

# Statistiques descriptives

# Analyse de données

# Algorithmes

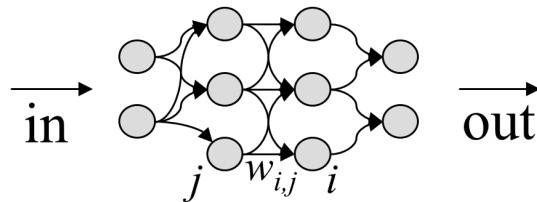
(5) MACHINES DE BOLTZMANN

Jérôme Lacaille  
Expert Émérite Safran

# INTRODUCTION

## ● Perceptron

- (Rosenblatt, 1962)
- (Rumelhart, McClelland 1986)



$$x_i(t+1) = \rho_i$$

$$\rho_i = \varphi \left( \sum_j w_{i,j} x_j(t) - \theta_i \right)$$

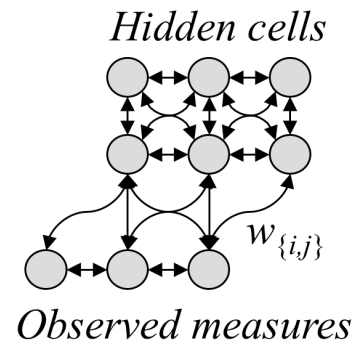
$$\varphi(u) = \frac{1}{1 + e^{-u}}$$

Calibrated by retro-propagation algorithm

**Supervised model**

## ● Gibbs Field

- (Hopfield, 1982)
- (Hinton, Sejnowski, 1983)



$$P[x_i(t+1) = 1 | x(t)] = \rho_i$$

$$P[x_i(t+1) = 0 | x(t)] = 1 - \rho_i$$

Calibrated by Hebbian learning rule

**Unsupervised model**



# CHAMPS DE GIBBS

1. **CHAMP MARKOVIAN**
2. **RÉSEAU DE NEURONES SYMÉTRIQUE**
3. **PROCESSUS MARKOVIAN ET DYNAMIQUE DE GLAUBER**

→ **RÈGLE DE HEBB**

# CHAMP MARKOVIAN (DE GIBBS)

On définit une structure de voisinage en associant à chaque site  $i$  un sous-ensemble des sites  $N_i$  connectés à  $i$ .

*Jusqu'à présent les sites  $i = 1 \dots p$  étaient vus comme des coordonnées de vecteurs aléatoires ou d'observation. Dans le cadre des champs, le site est placé dans une structure de voisinage (deux sites peuvent être connectés).*

Un champ aléatoire (variable aléatoire sur un espace muni d'un voisinage)  $X$  sur  $\mathbb{X}^{\{1 \dots p\}}$  est dit markovien si pour tout sous-ensemble fini  $V$  de  $\{1 \dots p\}$ , la loi de  $X$  sur  $V$  conditionnellement à  $\{1 \dots p\} \setminus V$  ne dépend que du voisinage de  $V$ .

$\mathbb{X}$  s'appelle la fibre du champ, l'ensemble des valeurs que peuvent prendre les activités. On a souvent  $\mathbb{X} = \{0,1\}$ , on parle alors de champ binaire.



# THÉORÈME DE DOBRUSHIN

Soient  $\Pi_i(x_{N_i}, x_i) \in [0,1]$  une famille de spécifications locales markoviennes.

Il existe une unique mesure  $M$  dont les probabilités conditionnelles sont définies par les  $\Pi_i$  et

$$P_M[X_i = x_i | X_{N_i} = x_{N_i}] = P_M[x_i | x_{N_i}] = \Pi_i(x_{N_i}, x_i)$$

La dynamique de Glauber issue de cette spécification définit un processus markovien qui converge vers  $M$ .

$$P[x_i(n+1) = y_i | x_{N_i}(n)] = \Pi_i(x_{N_i}(n), y_i)$$



# POTENTIEL D'INTERACTION

On appelle potentiel d'interaction une fonction qui associe à tous sous-ensemble de site  $V$  une fonction  $I_V: \mathbb{X}^V \mapsto \mathbb{R}$  ne dépendant que des activités  $x_i$  sur  $i \in V$ .

- Le potentiel  $I_V$  est dit d'ordre au plus  $k$  s'il est nul sur toute partie  $V$  de cardinal strictement supérieur à  $k$ .  
Exemple : un potentiel d'interaction d'ordre deux charge uniquement l'ensemble vide, les singletons et les paires.
- Un potentiel d'interaction d'ordre 2 symétrique sur une fibre binaire s'écrit :

$$I_{\{i,j\}}(x) = w_0 + w_i x_i + w_j x_j + w_{\{i,j\}} x_i x_j$$

On appelle énergie la somme du potentiel sur tous les sous-ensembles de sites.

$$H(x) = \sum_V I_V(x)$$

On note aussi usuellement pour toute partie  $W$  de sites  $H_W(x) = \sum_{V; V \cap W \neq \emptyset} I_V(x)$ .



