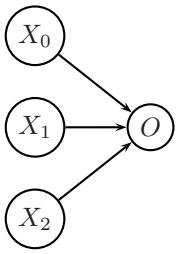
Ph. Leray

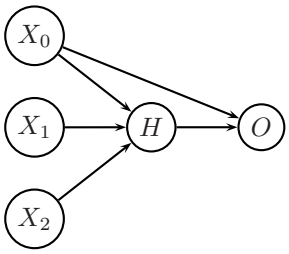
Réseaux de Neurones et Applications
Corrigé indicatif - Examen 2002/2003

4^{eme} année

1 Réseaux de neurones (15 points)



réseau 1



réseau 2

X_1	X_2	D
-1	-1	-1
-1	1	1
1	-1	1
1	1	-1

problème XOR

notations : $X_0 = 1$ (biais), fonction de coût quadratique $J = \frac{1}{2}(D-O)^2$

1. (1 point) Considérons le réseau 1, avec une sortie linéaire $O = A = \sum_{i=0}^{2} w_i X_i$.
   Calculer le (vecteur) gradient $\frac{\partial J}{\partial w_i}$ et la matrice Hessien $\frac{\partial^2 J}{\partial w_i \partial w_j}$

$$\begin{aligned} \frac{\partial J}{\partial w_i} &= \frac{\partial J}{\partial O}\frac{\partial O}{\partial w_i} = (O-D)X_i \\ \frac{\partial^2 J}{\partial w_i \partial w_j} &= \frac{\partial}{\partial w_j}\frac{\partial J}{\partial w_i} = X_i\frac{\partial O}{\partial w_j} = X_i X_j \end{aligned}$$

2. (2 points) Même question en changeant la fonction de transfert : $O = \frac{1-e^{-A}}{1+e^{-A}}$

$$\begin{aligned} \frac{\partial J}{\partial w_i} &= \frac{\partial J}{\partial O}\frac{\partial O}{\partial A}\frac{\partial A}{\partial w_i} = (O-D)\frac{2e^{-A}}{(1+e^{-A})^2}X_i = \frac{1}{2}(O-D)(1-O^2)X_i \\ \frac{\partial^2 J}{\partial w_i \partial w_j} &= \frac{\partial}{\partial w_j}\frac{\partial J}{\partial w_i} = \frac{1}{2}X_i\frac{\partial[(O-D)(1-O^2)]}{\partial w_j} \\ &= \frac{1}{2}X_i[(O-D)\frac{\partial(1-O^2)}{\partial w_j} + (1-O^2)\frac{\partial(O-D)}{\partial w_j}] \\ &= \frac{1}{2}X_i[(O-D)(-2O\frac{\partial O}{\partial w_j}) + (1-O^2)\frac{\partial O}{\partial w_j}] \\ &= \frac{1}{4}X_i X_j(1-O^2)[(1-O^2) - 2O(O-D)] \end{aligned}$$

3. (3 points) Considérons maintenant le réseau 2, avec les fonctions de transfert suivantes : $H = \frac{1-e^{-A}}{1+e^{-A}}$ et $O = w_3 X_0 + w_4 H$. Calculer le (vecteur) gradient et la matrice Hessien

couche de sortie : $u_i = \{w_3 \;\; w_4\}, Z_i = \{X_0 \;\; H\}$

$$\begin{aligned} \frac{\partial J}{\partial u_i} &= \frac{\partial J}{\partial O}\frac{\partial O}{\partial u_i} = (O-D)Z_i \\ \frac{\partial^2 J}{\partial u_i \partial u_j} &= Z_i Z_j \end{aligned}$$

couche cachée :

