

  ⇨ le **but** est de maximiser ce signal

Apprentissage Automatique

- Apprentissage supervisé : cours ML, puis en **M2-ID**, **M2 IASD**
  - on a des exemples (beaucoup !) et une étiquette pour chaque exemple
  - on veut utiliser un algorithme pour faire ensuite de la prédiction pour l'étiquette : on donne un exemple, il faut devinier l'étiquette
- Apprentissage non supervisé :
  - on a des exemples, mais pas d'étiquette
  - clustering ⇨ découvrir une structure cachée dans les données
- apprentissage par renforcement : cours en **M2 IASD**
  - un agent interagit avec un environnement
  - il reçoit une observation et un signal qui dépendent de l'action exécutée par l'agent
  - ⇨ le **but** est de maximiser ce signal

Dans ce cours, on va voir deux exemples d'apprentissage supervisé

- arbres de décision
- réseaux de neurones

Apprentissage supervisé

- on a un ensemble d'exemples
- on a une étiquette pour chaque exemple
  un exemple : données médicales sur un patient
  étiquette : est-ce que ce patient est atteint par une maladie X ?
- apprendre : pas apprendre "par coeur"

⇨ on voudra utiliser ce qu'on apprend pour de la prédiction
sur un exemple jamais vu, on devra deviner l'étiquette

⇨ l'apprentissage doit permettre de généraliser au delà des exemples donnés

⇨ comment valider une méthode ?

Réseaux de Neurones Artificiels

M1 IDD 2019–2020 *Représentation des connaissances et raisonnement*

Stéphane Airiau

LAMSADE

DAUPHINE
UNIVERSITÉ PARIS

- S'inspirer du cerveau humain pour construire un système
- Ici, on s'inspire des neurones
  ⇨ batir des réseaux de neurones artificiels
- vieille idée, McCulloch et Pitts ont présenté un modèle mathématique d'un neurone dès 1943.

Inspiration

- le cerveau : réseau interconnecté très complexe de neurones
- ⇨ Réseau de Neurones Artificiels (RNA) : réseau densémment connecté de simples unités de calcul
- le cerveau
  - $\approx 10^{11}$ neurones
  - chaque neurone est connecté à 10,000 autres neurones
  - le temps de réaction d'un neurone est autour de 10^{-3}s (lent par rapport à un ordinateur)
  - 0.1s pour réaliser une tâche de reconnaissance (ex : montre la photo d'une personnalité ou d'un ami)
    - ⇨ au plus quelques centaines d'étapes pour le calcul.
  - ⇨ calcul vraissemblablement massivement parallèle
- réseaux de neurones pour comprendre mieux le cerveau humain
- réseaux de neurones comme source d'inspiration pour un méchanisme d'apprentissage efficace

Réseau de neurones

- domaine de recherche ancien
- ex dès 1993, un RNA a été utilisé pour gérer le volant d'une voiture roulant sur une autoroute (aux Etats Unis)
  - input : images d'une caméra avec une résolution de 30x32 pixels
    ⇨ chaque pixel est connecté à un neurone d'entrée
  - 4 neurones cachés
  - 30 neurones pour la sortie (un neurone va correspondre à un angle du volant)

Situations pour utiliser un RNA

- entrée est un vecteur de large dimension de valeurs discrètes ou continues (e.g. sortie d'un capteur)
- la sortie est discrète ou continue
- la sortie est un vecteur de valeurs
- les valeurs d'entrées peuvent avoir du bruit (ou contenir des erreurs)
- on n'a pas d'hypothèse sur la fonction à apprendre
- il n'y a pas besoin qu'un humain puisse lire le résultat (par ex pour le vérifier ou comprendre ce qui est appris)
- un temps d'apprentissage long est acceptable
- l'utilisation du RNA est rapide