# THÉORÈME DE SULLIVAN

**Si  $M$  est un champ markovien, alors il existe un potentiel d'interaction  $I$  tel que pour tout sous-ensemble de sites  $V$  et tout couple de configurations  $(x, y)$**

$$P_M(x_V | y_{\setminus V}) = \frac{1}{Z_T(y)} \exp[-H_V(x_V \wedge y_{\setminus V})/T]$$

$Z_T(y)$  est une constante de normalisation dépendant du paramètre de température  $T$ .



# RÔLE DE LA TEMPÉRATURE

Si l'on observe deux configurations  $x_1$  et  $x_2$ , on peut calculer leur rapport de vraisemblance :

$$\frac{P_M(x_1)}{P_M(x_2)} = \exp \left[ - \frac{H(x_1) - H(x_2)}{T} \right]$$

En diminuant  $T$  on augmente le terme dans l'exponentielle donc on amplifie l'écart entre les deux probabilités.

Par exemple, si  $x_1$  est la configuration d'énergie minimale, le terme sous l'exponentielle est positif et d'autant plus grand que la température est petite.

**La température est donc un moyen d'influer sur le contraste entre deux configurations générées par la loi  $M$ .**

Si on augmente la température, on ramène le terme en exponentielle vers 0 et toutes les configurations deviennent équiprobables.





# RECUITS SIMULÉS, ALGORITHME GÉNÉTIQUE

## Un recuit simulé consiste à

1. partir d'une température élevée,
2. tirer aléatoirement sous la loi  $M$  une configuration initiale,
3. la modifier légèrement en ne changeant d'un site à la fois suivant les spécifications markoviennes,
4. puis de faire décroître doucement la température tout en continuant à modifier notre configuration suivant le processus markovien des spécifications,
5. pour converger progressivement vers une configuration d'énergie minimale (dite configuration fondamentale).

## Un recuit simulé parallèle consiste à

- suivre simultanément les trajectoires de plusieurs recuits simulés,
- en déduire la répartition empirique des configurations fondamentales.

## Un recuit simulé génétique (ou algorithme génétique) consiste à

- exécuter un recuit simulé parallèle,
- tout en autorisant des modifications locales supplémentaires par combinaison de configurations, les cross-over, ce qui augmente le nombre de configurations,
- mais tout en se limitant à une taille de population stable en sélectionnant depuis la nouvelles population augmentée les configurations par tirage aléatoire suivant leur probabilité, c'est ce que l'on appelle la pression de sélection.



# RÉSEAU DE NEURONE SYMÉTRIQUE

Une structure de voisinage est symétrique si quand un site  $i$  est connecté à un site  $j$  alors il existe une connexion réciproque de  $j$  vers  $i$ .

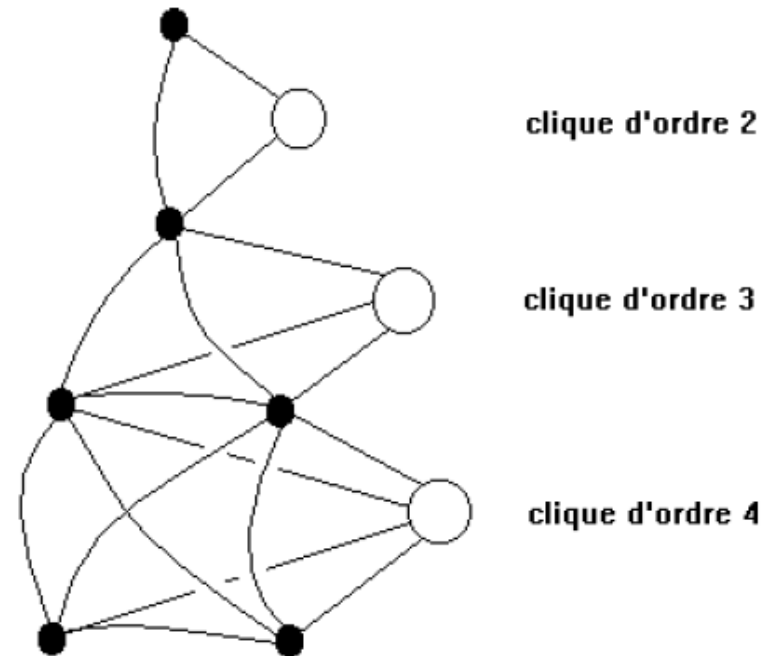
$$\forall(i, j), \quad j \in N_i \Leftrightarrow j \in N_j$$

On note  $i \leftrightarrow j$

On appelle clique tout sous-ensemble de sites formé de points simultanément voisins.

$$\exists \mathcal{C} \text{ clique telle que } \{i, j\} \subset \mathcal{C} \Leftrightarrow i \leftrightarrow j$$

Un sous ensemble de taille  $m$  est une clique si on peut vérifier  $\frac{m(m-1)}{2}$  relations de voisinages symétriques.



