Cours de probabilités discrètes

Magistère STIC

1re année — 1er semestre — 2004–2005

Table des matières

Auteurs

Dans ce document sont rassemblées des notes du cours de probabilités discrètes de Robert Azencott — et des travaux dirigés de Jérémie Jakubowicz — destinés aux étudiants de première année du magistère STIC de l'ENS de Cachan. Ces notes ont été rédigées par ces étudiants, c'est-à-dire Romain Bardou, Charles Bouillaguet, Pierre Chambart, Jean Daligault, Steven Gay et Stéphane Glondu. Des notes de bas de page détaillent la participation de chacun. Vous pouvez modifier et diffuser ce document si vous y faites apparaître ce paragraphe intact.

Cette version date du mercredi 26 janvier 2005, 08 : 05 (GMT).

1 Introduction[1]

On peut voir les probabilités sous deux aspects : la théorie de la mesure et les probabilités discrètes. Dans ce cours, on s'intéressera (presque) exclusivement aux probabilités discrètes.

1.1 Probabilités et informatique

On fait appel aux probabilités dans de multiples domaines de l'informatique. Ainsi, la compression fait appel à l'*entropie*, un concept probabiliste qui permet de produire des codages optimaux. De plus, de nombreux problèmes (notamment les fameux problèmes *NP-complets*) sont impossibles à résoudre exactement en un temps raisonnable. On est alors obligé d'utiliser des méthodes approchées dont on peut montrer la validité grâce aux probabilités (algorithmes *randomisés*). De manière plus générale, les probabilités permettent l'étude du comportement de certains algorithmes et structures de données afin de les optimiser.

1.2 Quelques définitions

1.2.1 Expériences, fréquences, probabilités

Définition 1.1. Une *expérience aléatoire* est caractérisée par l'ensemble Ω des résultats possibles, que l'on supposera toujours dénombrable. Ω est aussi appelé *espace* ou *univers probabiliste*. Un *événement* est une partie quelconque de Ω. Un événement A est *réalisé* si le résultat de l'expérience est dans A.

Par exemple, le nombre de clients connectés à un serveur à un instant donné est une expérience aléatoire (où $\Omega = \mathbb{N}$). Un exemple d'événement est $A = \{n \in \mathbb{N} \,|\, 1\,000 \leqslant n \leqslant 10\,000\}$.

Définition 1.2 (Fréquence empirique)**.** Soit $A \subset \Omega$. On répète une expérience aléatoire N fois, obtenant ainsi N résultats $\omega_1, \ldots, \omega_N$. La *fréquence empirique* de A est définie par :

$$f_N(A) = \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}_A(\omega_j) = \frac{\text{nombre de fois où } A \text{ est réalisé}}{N}.$$

On va s'intéresser intuitivement aux cas où $f_N(A) \xrightarrow[N\to\infty]{} \varphi(A)$. φ doit alors vérifier les propriétés de la définition suivante :

[1] Cours du 27 septembre, rédigé par Stéphane Glondu.

Définition 1.3 (Probabilité)**.** Une *probabilité* sur Ω est une application

$$\begin{array}{rccc} \mathbb{P} & : & \mathcal{P}(\Omega) & \to & [0,1] \\ & & A & \mapsto & \mathbb{P}[A] \end{array}$$

telle que :

1. $\mathbb{P}[\emptyset] = 0$;
2. $\mathbb{P}[\Omega] = 1$;
3. si $A \cap B = \emptyset$, alors $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B]$.

La *loi de probabilité* associée est la valeur de $\mathbb{P}$ sur chacun des singletons de Ω.

Le corollaire 2.5 formalisera l'intuition des probabilités.
Les probabilités vérifient quelques propriétés immédiates :

Proposition 1.1. *Une probabilité vérifie les propriétés suivantes :*

1. *Si $A \subset B$, alors :*
$$\mathbb{P}[A] \leqslant \mathbb{P}[B].$$
2. *Si $\Omega = \{\omega_1, \dots, \omega_n, \dots\}$ et que $\forall j$, $\mathbb{P}[\omega_j] = p_j$, alors :*
$$\sum_{j=1}^{\infty} p_j = 1.$$
3. *Si $A = \{x_1, \dots, x_n, \dots\} \subset \Omega$, alors :*
$$\mathbb{P}[A] = \sum_{j=1}^{\infty} \mathbb{P}[x_j].$$
4. $\mathbb{P}$ *est* dénombrablement additive, i.e. *si $(A_n)_{n\in\mathbb{N}^*}$ est une famille d'événements deux à deux disjoints, alors :*
$$\mathbb{P}\left[\bigcup_{j=1}^{\infty} A_j\right] = \sum_{j=1}^{\infty} \mathbb{P}[A_j].$$
5. *Si $(A_n)_{n\in\mathbb{N}^*}$ est décroissante et que $\bigcap_{j=1}^{\infty} A_j = \emptyset$, alors :*
$$\mathbb{P}[A_j] \xrightarrow[j\to\infty]{} 0.$$

1.2.2 Variables aléatoires

Définition 1.4 (Variable aléatoire)**.** Une *variable aléatoire* est une fonction $\Omega \to \mathbb{R}$.

Intuitivement, il s'agit d'un nombre — le concept est bien sûr généralisable — dont la valeur dépend du résultat d'une expérience aléatoire. Souvent, on ignorera l'expérience en question, on ne fera aucune allusion à Ω ou à l'un de ses éléments. La *loi* d'une variable aléatoire sera alors la probabilité qu'elle prenne une certaine valeur.