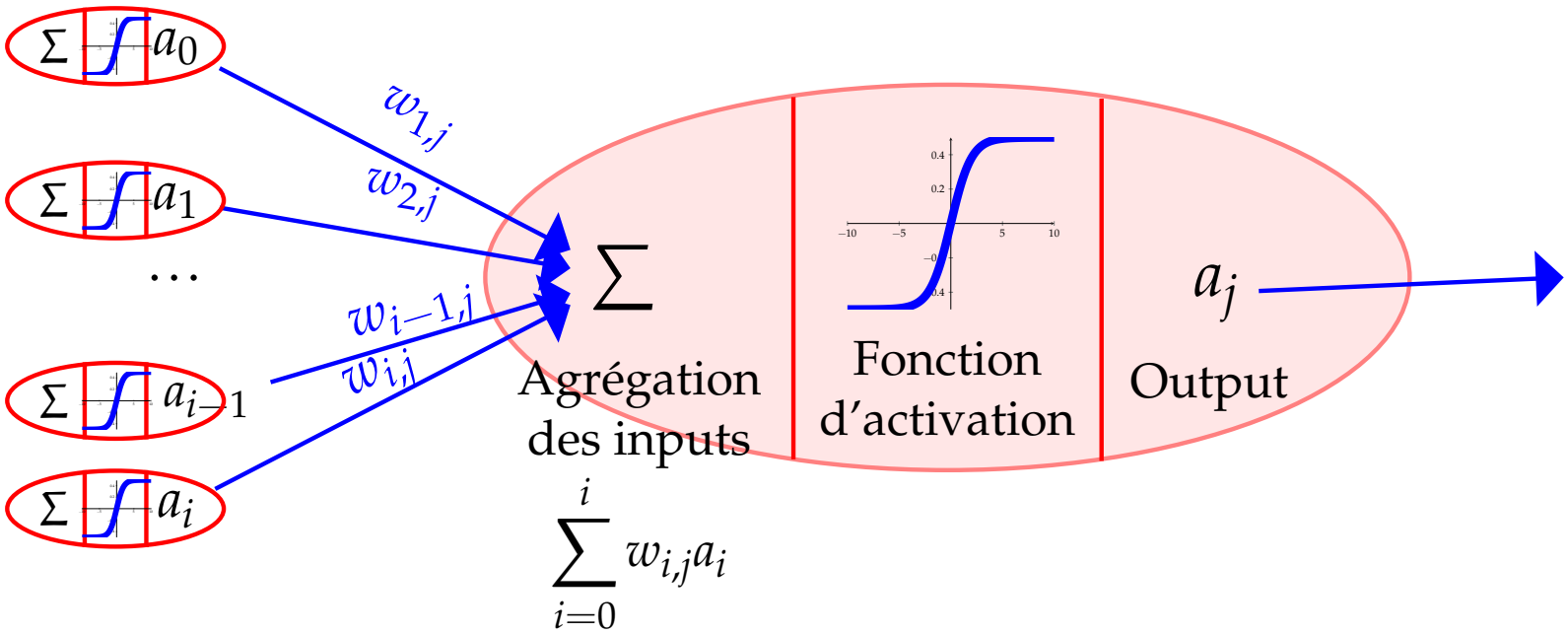
Exemples :

- Apprendre à reconnaître des sons
- Classification d'images
- Prédictions financiaires