*L'ordre d'une clique et le nombre d'éléments qu'elle contient.*



# THÉORÈME DE HAMERSLEY-CLIFFORD

Si la loi du champ  $M$  charge toutes les configurations  $x$ , alors il existe un potentiel d'interaction  $I$  non nul uniquement sur les cliques du réseau et tel que

$$P_M(x) = \frac{1}{Z_T} \exp[-H(x)/T]$$

avec

$$H(x) = \sum_{C \text{ clique}} I_C(x)$$

*L'énergie se concentre sur les cliques.*

**Ce résultat n'est malheureusement pas valable quand la connectivité du réseau n'est pas symétrique.**



# RÉSEAUX BINAIRE SYMÉTRIQUE

**On appelle réseau binaire symétrique un réseau de neurones binaires à connexions symétriques et à interactions d'ordre 2 au plus.**

En supprimant les constantes résiduelles (les intégrant dans la constante de normalisation  $Z_T$ ), et en considérant qu'il existe toujours un site constant ( $x_0 = 1$ ), on peut écrire

$$H(x) = - \sum_{i \leftrightarrow j} w_{\{i,j\}} x_i x_j$$

La loi du champ aléatoire s'écrit

$$P_M(x) = \frac{1}{Z_T} \exp \left( \frac{1}{T} \sum_{i \leftrightarrow j} w_{\{i,j\}} x_i x_j \right)$$



# SPÉCIFICATIONS DES RÉSEAUX BINAIRES SYMÉTRIQUES

## La spécification locale

$$\begin{aligned} P_M(x_i = 1 | x_{-i}) &= \frac{1}{Z_T(x_{-i})} \exp \left( \frac{1}{T} \sum_{j \in N_i} w_{\{i,j\}} x_j \right) \\ &= \frac{1}{1 + \exp \left( -\frac{1}{T} \sum_{j \in N_i} w_{\{i,j\}} x_j \right)} \\ &= \psi_T(v_i) \end{aligned}$$

On retrouve

$$v_i = \sum_{j \in N_i} w_{\{i,j\}} x_j = H(x_{-i}, x_i = 0) - H(x_{-i}, x_i = 1)$$



# CHAÎNE DE MARKOV HOMOGÈNE

Une chaîne de Markov sur un espace probabilisé  $\Omega$  est un processus discret donné par ses probabilités de transitions.

$$P(x(n+1) = y \mid x(n) = x) = Q_n(x, y)$$

La chaîne est homogène si la matrice de transition ne dépend pas du temps.

$$Q_n = Q$$

La donnée de la loi initiale  $P_0(x(0))$  et la matrice de transition  $Q$  définissent entièrement le processus.



# CHAÎNE DE MARKOV IRRÉDUCTIBLE

Une chaîne de Markov homogène est irréductible si pour tout couple d'états  $(x, y)$  il est possible d'aller de  $x$  à  $y$  en un temps fini.

$$\forall (x, y) \in \Omega^2, \exists n > 0, P[x(n) = y | x(0) = x] = Q^n(x, y) > 0$$

Si une chaîne de Markov est irréductible, alors on finit toujours par repasser par chaque état.

$$\forall x \exists n > 0, Q^n(x, x) > 0$$



# CHAÎNE DE MARKOV APÉRIODIQUE

Une chaîne de Markov est apériodique si pour tout état  $x$ , le PGCD des instants  $n$  tels que  $Q^n(x, x) > 0$  est égal à 1.

Une chaîne de Markov irréductible est apériodique si pour tout état  $x$ , la probabilité de ne pas changer est non nulle.

$$\forall x, \quad Q(x, x) > 0$$





# THÉORÈME (FELLER)

Toute chaîne de Markov homogène, irréductible et apériodique de transition  $Q$  admet une mesure invariante  $M$ .

$$\forall y, \quad M(y) = \sum_x M(x)Q(x, y)$$

Et la chaîne de Markov converge vers sa mesure invariante de manière indépendante des conditions initiales.

$$\forall (x, y), \quad \lim_{n \rightarrow \infty} P[x(n) = y | x(0) = x] = \lim_{n \rightarrow \infty} Q^n(x, y) = M(y)$$



# VITESSE DE CONVERGENCE

**Le rayon spectral d'une matrice est le maximum des valeurs absolues de ses valeurs propres.**

## Théorème de Perron

- Le rayon spectral d'une matrice positive est une valeur propre simple dominante.

## Théorème de Frobenius

- Le rayon spectral d'une matrice positive irréductible est une valeur propre simple dominante et admet un vecteur propre positif (toutes ses coordonnées sont positives).