$$\begin{aligned}
\frac{\partial J}{\partial w_i} &= \frac{\partial J}{\partial O}\frac{\partial O}{\partial H}\frac{\partial H}{\partial A}\frac{\partial A}{\partial w_i} = (O-D)w_4\frac{1}{2}(1-H^2)X_i \\
\frac{\partial^2 J}{\partial w_i \partial w_j} &= \frac{1}{2}w_4 X_i \frac{\partial[(O-D)(1-H^2)]}{\partial w_j} \\
&= \frac{1}{2}w_4 X_i[(O-D)\frac{\partial(1-H^2)}{\partial w_j} + (1-H^2)\frac{\partial(O-D)}{\partial w_j}] \\
&= \frac{1}{2}w_4 X_i[(O-D)(-2H\frac{\partial H}{\partial w_j}) + (1-H^2)\frac{\partial O}{\partial w_j}] \\
&= \frac{1}{2}w_4 X_i[(O-D)(-2H\frac{1}{2}(1-H^2)X_j) + (1-H^2)w_4\frac{1}{2}(1-H^2)X_j] \\
&= \frac{1}{4}w_4 X_i X_j(1-H^2)[w_4(1-H^2) - 2H(O-D)]
\end{aligned}$$

4. (4 points) On rajoute un terme à la fonction de coût :

$$J = \frac{1}{2}(D-O)^2 + \sum_i \frac{(w_i/\alpha)^2}{1+(w_i/\alpha)^2}$$

   (a) Quelle est l'utilité de ce terme supplémentaire ?

   Ce terme permet de limiter l'amplitude des poids lors de l'apprentissage pour garantir un meilleur estimateur (au sens de l'erreur en généralisation). On peut rapprocher cela à la régression ridge.

   (b) Quelles sont les répercussions sur le gradient (faites le calcul) ?

   Cela rajoute à $\frac{\partial J}{\partial w_i}$ un terme correspondant à la dérivée du terme de pénalisation

$$\frac{\partial}{\partial w_i}(\frac{(w_i/\alpha)^2}{1+(w_i/\alpha)^2}) = 2\frac{w_i/\alpha^2}{(1+(w_i/\alpha)^2)^2}$$

5. (5 points) On considère maintenant les réseaux 1 et 2 avec une fonction de transfert signe et le problème XOR.

   (a) Choix de la structure :
   Les réseaux 1 ou 2 sont-ils capables de résoudre ce problème. Si c'est le cas, expliquer pourquoi, sinon proposer un autre réseau du même genre résolvant le problème.

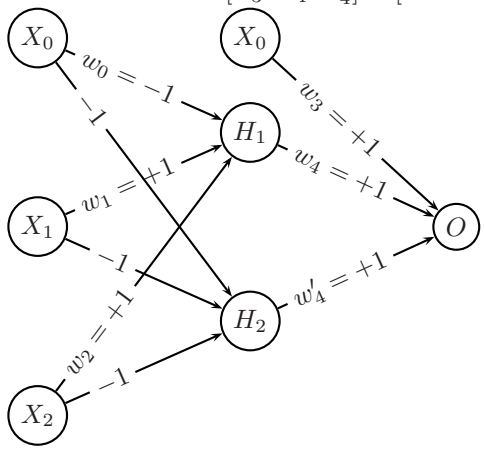
   L'équation du réseau 1 correspond à celle d'un plan dans l'espace $\{X_1, X_2\}$. Si la fonction de transfert est la fonction signe, le réseau 1 ne permet de résoudre que des problèmes linéairement séparables, ce qui n'est pas le cas du problème XOR.
   L'équation du réseau 2 permet seulement de déplacer le plan déterminé par la couche cachée. Ce n'est pas suffisant pour résoudre XOR (on peut même noter que tout problème linéairement séparable résolvable par le réseau 2 peut l'être par le réseau 1 tout seul !)
   Pour résoudre XOR, il faut donc combiner 2 frontières de décision, ce qui ne peut se faire qu'en utilisant 2 neurones dans la couche cachée.

   (b) Détermination des paramètres :
   Déterminer la valeur des poids du réseau choisi à la question précédente pour résoudre le problème XOR.