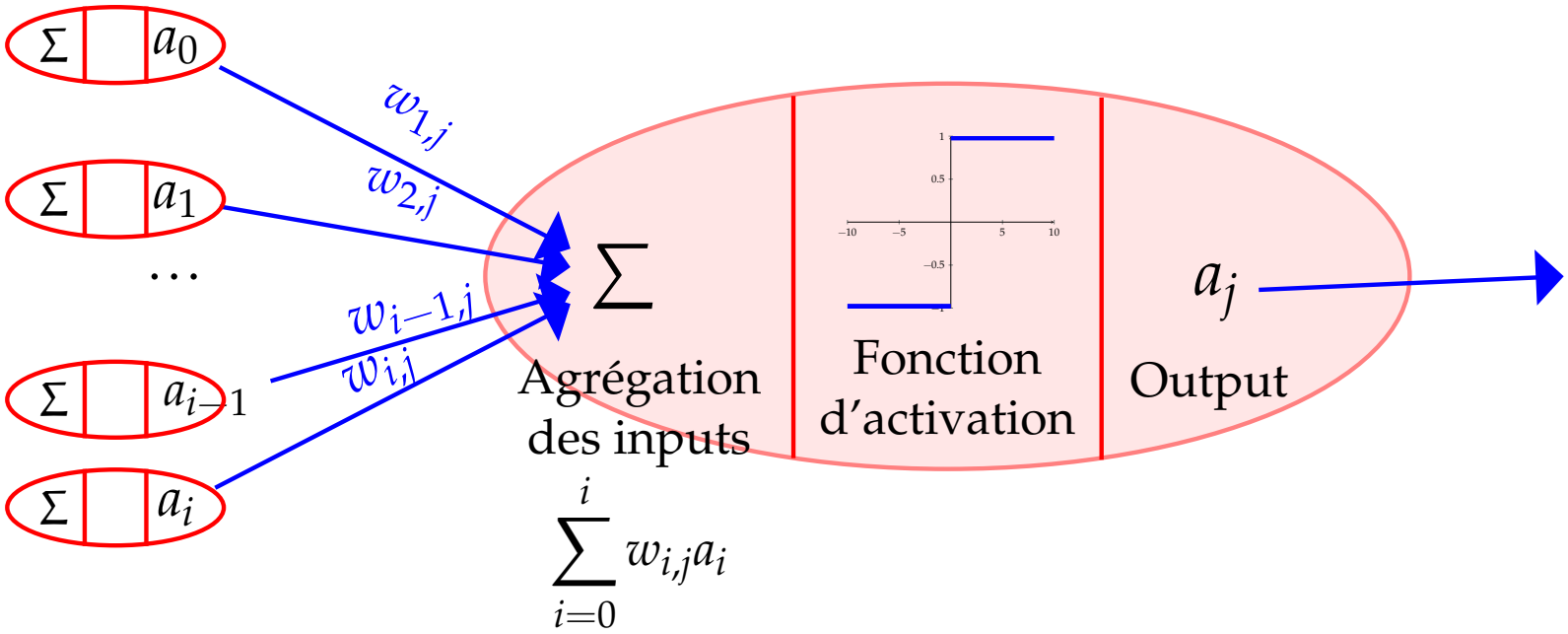
Deep Learning

- Coloration d'images en noir et blanc
- Traduction automatique
- classification d'objets dans des photos
- Game Playing.

Modèle d'un neurone : le perceptron



Modèle d'un neurone : le perceptron



- Fonction d'activation : fonction de Heaviside
- perceptron représente un hyperplan $\overrightarrow{w} \cdot \overrightarrow{x} = 0$ et les instances d'un côté de l'hyperplan seront positives et négatives de l'autre
- on peut représenter des fonction booléennes avec un seul perceptron, mais pas toutes (XOR).
- avec un réseau de perceptrons, on peut représenter toutes les fonctions booléennes

Apprendre pour **un** perceptron **linéaire**

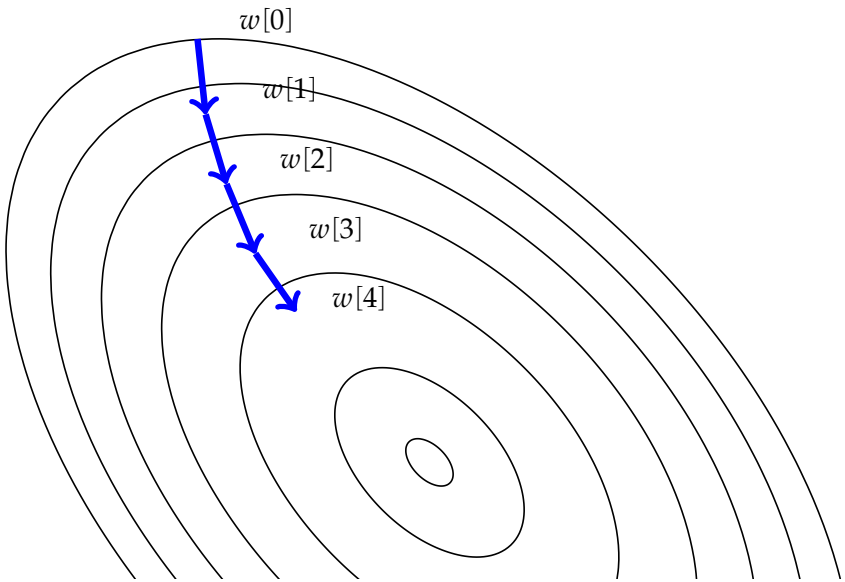
- On veut apprendre un vecteur de poids $\overrightarrow{w}$
- on va utiliser le principe la **descente de gradient** (la plus grande pente)
- L'erreur d'apprentissage est souvent mesurée par