**On note  $1 > |\mu_1| \geq |\mu_2| \geq \dots$  les valeurs propres de la matrice de transition  $Q$ .**

- $1$  est valeur propre de  $Q$  car  $M$  est un vecteur propre à gauche, et le vecteur unitaire à droite.

$$M = MQ \text{ et } Q\mathbb{1} = \mathbb{1}$$

- Ses valeurs propres  $\mu$  autres que  $1$  sont plus petites que  $1$  :

- Si  $Qv = \mu v$ , soit  $|v_k| = \max_i |v_i|$

$$|\mu||v_k| = |\mu v_k| = \left| \sum_j q_{k,j} v_j \right| \leq |v_k|$$



# THÉORÈME DE SENETA

$$\max_{x,y} |Q^n(x,y) - M(y)| \sim_{n \rightarrow \infty} C n^{m_1-1} |\mu_1|^n$$

où  $m_1$  est l'ordre de multiplicité de  $\mu_1$ .



# LOI FORTE DES GRANDS NOMBRES

Pour tout couple de fonctions  $f$  et  $g > 0$  intégrables (dans  $\mathcal{L}^1(\Omega)$ ), alors quelle que soit la condition initiale  $x(0)$

$$\frac{\sum_{m=0}^{n-1} f(x(m))}{\sum_{m=0}^{n-1} g(x(m))} \xrightarrow{n \rightarrow \infty} \frac{E_M[f(x)]}{E_M[g(x)]} \text{ ps}$$

En particulier si  $g = 1$

$$\frac{1}{n} \sum_{m=0}^{n-1} f(x(m)) \xrightarrow[n \rightarrow \infty]{ps} \sum_x M(x) f(x) = E_M[f(x)]$$



# THÉORÈME LIMITE CENTRALE

Soit  $f$  une fonction réelle de  $\Omega^2$  intégrable, alors quelle que soit la condition initiale

$$\frac{1}{n} \sum_{m=0}^{n-1} f(x(m), x(m+1)) \xrightarrow[n \rightarrow \infty]{ps} \sum_x M(x) Q(x, y) f(x, y) \\ = E_M[f(x, x^{+1})]$$

*Dans la dernière inégalité, le nombre +1 est symbolique, juste pour rappeler que les deux arguments sont successifs dans la chaîne.*



# TLC (CONVERGENCE NORMALE)

Si on a en plus  $f^2$  intégrable, alors

$$\frac{1}{\sqrt{n}} \sum_{m=0}^{n-1} (f(x(m), x(m+1)) - \bar{f}(x)) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} N(0, \sigma^2(f))$$

avec

$$\bar{f}(x) = \sum_y Q(x, y) f(x, y) = E_M[f(x, x^{+1})|x]$$

et

$$\begin{aligned} \sigma^2(f) &= \sum_{x,y} M(x) Q(x, y) (f(x, y) - \bar{f}(x))^2 \\ &= \text{var}_M[f(x, x^{+1})] \end{aligned}$$

**Pour une condition initiale  $x(0)$  quelconque, notons  $M^n(y|x(0)) = M Q^n(x(0), y)$ . On montre que**

$$\left| E_{M^n}[x|x(0)] - E_M(x) \right| = o\left(\frac{|\mu_1|^n}{n}\right)$$



# RÈGLE DE HEBB

1. **INTRODUCTION**
2. **APPRENTISSAGE**
3. **RÔLE DE LA TEMPÉRATURE**

→ **APPLICATIONS**

# HYPOTHÈSE SIMPLIFIÉE

On se limite au cas des réseaux binaires  $x_i \in \{0,1\}$ , symétrique à interaction d'ordre 2.

$$v_i = \sum_{j \neq i} w_{\{i,j\}} x_j \quad \text{et} \quad P[x_i(n+1) = 1 \mid x(n)] = \psi_T(v_i)$$

On construit une chaîne de Markov en décidant de ne changer à chaque étape qu'une seule cellule  $i_n$  à la fois (transitions asynchrones).

Dans ce cas on a montré que la chaîne de Markov ainsi construite converge vers une mesure limite  $M$  telle que

$$P_M(x) = \frac{1}{Z_T} e^{-\frac{H(x)}{T}} \quad \text{où} \quad H(x) = - \sum_{i \neq j} w_{\{i,j\}} x_i x_j$$

Nous avons vu plus haut que la loi limite était un champ de Gibbs, comme ce dernier est symétrique il s'exprime sur les cliques du réseau qui sont les paires  $\{i, j\}$ , dans le cas binaire symétrique l'expression s'écrit naturellement comme un produit et des sommes, mais en supposant comme depuis le début qu'on conserve une cellule constante d'activité 1, les singletons sont des produits avec 1 et les constantes sont entrées dans la constante de normalisation  $Z_T$ .