   Imaginons que l'on cherche seulement à distinguer le point en "haut à droite" (1,1) des 3 autres. Dans le plan $\{X_1, X_2\}$, cela correspond à la frontière de décision $X_2 = -X_1 + 1$ ou $X_1 + X_2 - 1 = 0$ (soit $[w_0 \;\; w_1 \;\; w_2] = [-1 \;\; 1 \;\; 1]$)
   Idem pour en "bas à gauche" pour distinguer le point (-1,-1) des 3 autres, avec la frontière de décision $X_2 = -X_1 - 1$ (soit $[w'_0 \;\; w'_1 \;\; w'_2] = [-1 \;\; -1 \;\; -1]$)

X_1	X_2	D	haut&droite	bas&gauche	(haut&droite+bas&gauche)
-1	-1	-1	-1	1	0
-1	1	1	-1	-1	-2
1	-1	1	-1	-1	-2
1	1	-1	1	-1	0

Supposons que nos 2 neurones cachés calculent les valeurs "haut&droite" et "bas&gauche", on s'apercoit facilement que $D = -1 - (haut\&droite + bas\&gauche)$, ce qui nous donne les poids de la couche cachée $[w_3 \;\; w_4 \;\; w'_4] = [-1 \;\; -1 \;\; -1]$)



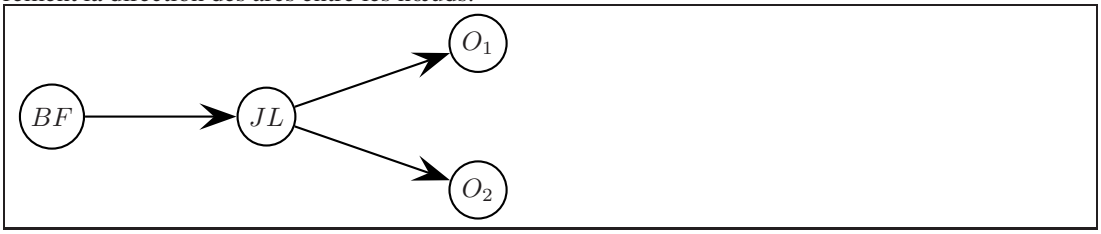
2 Juggler, le robot jongleur (15 pts)

Juggler, le robot jongleur, lache souvent les balles avec lesquelles il jongle quand sa batterie est faible. D'après les expériences précédentes, il a été déterminé que la probabilité qu'il lache une balle quand sa batterie est faible est 0.9. D'autre part, quand sa batterie n'est pas faible, la probabilité qu'il lache une balle est seulement 0.01. La batterie ayant été rechargée il y a peu de temps, il y a seulement 5% de chances que la batterie soit faible. Un premier système de vision (peu fiable) observe le robot et nous prévient lorsqu'il croit que Juggler a laché une balle. Un autre système (indépendant du premier) agit de la même façon.

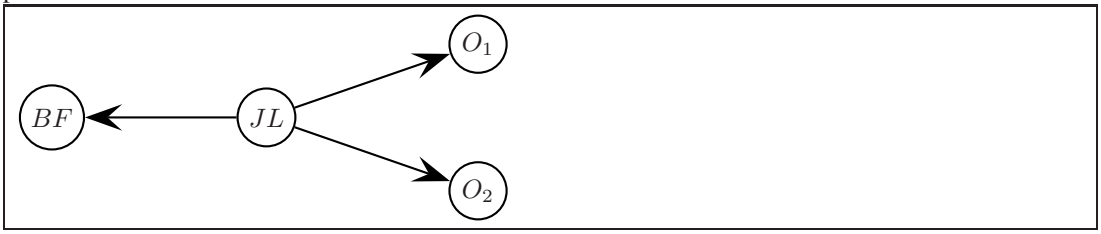
1. (5 points) Structure du réseau bayésien
   (a) En supposant que toutes vos variables sont booléennes, quelles variables allez vous choisir pour modéliser ce problème ?

