Part I

Théorie des réseaux de neurones

1 Backward propagation à travers le réseau

En utilisant la règle de la chaine :

$$\frac{\partial E}{\partial w} = \sum_j \frac{\partial E}{\partial y_j} \frac{\partial y_j}{\partial w}$$

Connaitre $\frac{\partial E}{\partial Y}$ et $\frac{\partial Y}{\partial w}$ nous assure donc de connaitre $\frac{\partial E}{\partial w}$.
Or on a, toujours d'après la règle de la chaine :

$$\frac{\partial E}{\partial x_i} = \sum_j \frac{\partial E}{\partial y_j} \frac{\partial y_j}{\partial x_i}$$

Connaitre $\frac{\partial E}{\partial Y}$ et $\frac{\partial Y}{\partial X}$ nous assure donc de connaitre $\frac{\partial E}{\partial X}$.
Et si $n \in \mathbf{N}$ alors $\frac{\partial E}{\partial X_{n+1}}$ correspond à $\frac{\partial E}{\partial Y_n}$.

2 Fully connected

Forward propagation :
$Y = XW + B$

Ici : $X = \begin{bmatrix} x_1 & x_2 & \cdots & x_i \end{bmatrix}$, $W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1j} \\ \vdots & \vdots & \ddots & \vdots \\ w_{i1} & w_{i2} & \cdots & w_{ij} \end{bmatrix}$, $B = \begin{bmatrix} b_1 & b_2 & \cdots & b_i \end{bmatrix}$

D'où $Y = \begin{bmatrix} \sum\limits_i x_i \times w_{i1} + b_1 & \cdots & \sum\limits_i x_i \times w_{ij} + b_j \end{bmatrix}$.

Backward propagation :

$$\frac{\partial E}{\partial W} = \begin{bmatrix} \frac{\partial E}{\partial w_{1,1}} & \cdots & \frac{\partial E}{\partial w_{1,j}} \\ \vdots & \ddots & \vdots \\ \frac{\partial E}{\partial w_{i,1}} & \cdots & \frac{\partial E}{\partial w_{i,j}} \end{bmatrix}$$

$$\frac{\partial E}{\partial X} = \frac{\partial E}{\partial Y} W^t$$

$$\frac{\partial E}{\partial W} = X^t \cdot \frac{\partial E}{\partial Y}$$

$$\frac{\partial E}{\partial B} = \frac{\partial E}{\partial Y}$$

Démonstration :

Règle de la chaine :

$$\frac{\partial E}{\partial w_{i,j}} = \frac{\partial E}{\partial y_1}\frac{\partial y_i}{\partial w_{i,j}} + \dots + \frac{\partial E}{\partial y_j}\frac{\partial y_j}{\partial w_{i,j}} = \frac{\partial E}{\partial y_j} x_i$$

D'où :

$$\frac{\partial E}{\partial W} = \begin{bmatrix} \frac{\partial E}{\partial y_1} x_1 & \cdots & \frac{\partial E}{\partial y_j} x_1 \\ \vdots & \ddots & \vdots \\ \frac{\partial E}{\partial y_1} x_i & \cdots & \frac{\partial E}{\partial y_j} x_i \end{bmatrix} = X^t \frac{\partial E}{\partial Y}$$

$$\frac{\partial E}{\partial B} = \begin{bmatrix} \frac{\partial E}{\partial b_1} & \cdots & \frac{\partial E}{\partial b_j} \end{bmatrix}$$

Règle de la chaîne :

$$\frac{\partial E}{\partial b_i} = \frac{\partial E}{\partial y_1}\frac{\partial y_1}{\partial b_i} + \dots + \frac{\partial E}{\partial y_j}\frac{\partial y_j}{\partial b_i} = \frac{\partial E}{\partial y_i}$$

D'où

$$\frac{\partial E}{\partial B} = \begin{bmatrix} \frac{\partial E}{\partial y_1} & \cdots & \frac{\partial E}{\partial y_j} = \frac{\partial E}{\partial Y} \end{bmatrix}$$

$$\frac{\partial E}{\partial X} = \begin{bmatrix} \frac{\partial E}{\partial x_1} & \cdots & \frac{\partial E}{\partial x_i} \end{bmatrix}$$

Règle de la chaîne :

$$\frac{\partial E}{\partial x_i} = \frac{\partial E}{\partial y_1}\frac{\partial y_1}{\partial x_i} + \dots + \frac{\partial E}{\partial y_j}\frac{\partial y_j}{\partial x_i} = \frac{\partial E}{\partial y_1} w_{i,1} + \dots + \frac{\partial E}{\partial y_j} w_{i,j}$$