# TRANSITIONS

Il est facile de vérifier que la transition unitaire prend l'allure souhaitée :

$$\begin{aligned} P_M[x_{i_n}(n+1) = 1 | x_{-i_n}(n)] &= \frac{e^{-\frac{1}{T}H(x_{-i_n}(n) \wedge x_{i_n}=1)}}{e^{-\frac{1}{T}H(x_{-i_n}(n) \wedge x_{i_n}=1)} + e^{-\frac{1}{T}H(x_{-i_n}(n) \wedge x_{i_n}=0)}} \\ &= \frac{1}{1 + e^{-\frac{1}{T}[H(x_{-i_n}(n) \wedge x_{i_n}=0) - H(x_{-i_n}(n) \wedge x_{i_n}=1)]}} \\ &= \frac{1}{1 + e^{-\frac{v_{i_n}}{T}}} \end{aligned}$$

- Une machine de Boltzmann binaire, symétrique, séquentielle est un processus markovien binaire symétrique qui converge vers une loi d'énergie  $H(x)$ .
- En diminuant la température  $T$  la loi limite charge les configurations fondamentales de  $M$  qui sont les minimums de  $H$ .



# APPRENTISSAGE

La « Machine de Boltzmann » est aussi un outil permettant de « régler » l'énergie  $H$  pour qu'elle réponde à nos « besoins ».

Soit donc une loi  $L(x)$  représentative de l'environnement que l'on cherche à modéliser.

Par exemple si on a des cellules dites visibles, on aimerait que la loi  $M$  charge ces cellules comme la loi  $L$ .

Soit un échantillon de données  $S = \{(z_n, y_n), n = 1 \dots N\}$ .

On suppose que la loi  $L$  définie sur l'ensemble des sites ait une répartition correspondant à notre observation.

On pose  $x = \begin{pmatrix} z \\ * \\ y \end{pmatrix}$ , les cellules cachées n'étant pas définies par l'environnement.

- $z$  : représente l'entrée ou le contexte exogène.
- $y$  : la sortie ou les observations endogènes.



# NOTATIONS

## On suppose que

- les entrées  $z_n$  correspondent à des sites  $i = 1 \dots p$ ,

$$z_n \sim (x_i(n))_{i=1\dots p}$$

- puis que l'on ait un certain nombre de sites inconnus (ou cachés) que l'on peut identifier par un vecteur aléatoire correspondant aux sites  $j = p + 1 \dots p + r$ ,

$$(x_j(n))_{j=p+1\dots p+r}$$

- et finalement les cellules de sorties représentées par les vecteurs  $y_n$  modélisés par les activités des sites  $k = p + r + 1 \dots p + r + q$ .

$$y_n \sim (x_k(n))_{k=p+r+1\dots p+r+q}$$



# L'ENVIRONNEMENT ET LA MACHINE

La loi  $M$  (champ de Gibbs) de la machine de Boltzmann va charger tous les sites, mais nous sommes particulièrement intéressés à ce qui se passe du point de vue de la relation entre les sites d'entrée et les sites de sortie.

Cette relation est celle observée par  $L(y|z)$ .

- Et  $L$  est connue par

$$L(z_n, y_n) = 1/N$$

- ou d'autres hypothèses dépendant du problème étudié.



# DIVERGENCE DE KULLBACK-LEIBLER

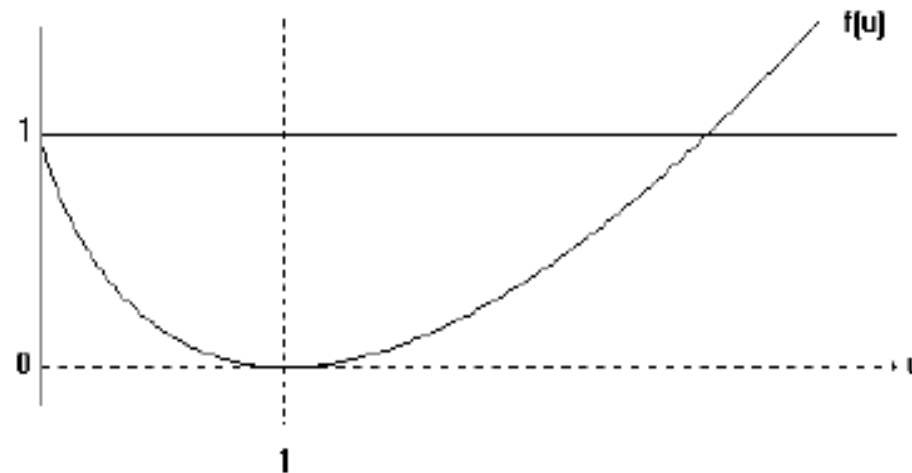
On utilise un écart entre distributions : l'information de Kullback

$$\lambda(M, L) = \sum_x L(x) \log \frac{L(x)}{M(x)}$$

C'est une fonction convexe en  $\frac{L}{M}$ , car

$$\lambda(M, L) = E_M \left[ f \left( \frac{L}{M} \right) \right]$$

avec  $\begin{cases} f(u) = u \log u + 1 - u \\ f(0) = 1 \end{cases}$



# CALCUL DU GRADIENT

$$\frac{\partial \lambda(M, L)}{\partial w} = - \sum_x \frac{L(x)}{M(x)} \frac{\partial M(x)}{\partial w}$$