BF	La Batterie est Faible
JL	Juggler lache une balle
O_1	L'observateur 1 affirme que Juggler a laché une balle
O_2	L'observateur 2 affirme que Juggler a laché une balle

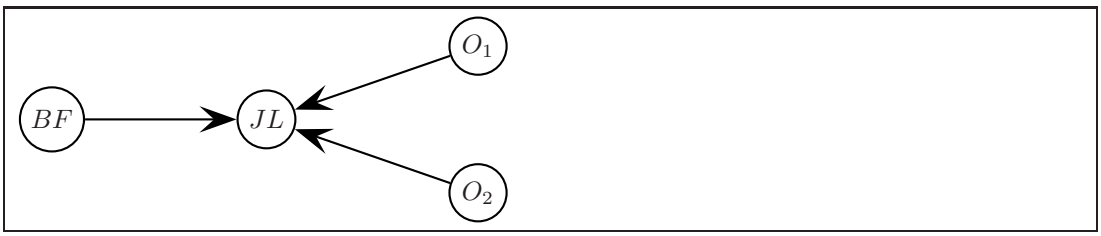
   (b) Proposer un réseau bayésien B_1 correspondant au problème. Etiqueter les nœuds de réseau et indiquer clairement la direction des arcs entre les nœuds.



   (c) Proposer une autre structure B_2 équivalente (au sens de Markov) à celle que vous avez précédemment proposée.



   (d) Proposer une structure B_3 possédant le même graphe non orienté que B_1, mais n'étant pas équivalente (au sens de Markov).



2. (2 points) Probabilités conditionnelles
   (a) Représenter les tables de probabilité correspondant à l'énoncé et à votre structure B_1.
   (b) Comment la qualité des deux observateurs est-elle "codée" dans le réseau bayésien ?

p(BF)	
BF=t	0.05
BF=f	0.95

p(JL | BF)	BF=t	BF=f
JL=t	0.90	0.01
JL=f	0.10	0.99

p(O1 | JL)	JL=t	JL=f
O1=t	precision1	1-p1
O1=f	1-p1	precision1

p(O2 | JL)	JL=t	JL=f
O2=t	precision2	1-p2
O2=f	1-p2	precision2

La précision des 2 observateurs peut se représenter sur la table de probabilité conditionnelle en fixant la valeur $precision1$ (resp. $precision2$). Notons que pour cet exercice nous supposons que la précision est la même sur la détection de l'évenement "Juggler lache une balle" et "Juggler n'a pas laché de balle" alors que nous aurions pu les distinguer en distinguant les précisions des 2 colonnes dans les tables de probas.

3. (8 points) Inférence (Message Passing)
   On suppose maintenant que la fiabilité de O_1 (resp. O_2) est de 70% (resp. 90%).
   (a) Calculer les messages λ et π circulant dans le réseau lorsqu'il n'y a aucune évidence ajoutée.

$\lambda(O_1) = [1\ \ 1]$ (feuille non instantiée)
$\lambda(O_2) = [1\ \ 1]$ (feuille non instantiée)

O_1 envoie un msg λ à son père JL :
$\lambda_{O_1}(JL) = \sum \lambda(O_1 = x)p(O1 = x|JL) = [1\ \ 1]$

O_2 envoie un msg λ à son père JL :
$\lambda_{O_2}(JL) = \sum \lambda(O_2 = x)p(O2 = x|JL) = [1\ \ 1]$

JL a reçu tous ces msgs λ et les "condense" : $\lambda(JL) = \lambda_{O_1}(JL).\lambda_{O_2}(JL) = [1\ \ 1]$

JL envoie un msg λ à son père BF :
$\lambda_{JL}(BF) = \sum \lambda(JL = x)p(JL = x|BF) = [1\ \ 1]$

BF a reçu tous ces msgs λ et les "condense" : $\lambda(BF) = \lambda_{JL}(BF) = [1\ \ 1]$

$\pi(BF) = [0.05\ \ 0.95]$ (racine)

$P(BF) = \lambda(BF).\pi(BF) = [0.05\ \ 0.95]$

mais ce n'est pas fini ...