$$E(\overrightarrow{w}) = \frac{1}{2} \sum_{d \in D} (t_d - o_d)^2$$

où

- D est l'ensemble de exemples d'apprentissage
- t_d est la vraie classification de l'instance d
- o_d est la réponse du peceptron pour l'instance d

Descente de gradient de l'erreur



le gradient indique la direction de la pente la plus forte

La règle de mise à jour sera donc :

$$\vec{w} \leftarrow \vec{w} + \Delta\vec{w} \qquad w_i \leftarrow w_i + \Delta w_i$$

$$\text{où } \Delta\vec{w} = -\eta\nabla E(\vec{w}) \qquad \Delta w_i = -\eta\frac{\partial E}{\partial w_i}$$

- η est le taux d'apprentissage qui determine la taille du pas que l'on fait en descendant
- le signe négatif indique que l'on veut descendre

Heureusement, calculer la dérivée est facile ici !

Descente de gradient avec fonction d'activation linéaire

$$
\begin{aligned}
\frac{\partial E}{\partial w_i} &= \frac{\partial}{\partial w_i}\left(\frac{1}{2}\sum_{d\in D}(t_d - o_d)^2\right) \\
&= \frac{1}{2}\sum_{d\in D}\frac{\partial}{\partial w_i}(t_d - o_d)^2 \\
&= \frac{1}{2}\sum_{d\in D}2(t_d - o_d)\frac{\partial}{\partial w_i}(t_d - o_d) \\
&= \sum_{d\in D}(t_d - o_d)\frac{\partial}{\partial w_i}(t_d - \vec{w}\cdot\vec{x}_d) \\
\frac{\partial E}{\partial w_i} &= \sum_{d\in D}(t_d - o_d)(-x_{i,d})
\end{aligned}
$$

La règle de mise à jour est donc

$$\Delta w_i = \eta\sum_{d\in D}(t_d - o_d)x_{i,d}$$

Algorithme de descente de gradient -v1 - avec fonction d'activation linéaire

```
1  initialise chaque w_i avec une valeur au hasard
2  tant que l'erreur est trop grande
3      Pour chaque i ∈ {1,...,n}
4          Δw_i = 0
5      Pour chaque example (x⃗, t) ∈ D
6          calculer la sortie o
7          Pour chaque i ∈ {1,...,n}
8              Δw_i = Δw_i + η(t − o)x_i
9      Pour chaque i ∈ {1,...,n}
10         w_i ← w_i + Δw_i
```

Algorithme pour **une** unité avec une fonction d'activation **linéaire**

Avec la fonction d'activation linéaire et un seul neurone, on a garanti d'un seul minimum global, mais ce ne sera pas toujours le cas.

Algorithme de descente de gradient stochastique

```
1  initialise chaque w_i avec une valeur au hasard
2  tant que l'erreur est trop grande
3      Pour chaque i ∈ {1,...,n}
4          Δw_i = 0
5      Pour chaque example (x⃗, t) ∈ D pris au hasard
6          calculer la sortie o
7          Pour chaque i ∈ {1,...,n}
8              Δw_i = Δw_i + η(t − o)x_i
9          Pour chaque i ∈ {1,...,n}
10             w_i ← w_i + Δw_i
```

Algorithme pour **une** unité avec une fonction d'activation **linéaire**

Avec la fonction d'activation linéaire et un seul neurone, on a garanti d'un seul minimum global, mais ce ne sera pas toujours le cas.