Cette formalisation permet de déterminer des caractéristiques intéressantes d'une loi :

Définition 1.5 (Espérance)**.** L'*espérance* d'une variable aléatoire X sur un espace Ω et à valeurs réelles est sa valeur moyenne :

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} \mathbb{P}[\omega]\, X(\omega).$$

On ne s'intéressera qu'aux cas où $\mathbb{E}[|X|] < \infty$ (*i.e.* $X \in \mathcal{L}^1$).

L'espérance vérifie quelques propriétés immédiates :

Proposition 1.2. *Soient X et Y deux variables aléatoires à valeurs réelles, et $\lambda \in \mathbb{R}$. On a :*

1. $\mathbb{E}[|X+Y|] \leqslant \mathbb{E}[|X|] + \mathbb{E}[|Y|]$;
2. $\mathbb{E}$ *est linéaire :* $\mathbb{E}[\lambda X] = \lambda \mathbb{E}[X]$;
3. $\mathbb{E}$ *est croissante : si* $X \leqslant Y$, *alors* $\mathbb{E}[X] \leqslant \mathbb{E}[Y]$.

Toutes ces propriétés — sauf la dernière — s'étendent aux variables aléatoires à valeurs complexes.

Le besoin de connaître l'écart entre une variable aléatoire et sa moyenne motive la définition suivante :

Définition 1.6 (Variance et écart-type)**.** La *variance* et l'*écart-type* d'une variable aléatoire sont définis respectivement par :

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$$
$$\text{et} \qquad \sigma(X) = \sqrt{\mathrm{Var}(X)}.$$

Remarque 1.1. Pour toute variable aléatoire X et tout réel λ, on a :

$$\mathrm{Var}(\lambda X) = \lambda^2 \mathrm{Var}(X) \qquad \text{et} \qquad \sigma(\lambda X) = |\lambda|\, \sigma(X).$$

L'écart-type donne déjà une idée de l'écart entre une variable aléatoire et sa moyenne :

Théorème 1.3 (Tchebychev)**.** *Soit X une variable aléatoire à valeurs réelles, $a > 0$, $m = \mathbb{E}[X]$ et $\sigma = \sigma(X)$. On a :*

$$\mathbb{P}[|X - m| \geqslant a] \leqslant \frac{\sigma^2}{a^2}.$$

Démonstration. Posons $Y = X - m$. On a $\mathbb{E}[Y] = 0$ et $\mathrm{Var}(Y) = \mathrm{Var}(X) = \sigma^2$. En posant $A = \{|Y| \geqslant a\}$, on a :

$$\begin{aligned}\mathrm{Var}(Y) &= \sum_{\omega \in A} \left|Y(\omega)^2\right| \times \mathbb{P}[\omega] \\ &\geqslant \sum_{\omega \in A} a^2 \, \mathbb{P}[\omega] \\ &= a^2 \, \mathbb{P}[A],\end{aligned}$$

d'où le résultat. □

On verra plus tard un résultat plus élaboré (théorème 2.7).

1.3 Exemples

On donne dans cette section quelques exemples de variables aléatoires et de lois qu'elles peuvent suivre. Ces exemples, et d'autres, seront étudiés plus en détail en TD.

1.3.1 Loi de Poisson

Cette loi modélise bien le nombre de clients $X \in \mathbb{N}$ connectés à un serveur.

Définition 1.7 (Loi de Poisson)**.** La *loi de Poisson* de paramètre $\lambda > 0$ est celle où :

$$\mathbb{P}[X = n] = \frac{\lambda^n}{n!} e^{-\lambda}.$$

1.3.2 Loi de Bernoulli et loi hypergéométrique[2]

Un lancer de pièce suit une loi de Bernoulli :

[2]Ce paragraphe a été approfondi en cours le 6 décembre et rédigé par Romain Bardou.

Définition 1.8 (Loi de Bernoulli)**.** La *loi de Bernoulli* de paramètre $p \in [0,1]$ est définie sur un univers à deux éléments $\Omega = \{0,1\}$ par :

$$\mathbb{P}[0] = 1 - p \qquad \text{et} \qquad \mathbb{P}[1] = p.$$

Concernant le lancer de pièce, on peut aussi s'intéresser au *temps d'attente* T de « pile », *i.e.* au nombre de fois qu'il faudra lancer la pièce pour obtenir « pile ». T est une variable aléatoire suivant une loi hypergéométrique :

Définition 1.9 (Loi hypergéométrique)**.** La *loi hypergéométrique* de paramètre $p \in [0,1]$ est celle où :

$$\mathbb{P}[T = k] = (1-p)^{k-1}p.$$

Dans le cas du lancer d'une pièce équilibrée, on a $\mathbb{P}[T = k] = \dfrac{1}{2^k}$.