et

$$\begin{aligned} \frac{\partial M(x)}{\partial w} &= - \frac{1}{T} \frac{\frac{\partial H(x)}{\partial w} e^{-\frac{1}{T} H(x)} Z - e^{-\frac{1}{T} H(x)} \sum_{x'} \frac{\partial H(x')}{\partial w} e^{-\frac{1}{T} H(x')}}{Z_T^2} \\ &= - \frac{1}{T} M(x) \left( \frac{\partial H(x)}{\partial w} - E_M \left[ \frac{\partial H}{\partial w} \right] \right) \end{aligned}$$

Soit finalement

$$\frac{\partial \lambda(M, L)}{\partial w} = \frac{1}{T} \left( E_L \left[ \frac{\partial H}{\partial w} \right] - E_M \left[ \frac{\partial H}{\partial w} \right] \right)$$



# SECOND ORDRE

De la même façon, on montre aussi que

$$\frac{\partial^2 \lambda(M, L)}{\partial w \partial w'} = \frac{1}{T^2} \text{cov}_M \left[ \frac{\partial H}{\partial w}, \frac{\partial H}{\partial w'} \right] \quad \text{dès que } \frac{\partial^2 H}{\partial w \partial w'} = 0$$

en particulier

$$\frac{\partial^2 \lambda(M, L)}{\partial w^2} = \frac{1}{T^2} \text{var}_M \left[ \frac{\partial H}{\partial w} \right] \geq 0$$



# ESTIMATION EMPIRIQUE DU GRADIENT

Dans le cadre symétrique, binaire et séquentiel choisi on a

$$\frac{\partial H(x)}{\partial w_{\{i,j\}}} = -x_i x_j$$

Ce qui fait que la loi de mise à jour des poids est définie par

$$w_{\{i,j\}}(n+1) = w_{\{i,j\}}(n) + \eta(E_L[x_i x_j] - E_M[x_i x_j])$$

Ou si l'on souhaite utiliser une formule de Newton :

$$w_{\{i,j\}}(n+1) = w_{\{i,j\}}(n) + \eta T \frac{(E_L[x_i x_j] - E_M[x_i x_j])}{\text{var}_M[x_i x_j]}$$





# RÈGLE DE HEBB

- L'estimation de l'espérance sous la loi  $M$  consiste à laisser observer des tirages issus de trajectoires réalisées par le champs de Gibbs. On a vu que la loi des grands nombres autorisait de faire une moyenne le long d'une seule trajectoire.
- L'estimation de l'espérance sous la loi  $L$  nécessite néanmoins d'avoir des observations intermédiaires pour les cellules cachées. D'après les calculs on doit fixer les cellules d'entrée et de sorties par l'environnement et d'estimer les cellules cachées par la loi  $M$  conditionnées par les observations visibles.

**Si l'on est intéressés par la modélisation de la fonction  $x \mapsto y$  alors on peut toujours laisser les entrées fixées puis réaliser des tirages avec sortie fixée (+) et sortie libre (-) et faire la différence des produits d'activités par connexion.**

$$w_{\{i,j\}}(n+1) = w_{\{i,j\}}(n) + \eta(x_i^+ x_j^+ - x_i^- x_j^-)$$



# A PROPOS DE LA TEMPÉRATURE

**La température  $T$  permet de régler le paramétrage d'un recuit simulé.**

- Pour  $T$  grand la loi est assez chaotique et l'on peut sauter facilement d'une configuration à l'autre. On assure donc un bon parcours de l'espace des états.
- Quand  $T$  est proche de 0, la loi limite est centrée sur les configurations fondamentales.

**Il est parfois pratique d'avoir une réponse déterministe à une entrée, bien que ce ne soit pas une très bonne idée quand le modèle n'est pas parfait. Pour cela on voudrait bien avoir un  $T$  petit, mais la convergence de la chaîne de Markov va être difficile car il sera très dur de sortir des minimums locaux.**

**On sent bien que ce sera plus facile de modéliser  $M$  pour  $T$  grand et l'on fera décroître la température progressivement après stabilisation.**

- Il faut savoir que dans ce cas on convergera vers 1 seul minimum global bien que plusieurs configurations fondamentales puissent exister.