D'où :

$$\frac{\partial E}{\partial X} = \begin{bmatrix} \frac{\partial E}{\partial y_1} w_{1,1} + \dots + \frac{\partial E}{\partial y_j} w_{1,j} & \cdots & \frac{\partial E}{\partial y_1} w_{i,1} + \dots + \frac{\partial E}{\partial y_j} w_{i,j} \end{bmatrix} = \frac{\partial E}{\partial Y} W^t$$

3 Activation Layer

Forward propagation :

$Y = \begin{bmatrix} f(x_1) & \cdots & f(x_i) \end{bmatrix}$

Backward propagation :

$$\frac{\partial E}{\partial X} = \frac{\partial E}{\partial Y} \odot f'(X) = \begin{bmatrix} \frac{\partial E}{\partial y_1} \times f'(x_1) & \cdots & \frac{\partial E}{\partial y_i} \times f'(x_i) \end{bmatrix}$$

Où $\odot$ est le produit d'Hadamard.

Démonstration :

On remarque i=j

$$\frac{\partial E}{\partial X} = \begin{bmatrix} \frac{\partial E}{\partial x_1} & \cdots & \frac{\partial E}{\partial x_i} \end{bmatrix}$$
$$= \begin{bmatrix} \frac{\partial E}{\partial y_1} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial E}{\partial y_i} \frac{\partial y_i}{\partial x_i} \end{bmatrix}$$
$$= \begin{bmatrix} \frac{\partial E}{\partial y_1} f^{'}(x_1) & \cdots & \frac{\partial E}{\partial y_i} f^{'}(x_i) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial E}{\partial y_1} & \cdots & \frac{\partial E}{\partial y_1} \end{bmatrix} \odot \begin{bmatrix} f^{'}(x_1) & \cdots & f^{'}(x_i) \end{bmatrix}$$
$$= \begin{bmatrix} \frac{\partial E}{\partial y_1} & \cdots & \frac{\partial E}{\partial y_1} \end{bmatrix} \odot f^{'}(X)$$

4 Théorie des GAN

4.1 Définitions

Définition 4.1.0.1 *Tribu :*
Soit E un ensemble. On appelle tribu tout sous-ensemble $\mathcal{A}$ de $\mathcal{P}(E)$ tel que :

- $\emptyset \in \mathcal{A}$
- *si $A \in \mathcal{A}$, alors $\bar{A} \in \mathcal{A}$*
- *si $(A_n)_{n\in N}$ est une suite de $\mathcal{A}$, alors $\cup_{n\in N} A_n \in \mathcal{A}$*

Un espace mesurable est un ensemble muni d'une tribu.

Définition 4.1.0.2 *Mesure :*
Soit $(E, \mathcal{A})$ un espace mesurable.
On appelle mesure sur E toute application $\mu : \mathcal{A} \to R^+$ vérifiant :

- $\mu(\emptyset) = 0$
- *Si $(A_n)_{n\in N}$ est une suite d'éléments de $\mathcal{A}$ deux à deux disjoints, alors $\mu(\sqcup_{n\in N} A_n) = \sum_{n\in N} \mu(A_n)$*

Définition 4.1.0.3 *Fonction mesurable : Soit $(E, \mathcal{T})$ et $(F, \mathcal{A})$ deux espaces mesurables.*
$f : E \to F$ est dite mesurable ssi $f^{-1}(\mathcal{A}) \subset \mathcal{T}$

Définition 4.1.0.4 *Variable aléatoire continue :*
$X : \Omega \to R$ mesurable.

Remarque 1 *Les densité de probabilités sont des mesures dont l'intégrale sur l'espace vaut 1.*

Théorème 4.1.0.1 *$(\Omega, \mathcal{T}, P)$ un espace probabilisé, $X : \Omega \to R$ de loi notée P_X.*
Soit $\phi : R \to R_+$ mesurable.
Alors $\int_\Omega \phi(X(\omega))dP(w) = \int_R \phi(x)dP_X(x)$

Corollaire 4.1.0.1.1 $E(\phi(X)) = \int_R \phi(x)f(x)dx$

4.2 Discriminant optimal

Démonstration :
On note p_g la densité créée par le générateur, p_r la densité de données réelles.