Proposition 1.4 (Espérance d'une loi hypergéométrique)**.** *Soit X une variable aléatoire dans $\{0,1\}$ suivant une loi de Bernoulli de paramètre p (donc $\mathbb{P}[X = 1] = p$). Le temps d'attente moyen de l'événement $X = 1$ est $\frac{1}{p}$. La variance de ce temps d'attente est $\frac{1-p}{p^2}$ et son écart-type $\frac{\sqrt{1-p}}{p}$.*

Démonstration. Notons X_i la variable aléatoire donnant la valeur de X au i-ème tirage. La probabilité que l'événement voulu arrive à la tentative k et pas avant est :

$$\begin{aligned}\mathbb{P}[X_1 = 0 \text{ et } \ldots \text{ et } X_{k-1} = 0 \text{ et } X_k = 1] &= \mathbb{P}[X_1 = 0] \cdots \mathbb{P}[X_{k-1} = 0]\,\mathbb{P}[X_k = 1] \\ &= (1-p)^{k-1}p.\end{aligned}$$

Le temps d'attente moyen est donc (en notant T le temps d'attente et en

posant $z = 1 - p$) :

$$\begin{aligned}
\mathbb{E}[T] &= \sum_{k\geqslant 1} k(1-p)^{k-1}p \\
&= p\sum_{k\geqslant 1} k(1-p)^{k-1} \\
&= p\sum_{k\geqslant 1} \frac{\mathrm{d}\left(z^k\right)}{\mathrm{d}z} \\
&= p\frac{\mathrm{d}\left(\sum_{k\geqslant 1} z^k\right)}{\mathrm{d}z} \\
&= p\frac{\mathrm{d}\left(\frac{1}{1-z} - 1\right)}{\mathrm{d}z} \\
&= p\frac{1}{(1-z)^2} \\
&= \frac{1}{p}.
\end{aligned}$$

De plus, la variance est :

$$\begin{aligned}
\mathrm{Var}(T) &= \mathbb{E}\left[T^2\right] - (\mathbb{E}[T])^2 \\
&= \mathbb{E}\left[T^2\right] - \frac{1}{p^2}.
\end{aligned}$$

On peut calculer l'espérance de T^2 :

$$\begin{aligned}
\mathbb{E}\left[T^2\right] &= p\sum_{k\geqslant 1} k^2 z^{k-1} \\
&= p\frac{\mathrm{d}\left(z\frac{\mathrm{d}\left(\sum_{k\geqslant 1} z^k\right)}{\mathrm{d}z}\right)}{\mathrm{d}z} \\
&= p\frac{\mathrm{d}\left(z\frac{\mathrm{d}\left(\frac{1}{1-z} - 1\right)}{\mathrm{d}z}\right)}{\mathrm{d}z} \\
&= p\frac{\mathrm{d}\left(\frac{z}{(1-z)^2}\right)}{\mathrm{d}z} \\
&= \frac{2-p}{p^2},
\end{aligned}$$

d'où :

$$Var(T) = \frac{1-p}{p^2}. \qquad \square$$

2 Variables indépendantes et lois jointes[3]

2.1 Loi d'une variable aléatoire X

Soient X et Y deux variables aléatoires sur un espace Ω.

Définition 2.1. La *loi* de X est définie par :

$$p(x) = \mathbb{P}[X = x] = \sum_{X(\omega)=x} \mathbb{P}[\omega], \qquad \text{pour } x \in \mathbb{R}.$$

On dit aussi que p est la *loi image* de $\mathbb{P}$ par X.

En d'autres termes, c'est la probabilité des événements ω qui vérifient $X(\omega) = x$, donc la probabilité que X prenne la valeur x lors d'une expérience aléatoire. Nécessairement, on a :

$$p(x) = 0 \qquad \text{si } x \notin X(\Omega),$$
$$\text{et :} \qquad \sum_{x\in\mathbb{R}} p(x) = 1 \qquad \text{car } X \text{ prend forcément une valeur.}$$

p est en quelque sorte une probabilité sur $\mathbb{R}$ portée par un sous-ensemble dénombrable.

2.2 Loi jointe de deux variables aléatoires X et Y

Notons $p(x) = \mathbb{P}[X = x]$ et $q(y) = \mathbb{P}[Y = y]$.

Définition 2.2. La *loi jointe* de deux variables aléatoires X et Y est définie par :

$$\phi(x, y) = \mathbb{P}[X = x \text{ et } Y = y], \qquad \text{pour } x, y \in \mathbb{R}.$$

C'est la probabilité des événements ω qui vérifient simultanément $X(\omega) = x$ et $Y(\omega) = y$. On a de toute évidence les égalités suivantes :

$$\sum_{(x,y)\in\mathbb{R}^2} \phi(x, y) = 1,$$
$$\sum_{y\in\mathbb{R}} \phi(x, y) = p(x), \qquad \text{et} \qquad \sum_{x\in\mathbb{R}} \phi(x, y) = q(y).$$

[3]Cours du 4 octobre, rédigé par Charles Bouillaguet et Pierre Chambart.

Remarque 2.1. La loi jointe n'est pas déterminée par les deux lois de X et de Y séparément. Les lois p et q sont parfois appelées les lois *marginales* de X et Y.

Remarque 2.2. On peut aussi définir une variable aléatoire $Z : \Omega \to \mathbb{R}^2$ par $Z = (X, Y)$. Z est aussi appelé *vecteur aléatoire*.

2.3 Variables aléatoires indépendantes

Définition 2.3. X et Y sont des variables *indépendantes* si :

$$\forall x, y \in \Omega, \quad \mathbb{P}\left[X = x \text{ et } Y = y\right] = \mathbb{P}\left[X = x\right]\mathbb{P}\left[Y = y\right].$$

Dans ces conditions, $\phi(x, y) = p(x)q(y)$, et on a les théorèmes suivants :

Théorème 2.1. *Soient X et Y des variables aléatoires indépendantes, et f et g deux fonctions. Alors :*

$$\mathbb{E}\left[f\left(X\right)g\left(Y\right)\right] = \mathbb{E}\left[f(X)\right]\mathbb{E}\left[g(Y)\right].$$

Démonstration.