... ca repart dans l'autre sens :
BF envoie un msg π à son fils JL :
$\pi_{JL}(BF) \propto \pi(BF) = [0.05 \;\; 0.95]$

JL calcule son message π :
$\pi(JL) = [0.9 * 0.05 + 0.01 * 0.95 \;\; 0.1 * 0.05 + 0.99 * 0.95] = [0.0545 \;\; 0.9455]$

$P(JL) = \lambda(JL).p(JL) = [0.0545 \;\; 0.9455]$

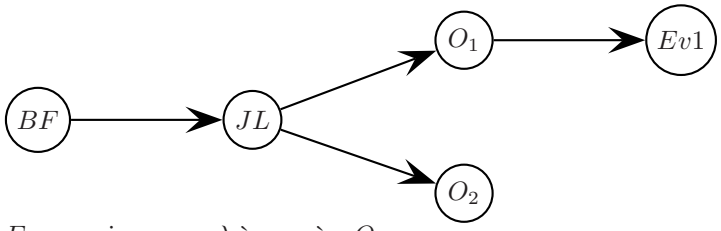
JL envoie un msg π à ses fils O_1 et O_2 :
$\pi_{O_1}(JL) \propto \pi(JL).\lambda_{O_2}(JL) = [0.0545 \;\; 0.9455]$
$\pi_{O_2}(JL) \propto \pi(JL).\lambda_{O_1}(JL) = [0.0545 \;\; 0.9455]$

O_1 calcule son message π :
$\pi(O_1) = [0.7 * 0.0545 + 0.3 * 0.9455 \;\; 0.3 * 0.0545 + 0.7 * 0.9455] = [0.3218 \;\; 0.6782]$

$P(O_1) = \lambda(O_1).p(O_1) = [0.3218 \;\; 0.6782]$

O_2 calcule son message π :
$\pi(O_2) = [0.9 * 0.0545 + 0.1 * 0.9455 \;\; 0.1 * 0.0545 + 0.9 * 0.9455] = [0.1436 \;\; 0.8564]$
$P(O_2) = \lambda(O_2).p(O_2) = [0.1436 \;\; 0.8564]$

(b) O_1 observe que Juggler a laché une balle. Quelle est la probabilité que la batterie soit faible sachant cela ?



Ev_1 envoie un msg λ à son père O_1 :
$\lambda_{Ev_1}(O_1) = [1 \;\; 0]$

O_1 a reçu tous ces msgs λ et les "condense" : $\lambda(O_1) = \lambda_{Ev_1}(O_1) = [1 \;\; 0]$

O_1 envoie un msg λ à son père JL :
$\lambda_{O_1}(JL) = \sum \lambda(O_1 = x)p(O1 = x|JL) = [0.7 \;\; 0.3]$

O_2 envoie un msg λ à son père JL :
$\lambda_{O_2}(JL) = \sum \lambda(O_2 = x)p(O2 = x|JL) = [1 \;\; 1]$

JL a reçu tous ces msgs λ et les "condense" : $\lambda(JL) = \lambda_{O_1}(JL).\lambda_{O_2}(JL) = [0.7 \;\; 0.3]$

JL envoie un msg λ à son père BF :
$\lambda_{JL}(BF) = \sum \lambda(JL = x)p(JL = x|BF) = [0.66 \;\; 0.304]$

BF a reçu tous ces msgs λ et les "condense" : $\lambda(BF) = \lambda_{JL}(BF) = [0.66 \;\; 0.304]$

$\pi(BF) = [0.05 \;\; 0.95]$ (racine)
$P(BF|EV_1) \propto \lambda(BF).\pi(BF) = [0.033 \;\; 0.2888]$
$P(BF|EV_1) = [0.1025 \;\; 0.8975]$

mais ce n'est pas fini ...