Descente de gradient stochastique

- la descente peut être lente
- s'il y a plusieurs minimaux locaux, on n'a pas de garantie de trouver le minimum global

On utilise souvent une variante qui consiste à mettre à jour les poids après l'examen de chaque point (et pas après de les avoir tous examinés)

- c'est souvent une bonne approximation
- demande un peu moins de calculs
- permet parfois d'éviter de tomber dans un minimum local
  ⇨ plus de chances de tomber dans me minimum global

On veut apprendre des fonctions **non linéaires**

- on veut représenter des fonctions non linéaires
- avec la discontinuité du perceptron, on ne peut pas calculer de dérivées
- on remplace la fonction d'activation par une fonction **sigmoïde** (ou fonction logistique) qui est une approximation continue et différentiable de la fonction seuil

$$\sigma(y) = \frac{1}{1+e^{-y}}$$

- on peut aussi utiliser la fonction tangente hyperbolique

$$tanh(y) = \frac{e^{y} - e^{-y}}{e^{y} + e^{-y}}$$

- ReLU : pour éviter le coût de calcul lié à calculer la fonction logistique ou tanh, on peut utiliser une fonction linéaire par morceaux quand même : en pratique, ça marche bien !
  $x \mapsto 0$ pour $x < 0$, $x \mapsto \lambda x$ pour $x \geqslant 0$

On peut refaire les calculs précédents pour calculer la nouvelle fonction de mise à jour pour ces fonctions d'activation.

Il faut passer au **réseau** de neurones!

Le **deep** learning consiste à avoir beaucoup de couches de neurones (ex α-Go utilise des réseaux de 13 couches).

On va avoir des réseaux **multi-couches**

- une "couche" pour accéder aux données (ex pixels d'une image)
- couche cachée 1 : une couche de neurones connectés aux neurones de la couche d'entrée
- couche cachée 2 : couche de neurones connectés aux neurones de la couche d'entrée ou au neurones de la couche cachée 1
- …
- couche cachée k : couche de neurones connectés aux neurones de la couche d'entrée, des couches 1, 2, …, k-1
- …
- couche de sortie

Problème : on sait mesurer l'erreur pour la couche de sortie (en comparant avec les étiquettes)

⇨ **comment mesurer l'erreur pour les neurones des couches cachées ?**

Fonction d'erreurs

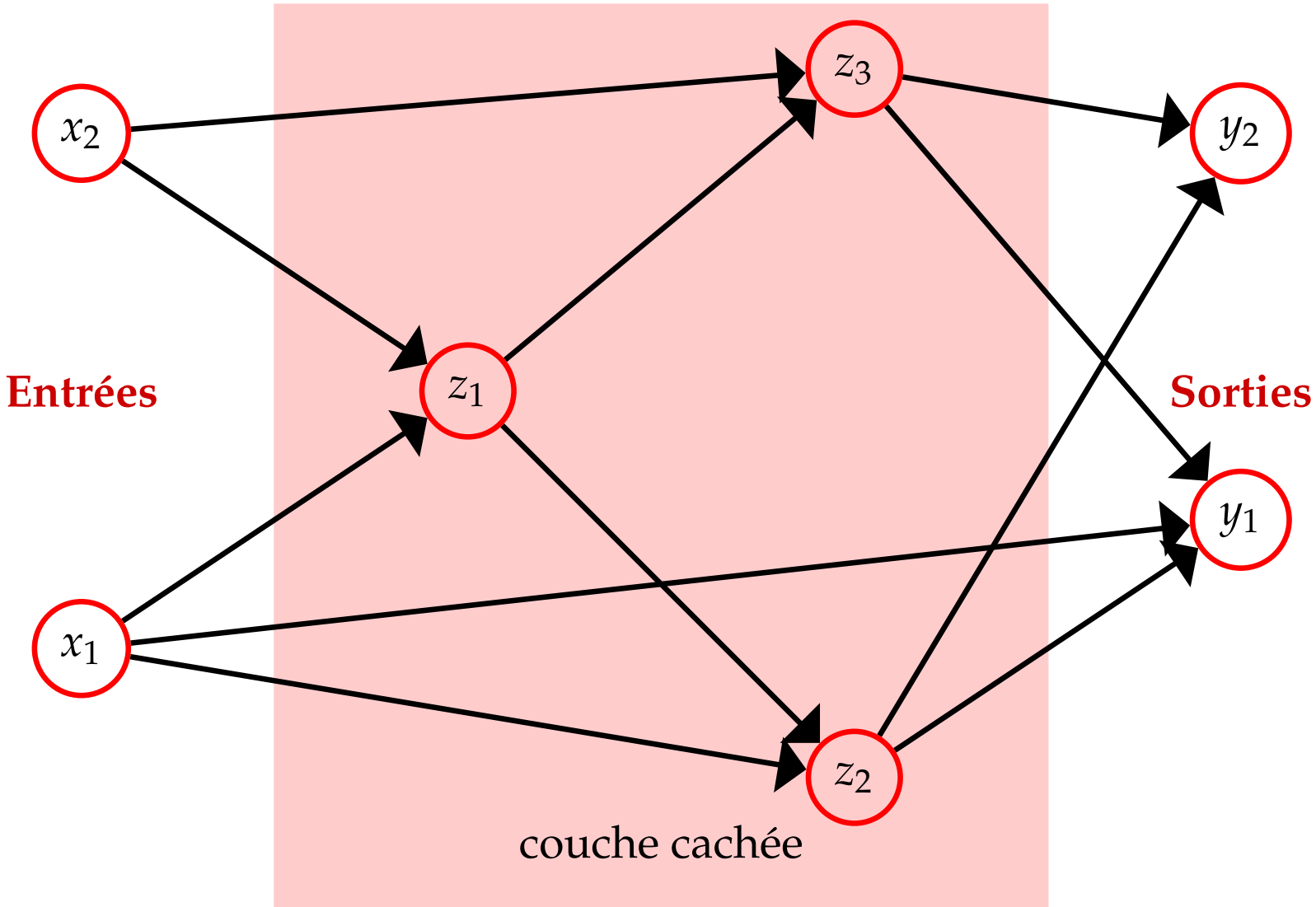
L'erreur d'apprentissage est mesurée par