$$\begin{aligned}
\mathbb{E}\left[f\left(X\right)g\left(Y\right)\right] &= \sum_{\omega\in\Omega} f\left(X\left(\omega\right)\right)g\left(Y(\omega)\right)\mathbb{P}\left[\omega\right] \\
&= \sum_{(x,y)\in\mathbb{R}^2} f(x)g(y) \sum_{\substack{X(\omega)=x \\ Y(\omega)=y}} \mathbb{P}\left[\omega\right] \\
&= \sum_{(x,y)\in\mathbb{R}^2} f(x)g(y)\phi(x,y) \\
&= \sum_{(x,y)\in\mathbb{R}^2} f(x)g(y)p(x)q(y) \\
&= \left(\sum_{x\in\mathbb{R}} f(x)p(x)\right)\left(\sum_{y\in\mathbb{R}} g(y)q(y)\right) \\
&= \mathbb{E}\left[f(X)\right]\mathbb{E}\left[g(Y)\right]
\end{aligned}$$

□

Théorème 2.2. *Si, pour toutes fonctions f et g, $f(X)$ et $g(Y)$ sont indépendantes, alors X et Y sont indépendantes.*

Démonstration. Soient x, y. Posons $f = \mathbf{1}_{\{x\}}$ et $g = \mathbf{1}_{\{y\}}$, *i.e.* $f(x_0) = 1$ si et seulement si $x_0 \in \{x\}$. On a d'une part :

$$\begin{aligned}\mathbb{E}\left[f(X)g(Y)\right] &= \mathbb{E}\left[f(X)\right]\mathbb{E}\left[g(Y)\right] \\ &= \left(\sum_{\omega\in\Omega} \mathbf{1}_{\{x\}}(X(\omega))\mathbb{P}\left[\omega\right]\right)\left(\sum_{\omega\in\Omega} \mathbf{1}_{\{y\}}(Y(\omega))\mathbb{P}\left[\omega\right]\right) \\ &= \left(\sum_{X(\omega)=x} \mathbb{P}\left[\omega\right]\right)\left(\sum_{Y(\omega)=y} \mathbb{P}\left[\omega\right]\right) \\ &= \mathbb{P}\left[X = x\right]\mathbb{P}\left[Y = y\right],\end{aligned}$$

et d'autre part :

$$\begin{aligned}\mathbb{E}\left[f(X)g(Y)\right] &= \sum_{\omega\in\Omega} \mathbf{1}_{\{x\}}(X(\omega))\mathbf{1}_{\{y\}}(Y(\omega))\mathbb{P}\left[\omega\right] \\ &= \sum_{\substack{X(\omega)=x \\ Y(\omega)=y}} \mathbb{P}\left[\omega\right] \\ &= \mathbb{P}\left[X = x \text{ et } Y = y\right],\end{aligned}$$

donc $\mathbb{P}\left[X = x \text{ et } Y = y\right] = \mathbb{P}\left[X = x\right]\mathbb{P}\left[Y = y\right]$, ce qui traduit exactement l'indépendance de X et Y. □

2.4 Variance de $X + Y$

Soient X et Y deux variables aléatoires sur Ω. Posons $Z = X + Y$. Z est bien une variable aléatoire sur Ω. On a toujours $\mathbb{E}\left[Z\right] = \mathbb{E}\left[X\right] + \mathbb{E}\left[Y\right]$, mais, dans certains cas, on a un résultat supplémentaire :

Théorème 2.3. *Soient X et Y deux variables aléatoires indépendantes. En posant $Z = X + Y$, on a :*

$$\operatorname{Var} Z = \operatorname{Var} X + \operatorname{Var} Y.$$

Démonstration. En posant :

$$\tilde{X} = X - \mathbb{E}\left[X\right], \qquad \tilde{Y} = Y - \mathbb{E}\left[Y\right], \qquad \text{et} \qquad \tilde{Z} = \tilde{X} + \tilde{Y},$$

$\tilde{X}$ et $\tilde{Y}$ sont indépendantes. De plus,

$$\operatorname{Var} X = \operatorname{Var} \tilde{X}, \qquad \operatorname{Var} Y = \operatorname{Var} \tilde{Y}, \qquad \text{et} \qquad \operatorname{Var} Z = \operatorname{Var} \tilde{Z},$$

(en effet, la variance n'est pas modifiée par l'ajout d'une constante). Remarquons que : $\mathbb{E}\left[\tilde{X}\right] = \mathbb{E}\left[\tilde{Y}\right] = \mathbb{E}\left[\tilde{Z}\right] = 0$. On a alors :

$$\begin{aligned}
\operatorname{Var}\tilde{Z} &= \mathbb{E}\left[\left(\tilde{Z} - \mathbb{E}\left[\tilde{Z}\right]\right)^2\right] \\
&= \mathbb{E}\left[\tilde{Z}^2\right] \\
&= \mathbb{E}\left[\tilde{X}^2 + 2\tilde{X}\tilde{Y} + \tilde{Y}^2\right] \\
&= \mathbb{E}\left[\tilde{X}^2\right] + 2\mathbb{E}\left[\tilde{X}\tilde{Y}\right] + \mathbb{E}\left[\tilde{Y}^2\right] \\
&= \operatorname{Var}\tilde{X} + 2\underbrace{\mathbb{E}\left[\tilde{X}\right]\mathbb{E}\left[\tilde{Y}\right]}_{=0} + \operatorname{Var}\tilde{Y} \\
&= \operatorname{Var}\tilde{X} + \operatorname{Var}\tilde{Y}.
\end{aligned}$$