... ca repart dans l'autre sens : (pas demandé dans la question ...)
BF envoie un msg π à son fils JL :
$\pi_{JL}(BF) \propto \pi(BF) = [0.05 \;\; 0.95]$

JL calcule son message π :
$\pi(JL) = [0.9 * 0.05 + 0.01 * 0.95 \;\; 0.1 * 0.05 + 0.99 * 0.95] = [0.0545 \;\; 0.9455]$

$P(JL|EV_1) \propto \lambda(JL).p(JL) = [0.1186 \;\; 0.8814]$

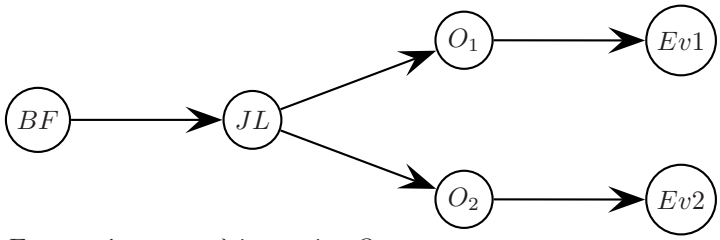
JL envoie un msg π à ses fils O_1 et O_2 :
$\pi_{O_1}(JL) \propto \pi(JL).\lambda_{O_2}(JL) = [0.0545 \;\; 0.9455]$
$\pi_{O_2}(JL) \propto \pi(JL).\lambda_{O_1}(JL) = [0.1186 \;\; 0.8814]$

O_1 calcule son message π :
$\pi(O_1) = [0.7 * 0.0545 + 0.3 * 0.9455 \;\; 0.3 * 0.0545 + 0.7 * 0.9455] = [0.3218 \;\; 0.6782]$

$P(O_1|EV_1) \propto \lambda(O_1).p(O_1) = [1 \;\; 0]$

O_2 calcule son message π :
$\pi(O_2) = [0.9 * 0.1186 + 0.1 * 0.8814 \;\; 0.1 * 0.1186 + 0.9 * 0.8814] = [0.1949 \;\; 0.8051]$
$P(O_2|EV_1) \propto \lambda(O_2).p(O_2) = [0.1949 \;\; 0.8051]$

(c) On ajoute alors une information supplémentaire : O_2 n'a rien vu (à la différence d'O_1). Quelle est la probabilité que la batterie soit faible sachant ces deux informations ?



Ev_1 envoie un msg λ à son père O_1 :
$\lambda_{Ev_1}(O_1) = [1\ \ 0]$

O_1 a reçu tous ces msgs λ et les "condense" : $\lambda(O_1) = \lambda_{Ev_1}(O_1) = [1\ \ 0]$

O_1 envoie un msg λ à son père JL :
$\lambda_{O_1}(JL) = \sum \lambda(O_1 = x)p(O1 = x|JL) = [0.7\ \ 0.3]$

Ev_2 envoie un msg λ à son père O_2 :
$\lambda_{Ev_2}(O_2) = [0\ \ 1]$

O_2 a reçu tous ces msgs λ et les "condense" : $\lambda(O_2) = \lambda_{Ev_2}(O_2) = [0\ \ 1]$

O_2 envoie un msg λ à son père JL :
$\lambda_{O_2}(JL) = \sum \lambda(O_2 = x)p(O2 = x|JL) = [0.1\ \ 0.9]$

JL a reçu tous ces msgs λ et les "condense" : $\lambda(JL) = \lambda_{O_1}(JL).\lambda_{O_2}(JL) = [0.07\ \ 0.27]$

JL envoie un msg λ à son père BF :
$\lambda_{JL}(BF) = \sum \lambda(JL = x)p(JL = x|BF) = [0.09\ \ 0.0268]$

BF a reçu tous ces msgs λ et les "condense" : $\lambda(BF) = \lambda_{JL}(BF) = [0.09\ \ 0.0268]$

$\pi(BF) = [0.05\ \ 0.95]$ (racine)
$P(BF|EV_1, EV_2) \propto \lambda(BF).\pi(BF) = [0.0045\ \ 0.2546]$
$P(BF|EV_1, EV_2) = [0.0174\ \ 0.9826]$