# QUELLE EST LA MEILLEURS TEMPÉRATURE POUR L'APPRENTISSAGE ?

La question est logique car l'apprentissage règle l'énergie  $H$  qui ne dépend pas de la température. On peut se poser la question du gradient de l'erreur en fonction de la température.

$$\lambda = g\left(\frac{w}{T}\right) \text{ où } w \text{ est le vecteur des poids}$$

donc

$$\frac{\partial \lambda}{\partial w} = \frac{1}{T} g'\left(\frac{w}{T}\right)$$

$$\frac{\partial \lambda}{\partial T} = -\frac{1}{T^2} \left\langle w, g'\left(\frac{w}{T}\right) \right\rangle = -\frac{1}{T} \left\langle w, \frac{d\lambda}{dw} \right\rangle$$

Soit finalement

$$\frac{\partial \lambda(M, L)}{\partial T} = -\frac{1}{T} \sum_{\{i,j\}} w_{\{i,j\}} \frac{\partial \lambda(M, L)}{\partial w_{\{i,j\}}}$$



# STRATÉGIE DE NORMALISATION DES POIDS

Le calcul précédent permet éventuellement d'appliquer une règle du type

$$T(n + 1) = T(n) + \gamma \frac{1}{T} \sum_{\{i,j\}} w_{\{i,j\}}(n) (w_{\{i,j\}}(n + 1) - w_{\{i,j\}}(n))$$

*Si la valeur des poids augmentent, on aura intérêt à augmenter la valeur de la température pour compenser.*

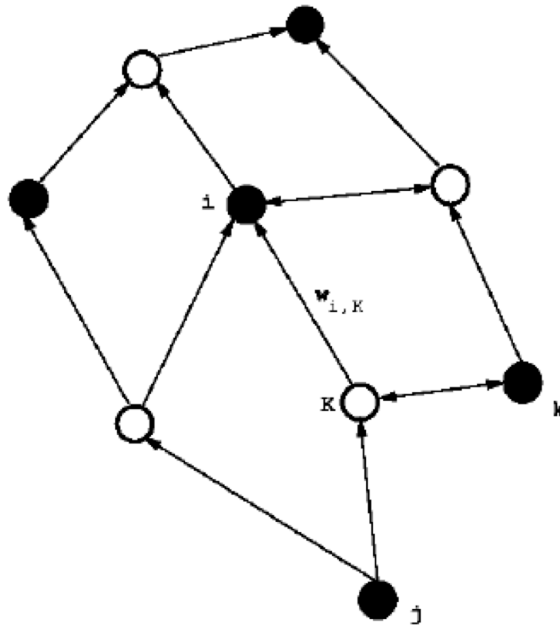
**Dans la pratique on essaye de normaliser le vecteur des poids en rajoutant une contrainte sur leur norme.**



# ANNEXE

## 1. CALCUL PARALLÈLE

# ECHANTILLONNEUR PARALLÈLE (1)



Système de voisinage sur un réseau de neurones à cliques

$$N_i = \bigcup_{w_{i,K} \neq 0} K$$

Transitions markoviennes

$$Q(x,y) = P[x^{t+1} = y \mid x^t = x] = \prod_{i \in S} \Pi_i(y_i \mid x_{N_i})$$

Spécifications locales

$$\Pi_i(y_i \mid x_{N_i}) \propto \exp\left[\frac{1}{T} \sum_K w_{i,K} J_K(x_K, y_i)\right]$$




# ECHANTILLONNEUR PARALLÈLE (2)

## Minimisation d'une erreur moyenne

$$\ell = \mathbb{E}[\lambda(x)] = \sum_{x \in \Omega} \lambda(x) M(x)$$

$$\frac{\partial \ell}{\partial w_{i,K}} = \sum_{x \in \Omega} \lambda(x) \frac{\partial M}{\partial w_{i,K}}$$

$$M = MQ \implies M' = MQ' + M'Q \quad M' = \lim_{n \rightarrow +\infty} MQ'(I + Q + \dots + Q^n)$$

$$\begin{aligned} \frac{\partial \log Q(x,y)}{\partial w_{i,K}} &= \sum_{i \in S} \frac{\partial \log \Pi_i(y_i | x_{N_i})}{\partial w_{i,K}} \\ &= \frac{\partial \log \Pi_i(y_i | x_{N_i})}{\partial w_{i,K}} \\ &= \frac{1}{T} J_K(x_K, y_i) - \frac{1}{T} \sum_{z_i \in \Phi_i} J_K(x_K, z_i) \Pi_i(z_i | x_{N_i}) \end{aligned}$$


$$\bar{J}_K(x_K) = \mathbb{E}[J_K(x_K^t, x_i^{t+1}) | x^t] \quad u_{i,K}(x,y) = J_K(x_K, y_i) - \bar{J}_K(x_K)$$