A G fixé, on cherche D_G^* le discriminant optimal (qui maximise la fonction $D \mapsto V(G, D)$)

$V(G, D) = E_{x \sim p_r}[log(D(x)] + E_{z \sim p_z}[log(1 - D(G(z))]$
$= \int_x p_r(x) log(D(x)) dx + \int_z p_z(z) log(1 - D(G(z))) dz$

changement de variable : x = G(z) (par Théorème de transfert)

$\int_z p_z(z) log(1 - D(G(z))) dz = \int_x p_g(x) log(1 - D(x)) dx$

Alors :
$V(G, D) = \int_x p_g(x) log(1 - D(x)) + p_r(x) log(D(x)) dx$
Soit $f : x \mapsto a \cdot log(x) + b \cdot log(1 - x)$, $a, b > 0$

f est maximale en $\frac{a}{a+b}$ sur [0, 1]

Alors $V(G, D) \leq \int_x p_g(x) log(1 - \frac{p_r(x)}{p_g(x)+p_r(x))}) + p_r(x) log(\frac{p_r(x)}{p_g(x)+p_r(x))}) dx$

$D_G^*(x) \in [0, 1]$ (C'est bien une probabilité, cohérent)

On pose $C(G) = V(G, D_G^*)$

$$D_G^*(x) = \frac{p_r(x)}{p_g(x) + p_r(x))}$$

4.3 Divergence de Kullblack-Leibler

Caractère positif :
$D(p \mid\mid q) = \int_x p(x) \cdot log(\frac{p(x)}{q(x)}) dx = - \int_x p(x) \cdot log(\frac{q(x)}{p(x)}) dx$
$\geq -log(\int_x p(x) \cdot \frac{q(x)}{p(x)} dx)$ (Inégalité de Jensen)
$= -log(\int_x q(x) dx)$
$= 0$

Cas d'égalité:
La divergence de Kullblack-Leibler est positive.
De plus le log est strictement concave : il ya donc égalité dans l'égalité précédente si et seulement si $x \mapsto \frac{p(x)}{q(x)}$ est constante p-presque partout.
Alors il existe $\alpha \in R_+^*$ tel que $p = \alpha \cdot q$.
De plus, $\int_x p(x) dx = \int_x q(x) dx = 1$, donc $\alpha = 1$ i.e. p = q p-presque partout.

4.4 Divergence de Jensen-Shannon

Caractère fini et distance :
$\forall p, q, JS(p \,||\, q)$ est fini :
$D(p \,||\, \frac{p+q}{2}) = \int_x p(x) \cdot log(\frac{2 \cdot p(x)}{p(x)+q(x)})dx$
$\leq \int_x p(x) \cdot log(\frac{2 \cdot p(x)}{p(x)})dx$
$\leq log(2)$

La divergence de Jensen-Shannon est bien une distance, car symétrique, positive, et nulle si et seulement si q et p sont égales p-presque partout.

Calcul de $C(G)$:
$\int_x p_r(x) log(\frac{p_r(x)}{p_g(x)+p_r(x))})dx = \int_x p_r(x)[log(\frac{p_r(x)}{\frac{p_g(x)+p_r(x))}{2}}) - log(2)]dx$
$= \int_x p_r(x) log(\frac{p_r(x)}{\frac{p_g(x)+p_r(x))}{2}})dx - log(2)$
$= D(p_r \,||\, \frac{p_r+p_g}{2}) - log(2)$

$\int_x p_g(x) log(\frac{p_r(x)}{p_g(x)+p_r(x))})dx = \int_x p_g(x)[log(\frac{p_r(x)}{\frac{p_g(x)+p_r(x))}{2}}) - log(2)]dx$
$= \int_x p_g(x) log(\frac{p_r(x)}{\frac{p_g(x)+p_r(x))}{2}})dx - log(2)$
$= D(p_g \,||\, \frac{p_r+p_g}{2}) - log(2)$

Alors $C(G) = 2 \cdot JS(p_r \,||\, p_g) - log(4)$

5 Distance de Wassertein

Soit $\theta \in [-1; 1]$ Si $\forall (x, y) \in P, x = 0, y \sim U([0; 1])$ et $\forall (x, y) \in Q, x = \theta, y \sim U([0; 1])$

$$|\theta| \neq 0 :$$

$D_{KL}(P||Q) = \sum_{x=0, y \sim U([0;1])} (1 \cdot log\frac{1}{0}) = +\infty$
$D_{KL}(Q||P) = \sum_{x=\theta, y \sim U([0;1])} (1 \cdot log\frac{1}{0}) = +\infty$
$D_{JS}(Q||P) = \frac{1}{2} \cdot (\sum_{x=0, y \sim U([0;1])} 1 \cdot log\frac{1}{1/2} + \sum_{x=\theta, y \sim U([0;1])} 1 \cdot log\frac{1}{1/2}) = log(2)$
$W(P, Q) = |\theta|$

$$|\theta| = 0 :$$

$W(P, Q) = D_{KL}(P|Q) = D_{KL}(Q|P) = D_{JS}(Q|P) = |\theta| = 0$