$$E(\overrightarrow{w}) = \frac{1}{2} \sum_{d \in D} \sum_{k \in sorties} (t_{k,d} - o_{k,d})^2$$

où

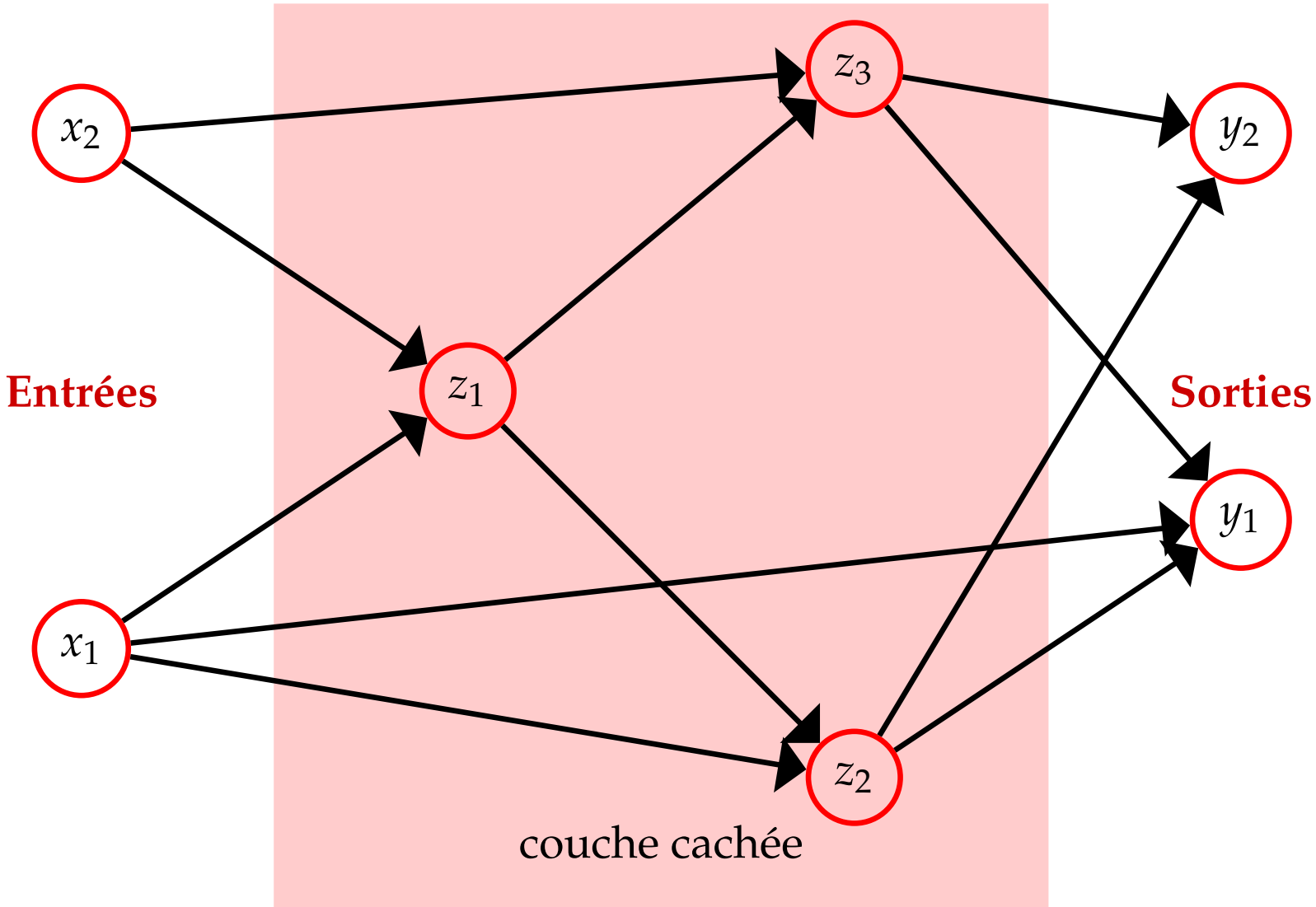
- D est l'ensemble des exemples d'apprentissage
- $t_{k,d}$ est la vraie classification de l'instance d pour la sortie k