# ECHANTILLONNEUR PARALLÈLE (3)

## Minimisation d'une erreur moyenne (suite)

$$\frac{\partial Q(x,y)}{\partial w_{i,K}} = \frac{1}{T} Q(x,y) u_{i,K}(x,y)$$

$$\frac{\partial M(z)}{\partial w_K} = \frac{1}{T} \sum_{k=0}^{+\infty} \sum_x \sum_y M(x) u_{i,K}(x,y) Q(x,y) Q^k(y,z)$$

$$\sum_z \lambda(z) \frac{\partial M}{\partial w_{i,K}}(z) = \frac{1}{T} \sum_{k=0}^{+\infty} \sum_{x,y,z} M(x) u_{i,K}(x,y) Q(x,y) Q^k(y,z) \lambda(z)$$

$$\begin{aligned} \frac{\partial \ell}{\partial w_{i,K}} &= \frac{1}{T} \sum_{k=0}^{+\infty} \mathbb{E}[\lambda(x^{t+k+1}) u_{i,K}(x^t, x^{t+1})] \\ &= \frac{1}{T} \sum_{k=0}^{+\infty} \mathbb{E}[\lambda(x^t) u_{i,K}(x^{t-k-1}, x^{t-k})] \end{aligned}$$





# ECHANTILLONNEUR PARALLÈLE (4)

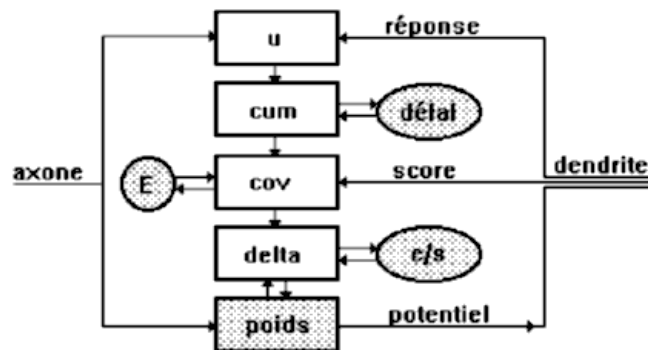
$$\frac{\partial \bar{\lambda}}{\partial w_{i,K}} \approx \frac{1}{T} E_M \left[ \lambda(x^{\tau+\nu}) \text{cum}_{i,K}^{\tau}(\nu) \right]$$

$$\text{cum}_{i,K}^{\tau}(\nu) = \sum_{t=\tau+1}^{\tau+\nu} u_{i,K}(x^{t-1}, x^t)$$

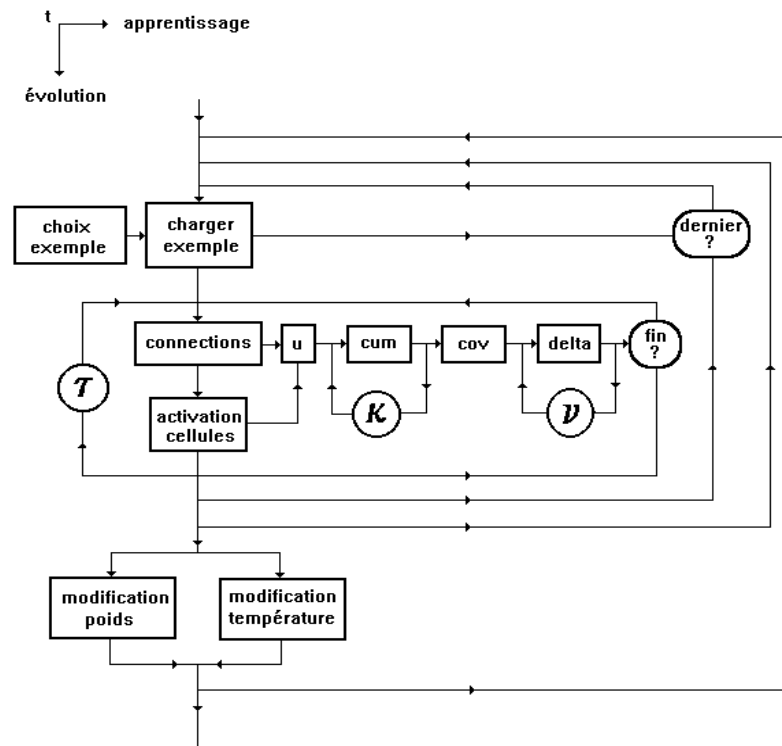
$$u_{i,K}(x, y) = J_K(x_K, y_i) - \bar{J}_{i,K}(x_K)$$

$$\bar{J}_{i,K}(x_K) = E \left[ J_K(x_K^t, x_i^{t+1}) | x^t \right]$$

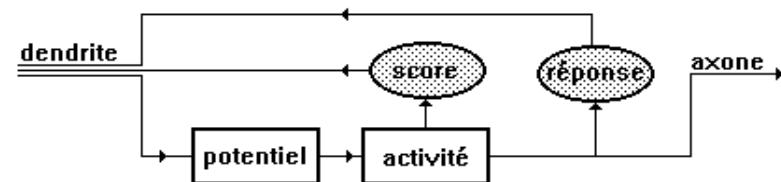
t → évolution  
apprentissage



Dynamique d'une synapse



Dynamique générale



Dynamique d'une cellule



# A SUIVRE

## EXEMPLES DE DONNÉES