□

2.5 Loi des grands nombres

Définition 2.4. Des variables aléatoires d'une famille (X_i) sont *identiquement distribuées* selon la loi q si elles suivent toutes la même loi de probabilité q, *i.e.* si :

$$\forall i, \quad \mathbb{P}\left[X_i = x\right] = q(x).$$

Théorème 2.4 (Loi des grands nombres)**.** *Soit $X_1, X_2, \ldots, X_n, \ldots$ une suite de variables aléatoires $\Omega \to \mathbb{R}$, indépendantes et identiquement distribuées selon la loi q. Posons :*

$$\overline{X}_n = \frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right)$$

(il s'agit de la moyenne empirique*). Supposons de plus que, pour tout j, on ait :*

$$\mathbb{E}\left[X_j\right] = a < +\infty \qquad \textit{et} \qquad \operatorname{Var} X_j = \sigma^2 < +\infty.$$

Alors :

1. $\overline{X}_n \xrightarrow[n\to+\infty]{} a$;
2. $\forall \epsilon > 0, \quad \mathbb{P}\left[\left|\overline{X}_n - a\right| > \epsilon\right] \xrightarrow[n\to+\infty]{} 0.$

Moralement, le premier point signifie que, conformément à l'intuition, quand on répète une expérience aléatoire un grand nombre de fois, la moyenne de X tend vers $\mathbb{E}\left[X\right]$. Le deuxième point (qui implique que la variance de $\overline{X}_n$ tend vers 0), signifie que, en plus de converger vers $\mathbb{E}\left[X\right]$, $\overline{X}_n$ ne fluctue pas indéfiniment autour de sa moyenne, mais, au contraire, s'en rapproche infiniment.

Démonstration. Premier point

$$\begin{aligned}\mathbb{E}\left[\overline{X}_n\right] &= \frac{1}{n}\left(\underbrace{\mathbb{E}\left[X_1\right]}_{a} + \cdots + \underbrace{\mathbb{E}\left[X_n\right]}_{a}\right)\\ &= \frac{1}{n}\left(na\right)\\ &= a\end{aligned}$$

Second point

$$\begin{aligned}\operatorname{Var}\overline{X}_n &= \frac{1}{n^2}\operatorname{Var}\left(\overbrace{X_1 + \cdots + X_n}^{\text{indépendantes}}\right)\\ &= \frac{1}{n^2}\left(\underbrace{\operatorname{Var} X_1}_{\sigma^2} + \cdots + \underbrace{\operatorname{Var} X_n}_{\sigma^2}\right)\\ &= \frac{1}{n^2}\left(n\sigma^2\right)\\ &= \frac{\sigma^2}{n}\end{aligned}$$

D'après le théorème de Tchebychev, on a donc :

$$\begin{aligned}\mathbb{P}\left[\left|\overline{X}_n - a\right| \geqslant \epsilon\right] &\leqslant \frac{\operatorname{Var}\overline{X}_n}{\epsilon^2}\\ &\leqslant \frac{\sigma^2}{n\epsilon^2} \xrightarrow[n\to+\infty]{} 0.\end{aligned}$$

□

Remarque 2.3. On verra en TD un résultat plus fort : si on note $A_n = \left\{\left|\overline{X}_n - a\right| > \epsilon\right\}$, alors il existe $N \in \mathbb{N}$ tel que A_n ne soit plus jamais réalisé pour $n \geqslant N$ (conséquence du théorème de Borel-Cantelli).

Corollaire 2.5. *On répète n fois une expérience. Soit $A \subset \Omega$. Alors la fréquence empirique de A tend vers sa probabilité :*

$$f_n(A) = \frac{\textit{nombre de fois où } A \textit{ est réalisé}}{n} \xrightarrow[n\to+\infty]{} \mathbb{P}\left[A\right].$$

Démonstration. En notant $\omega_1, \ldots, \omega_n$ les résultats successifs, et en appliquant la loi des grands nombres aux variables aléatoires $X_i = \mathbf{1}_A(\omega_i)$, on a le résultat. □

2.6 Transformée de Fourier d'une loi de probabilité[4]

Soit X une variable aléatoire sur un univers $\Omega = \mathbb{R}$ tel que $X(\Omega)$ soit dénombrable. Soit μ la loi de X ($\mathbb{P}[X = x] = \mu(x)$). Posons alors $Z = e^{itX}$, où $i^2 = -1$ et $t \in \mathbb{R}$. $Z : \Omega \to \mathbb{C}$ est une variable aléatoire.

Définition 2.5. La *transformée de Fourier* $\hat{\mu}$ de μ est définie par :

$$\begin{aligned}\hat{\mu}(t) &= \mathbb{E}\left[e^{itX}\right] \\ &= \mathbb{E}\left[\cos(tX) + i\sin(tX)\right] \\ &= \sum_{x\in\mathbb{R}} \cos(tx)\mu(x) + i\sum_{x\in\mathbb{R}} \sin(tx)\mu(x).\end{aligned}$$

Proposition 2.6. *Pour toutes lois de probabilité θ et μ, on a :*

$$\hat{\theta} = \hat{\mu} \iff \theta = \mu.$$

Cela signifie que l'on peut caractériser une loi de probabilité par sa transformée de Fourier.