- $o_{k,d}$ est la valeur de la sortie k pour l'instance d

Backpropagation



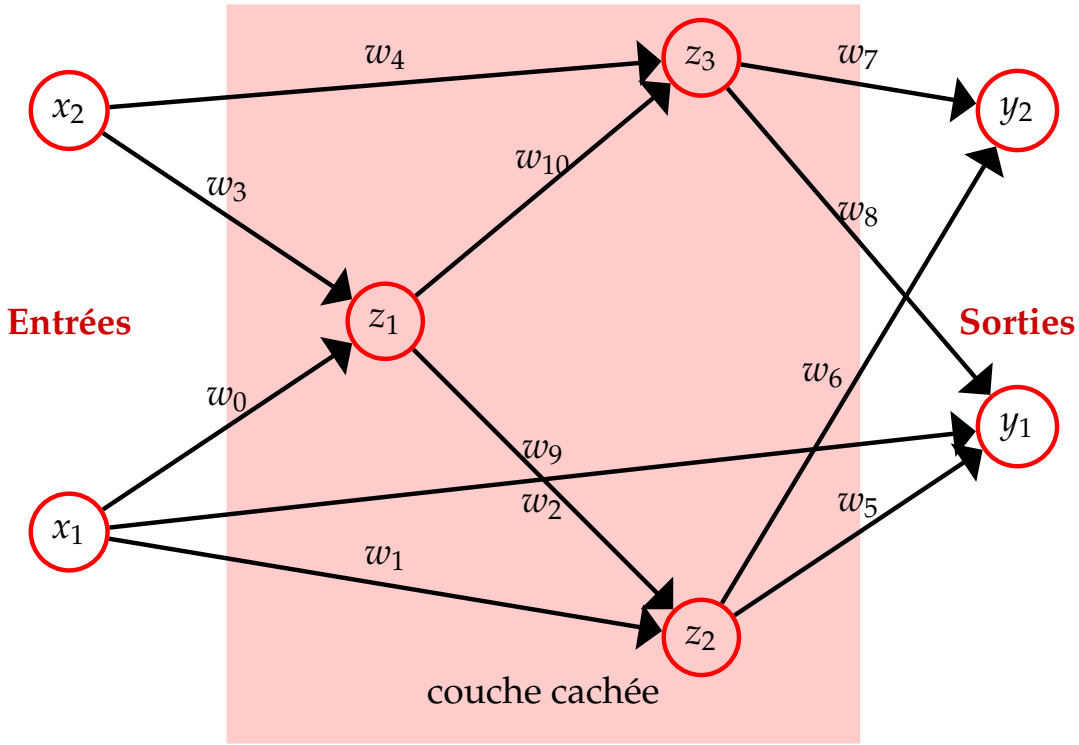
on peut calculer l'erreur aux sorties
mais il faut calculer une erreur pour les neurones de la couche interne !