Remarque 2.4. Soient $X_1, \ldots, X_n$ n variables aléatoires indépendantes de lois respectives $\mu_1, \ldots, \mu_n$, et $S_n = X_1 + \cdots + X_n$. Soit ν la loi de probabilité de S_n. Alors :

$$\forall t \in \mathbb{R}, \quad \hat{\nu}(t) = \hat{\mu}_1(t)\,\hat{\mu}_2(t)\cdots\hat{\mu}_n(t).$$

Cela provient de la propriété de l'espérance d'un produit de fonctions de variables aléatoires indépendantes, et du fait que $e^{a+b} = e^a e^b$. En particulier, si $\forall j, \mu_j = \mu$ (les variables suivent la même loi), alors $\hat{\nu}(t) = \hat{\mu}(t)^n$.

2.7 Théorème de la limite centrale

Le théorème suivant affine la loi des grands nombres :

Théorème 2.7 (Théorème de la limite centrale)**.** *Soit $(X_j)_{j\geqslant 1}$ une suite de variables aléatoires indépendantes de même loi μ et de variance σ^2. On note :*

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}.$$

On a :

1. $\overline{X}_n \xrightarrow[n\to+\infty]{} m = \mathbb{E}[X_1] = \sum_x x\,\mu(x)$,
2. $\mathbb{P}\left[a \leqslant \sqrt{n}\left(\dfrac{\overline{X}_n - m}{\sigma}\right) \leqslant b\right] \xrightarrow[n\to+\infty]{} \dfrac{1}{\sqrt{2\pi}}\displaystyle\int_a^b e^{-\frac{x^2}{2}}\mathrm{d}x.$

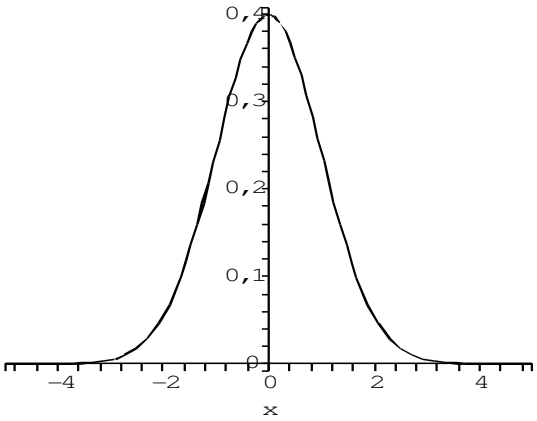

FIG. 1 – Graphe de la fonction $x \mapsto \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}$

La fonction $\Phi : u \mapsto \frac{1}{\sqrt{2\pi}} \int_{-u}^{u} e^{-\frac{x^2}{2}} \mathrm{d}x$ est tabulée, et on a :

$$\Phi(1,96) \approx 0,95$$
$$\Phi(2,58) \approx 0,99$$
$$\lim_{+\infty} \Phi = 1.$$

En général, si $n \geqslant 12$, on est déjà très proche de la limite.

Démonstration. On va juste donner quelques grandes lignes. Supposons pour simplifier que $m = 0$ et que $\sigma = 1$. On va utiliser les transformées de Fourier. Posons $U_n = \frac{X_1 + \cdots + X_n}{\sqrt{n}}$. D'après la remarque 2.4, on a :

$$\mathbb{E}\left[e^{it\sqrt{n}U_n}\right] = \left(\hat{\mu}(t)\right)^n,$$
$$\text{donc} \qquad \hat{\varphi}_n(t) = \mathbb{E}\left[e^{itU_n}\right] = \left(\hat{\mu}\left(\frac{t}{\sqrt{n}}\right)\right)^n.$$

À t fixé, $\frac{t}{\sqrt{n}} \xrightarrow[n\to+\infty]{} 0$, et, lorsque $t \to 0$, on a :

$$\hat{\mu}(t) = \hat{\mu}(0) + t\,\hat{\mu}'(0) + \frac{t^2}{2}\,\hat{\mu}''(0) + \cdots$$

[4]Cours du 18 octobre, rédigé par Romain Bardou.

Or :

$$\begin{aligned}\hat{\mu}(0) &= 1\\ \hat{\mu}'(0) &= i\mathbb{E}\left[X_1\right] = 0\\ \hat{\mu}''(0) &= -\mathbb{E}\left[X_1^2\right] = -1,\end{aligned}$$

d'où :

$$\hat{\varphi}_n(t) \approx \left(1 - \frac{t^2}{n}\right)^n \xrightarrow[n\to+\infty]{} e^{-t^2}.$$

Un analogue continu de la proposition 2.6 et le résultat

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} e^{itx} \mathrm{d}x = e^{-\frac{t^2}{2}}$$

permettent de conclure. □

3 Probabilités conditionnelles[5]

Considérons un espace Ω, ainsi que deux événements A et B ($A, B \subset \Omega$). On répète n fois une expérience aléatoire. On s'intéresse seulement aux situations où B est réalisé. Dans ces cas, quelle est la fréquence de A ?

C'est, en notant N_B le nombre de réalisations de B et $N_{A\cap B}$ le nombre de réalisations de A et B simultanément :

$$\frac{N_{A\cap B}}{N_B} = \frac{\frac{N_{A\cap B}}{n}}{\frac{N_B}{n}} \xrightarrow[n\to+\infty]{} \frac{\mathbb{P}\left[A \cap B\right]}{\mathbb{P}\left[B\right]}.$$

On définit donc :

Définition 3.1. La *probabilité conditionnelle de A sachant B* est :

$$\mathbb{P}\left[A \,|\, B\right] = \frac{\mathbb{P}\left[A \cap B\right]}{\mathbb{P}\left[B\right]}.$$

3.1 Événements indépendants

Si la réalisation de B n'apporte pas d'information sur A, on devrait alors avoir :

$$\mathbb{P}\left[A \,|\, B\right] = \mathbb{P}\left[A\right]$$

(A se réalise à sa fréquence « normale » sans que la connaissance de B change quoi que ce soit). Cela motive la définition suivante :

[5]Cours du 4 octobre, rédigé par Charles Bouillaguet et Pierre Chambart.

Définition 3.2. Deux événements A et B sont *indépendants* si :

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B].$$

Remarque 3.1. Intuitivement, deux événements sont indépendants si la réalisation de l'un n'apporte pas d'information sur la réalisation de l'autre. On peut remarquer que :

$$\begin{aligned}\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B] &\implies \mathbb{P}[A] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} \\ &\implies \mathbb{P}[A] = \mathbb{P}[A \mid B].\end{aligned}$$

La définition de l'indépendance de deux événements est donc bien conforme à l'intuition.

Par exemple, avec deux variables aléatoires X et Y indépendantes, et les deux événements $A = \{X = x\}$ et $B = \{Y = y\}$, on a :

$$\begin{aligned}\mathbb{P}[A \cap B] &= \mathbb{P}[X = x \text{ et } Y = y] \\ &= \mathbb{P}[X = x]\,\mathbb{P}[Y = y] \\ &= \mathbb{P}[A]\,\mathbb{P}[B],\end{aligned}$$

donc A et B sont deux événements indépendants.

3.2 Loi conditionnelle de X sachant Y

La loi conditionnelle de X sachant Y est la probabilité d'observer $X = x$ sachant que $Y = y$. Il s'agit de :

$$\mathbb{P}\left[\underbrace{X = x}_{A} \,\middle|\, \underbrace{Y = y}_{B}\right] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\phi(x, y)}{q(y)}$$

Remarque 3.2. On voit alors que la loi jointe de X et Y est entièrement déterminée par la loi de Y et la loi conditionnelle de X sachant Y.

4 Chaînes de Markov sur espace d'états fini[6]

4.1 Automates aléatoires

On observe l'évolution d'un automate qui prend un nombre fini d'états que l'on identifiera à $F = \{1, 2, \ldots, r\}$. L'évolution se fait en *temps discret*,

[6]Cours du 11 octobre, rédigé par Steven Gay et Stéphane Glondu.

i.e. on regarde l'état de l'automate aux instants $T \in \mathbb{N}$. À chacun de ces instants, l'automate est dans un état $X_0, X_1, \ldots, X_n, \ldots$ avec $\forall j,\ X_j \in F$. On suppose la séquence $X_0, X_1, \ldots, X_n, \ldots$ aléatoire $(X_n : \Omega \to F)$.

On s'intéresse aux *transitions*, *i.e.* quand on passe de i à j en un seul coup. Soit $q_{i,j} = \mathbb{P}[X_{k+1} = j \mid X_k = i]$. On suppose que l'évolution aléatoire est *stationnaire* dans le temps, *i.e.* $q_{i,j}$ ne dépend pas de k. La *matrice de transition* $Q = (q_{i,j})$ vérifie les propriétés suivantes :

$$\forall i \in F, \quad \sum_{j=1}^{r} q_{i,j} = 1 \qquad \text{et} \qquad \forall i, j \in F, \quad 0 \leqslant q_{i,j} \leqslant 1.$$

Une telle matrice est dite *stochastique*. C'est un automate à *mémoire courte* :

$$\mathbb{P}[X_{k+1} = j \mid X_k = i, X_{k-1} = i_1, \ldots, X_{k-m} = i_m] = \mathbb{P}[X_{k+1} = j \mid X_k = i].$$

Définition 4.1. La propriété de mémoire courte s'appelle aussi la *propriété de Markov*. On dit alors que (X_n) est une *chaîne de Markov* (ou *processus de Markov*).

4.2 Étude des chaînes de Markov

On suppose X_0 connu. Quelle est la loi de probabilité de X_n ? Que se passe-t-il lorsque $n \to +\infty$?

4.2.1 Cas particulier

Dans le cas le plus simple, les $q_{i,j}$ ne dépendent pas de i. Ainsi, pour tous i et j,

$$\begin{aligned} q_{i,j} &= \mathbb{P}[X_{k+1} = j | X_k = i] \\ &= \mathbb{P}[X_{k+1} = j] \end{aligned}$$

Cela correspond à des observations indépendantes. Les X_k ont même loi de probabilité :

$$\forall k, \quad \mathbb{P}[X_k = j] = \mu_j.$$

4.2.2 Cas général

Théorème 4.1. *Soit (X_n) une chaîne de Markov, et θ la loi de X_0 :*

$$\mathbb{P}[X_0 = j] = \theta_j.$$

Avec les notations :

$$\begin{aligned}\mu_n(j) &= \mathbb{P}\left[X_n = j\right],\\ \vec{\mu}_n &= (\mu_n(1), \mu_n(2), \ldots, \mu_n(r)),\\ \vec{\mu}_0 &= (\theta_1, \theta_2, \ldots, \theta_r),\end{aligned}$$

on a :

$$\forall n \in \mathbb{N}, \quad \vec{\mu}_n = \vec{\mu}_0 \times Q^n.$$

Démonstration. On procède par récurrence sur n. Pour $n = 0$, c'est évident. Supposons le résultat vérifié pour n.

$$\begin{aligned}\mu_{n+1}(j) &= \mathbb{P}\left[X_{n+1} = j\right]\\ &= \sum_{i=1}^{r} \mathbb{P}\left[X_{n+1} = j \text{ et } X_n = i\right]\\ &= \sum_{i=1}^{r} \mathbb{P}\left[X_{n+1} = j \,|\, X_n = i\right] \mathbb{P}\left[X_n = i\right]\\ &= \sum_{i=1}^{r} q_{i,j}\, \mu_n(i).\end{aligned}$$