Backpropagation



idée : on partage l'erreur en fonction des poids

Backpropagation



- z_3 contribue à la décision sur y_2 avec un poids de w_7
  ⇨on attribue à z_3 une partie de l'erreur de y_2 avec un poids de w_7
- z_3 contribue à la décision sur y_1 avec un poids de w_8
  ⇨on attribue à z_3 une partie de l'erreur de y_1 avec un poids de w_8

L'erreur de z_3 sera donc $w_7 \times erreur(y_2) + w_8 * erreur(y_1)$

Backpropagation algorithm

- Cet algorithme suppose que la fonction d'activation est la fonction sigmoïde.
- x_{ji} est l'entrée du noeud j venant du noeud i et w_{ji} est le poids correspondant.
- δ_n est l'erreur associée à l'unité n. Cette erreur joue le rôle du terme $t-o$ dans le cas d'une seule unité.

initialise chaque w_i avec une valeur au hasard
tant que l'erreur est trop grande
    Pour chaque example $(\overrightarrow{x}, t) \in D$
        1- calcul de l'état du réseau par propagation
        2- rétro propagation des erreurs
            a- **pour chaque** unité de sortie k, calcul du terme d'erreur
                $\delta_k \leftarrow o_k(1-o_k)(t_k-o_k)$
            b- **pour chaque** unité cachée h, calcul du terme d'erreur
                $\delta_h \leftarrow o_h(1-o_h) \sum_{k \in \text{Neurones en aval de } j} w_{k,h}\delta_k$
            c- mise à jour des poids w_{ji}
                $w_{j,i} \leftarrow w_{j,i} + \eta\delta_j x_{ji}$

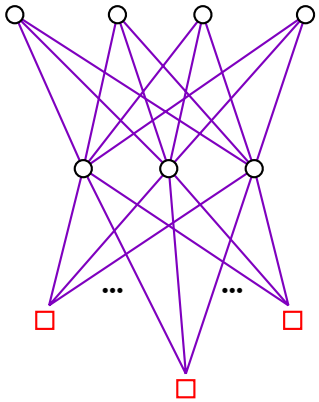
Propriétés

- backpropagation converge vers un minimum local (aucune garantie que le minimum soit global)
- en pratique, il donne de très bons résultats
  dans la pratique, le grand nombre de dimensions peut donner des opportunités pour ne pas être bloqué dans un minimum local
- pour le moment, on n'a pas assez de théorie qui explique les bons résultats en pratique
  - ajouter un terme de "momentum"
  - entrainer plusieur RNA avec les mêmes données mais différents poids de départ
    (boosting)

Quelles fonctions peut on représenter avec un RNA ?

- fonctions booléennes
- fonctions continues (théoriquement : toute fonction continue peut être approximée avec une erreur arbitraire par un réseau avec deux niveaux d'unités)
- fonctions arbitraires (théoriquement : toute fonction peut être approximée avec une précision aribitraire par un réseau avec 3 niveaux d'unités)

Exemple : reconnaissance de forme



20 personnes, environ 32 photos par personnes, leur tête peut être de face, tournée à gauche, droite, regardant en haut, différentes expressions (content, triste, en colère, neutre), avec ou sans lunettes de soleil. image de 120x128 pixels, noire et blanc, intensité entre 0 et 255.

taux de réussite de 90% pour apprendre l'orientation de la tête et reconnaître une des 20 personnes