On a donc finalement :

$$\vec{\mu}_{n+1} = \vec{\mu}_n \times Q.$$ □

4.2.3 Loi limite

On s'intéresse au cas $\vec{\mu}_n \xrightarrow[n\to+\infty]{} \vec{\nu}$. Dans ce cas, $\vec{\nu}$ est unique ($\vec{\nu}$ ne dépend pas de $\vec{\mu}_0$). $\vec{\nu}$ ne dépend que de Q et doit vérifier l'équation :

$$\vec{\nu} = \vec{\nu} \times Q.$$

$\vec{\nu}$ est donc un vecteur propre à gauche pour la valeur propre 1. L'équation ci-dessus est en fait un système de r équations linéaires à r inconnues.

L'existence d'une telle loi limite est assurée lorsque l'on fait l'hypothèse d'ergodicité :

Définition 4.2. Une chaîne de matrice Q est dite *ergodique* si pour tous états i et j, il existe un entier q et une chaîne $i_1, i_2, \ldots, i_q$ tels que :

$$i \to i_1 \to i_2 \to \cdots \to i_q \to j,$$

où $q_{i,i_1} > 0$, $q_{i_1,i_2} > 0, \ldots,$ $q_{i_q,j} > 0$. On peut vérifier que cela équivaut à dire qu'il existe un entier $s \geqslant 0$ tel que Q^s ait tous ses coefficients strictement positifs.

Théorème 4.2. *Soit Q la matrice de transition d'une chaîne de Markov ergodique. Il existe un unique vecteur $\vec{\theta}$ tel que $\vec{\mu}_n \xrightarrow[n\to+\infty]{} \vec{\theta}$.*

Démonstration. Allégeons les notations en notant α au lieu de $\vec{\alpha}$. Notons S l'ensemble des lois de probabilité sur F. C'est un fermé de $\mathbb{R}^F$, qui est complet. L'ergodicité nous permet de supposer sans perte de généralité que $Q > 0$. Posons $a = \min Q \in]0, 1[$.

Soient $\mu, \nu \in S$, et posons $\theta = \mu - \nu$. La somme des coordonnées de θ est nulle. Adoptons les notations suivantes :

$$\theta^+ = \{i \in F \,|\, \theta(i) \geqslant 0\} \qquad \text{et} \qquad \theta^- = \{i \in F \,|\, \theta(i) < 0\},$$

et considérons la norme :

$$\|\theta\| = \sum_{i\in F} |\theta(i)| = \sum_{i\in\theta^+} \theta(i) - \sum_{i\in\theta^-} \theta(i).$$

Posons $\hat{\theta} = \theta \times Q$. On a :

$$\hat{\theta}(j) = \sum_{i\in F} \theta(i)\, q_{i,j} = \underbrace{\sum_{i\in\theta^+} \theta(i)\, q_{i,j}}_{\geqslant 0} + \underbrace{\sum_{i\in\theta^-} \theta(i)\, q_{i,j}}_{\leqslant 0}.$$

Si $\hat{\theta} \neq 0$, la nullité de la somme des coordonnées de $\hat{\theta}$ nous permet de dire que $\hat{\theta}^+$ et $\hat{\theta}^-$ sont des parties strictes de F. Dans ce cas, on a d'une part :

$$\begin{aligned}\sum_{j\in\hat{\theta}^+} \hat{\theta}(j) &\leqslant \sum_{j\in\hat{\theta}^+} \sum_{i\in\theta^+} \theta(i)\, q_{i,j} = \sum_{i\in\theta^+} \theta(i) \sum_{j\in\hat{\theta}^+} q_{i,j} \\ &\leqslant (1-a) \sum_{i\in\theta^+} \theta(i),\end{aligned}$$

et d'autre part :

$$\begin{aligned}\sum_{j\in\hat{\theta}^-} \hat{\theta}(j) &\geqslant \sum_{j\in\hat{\theta}^-} \sum_{i\in\theta^-} \theta(i)\, q_{i,j} = \sum_{i\in\theta^-} \theta(i) \sum_{j\in\hat{\theta}^-} q_{i,j} \\ &\geqslant (1-a) \sum_{i\in\theta^-} \theta(i),\end{aligned}$$

et en soustrayant les deux inégalités, on obtient :

$$\begin{aligned}\left\|\hat{\theta}\right\| = \sum_{j\in\hat{\theta}^+} \hat{\theta}(j) - \sum_{j\in\hat{\theta}^-} \hat{\theta}(j) &\leqslant (1-a) \left(\sum_{i\in\theta^+} \theta(i) - \sum_{i\in\theta^-} \theta(i) \right) \\ &\leqslant (1-a)\, \|\theta\|,\end{aligned}$$

inégalité encore vraie pour $\hat{\theta} = 0$.

On a donc $\|\mu \times Q - \nu \times Q\| \leqslant (1-a)\,\|\mu - \nu\|$. Par conséquent, l'application $\mu \mapsto \mu \times Q$ est contractante et admet un unique point fixe dans S vers lequel converge (μ_n). □

Définition 4.3. L'unique vecteur introduit précédemment est appelé *loi limite* de la chaîne.

Remarque 4.1. Dans le cas où la chaîne n'est pas ergodique, on peut avoir plusieurs, voire une infinité de « lois limites ».

4.2.4 Temps d'atteinte et de premier retour

Définition 4.4. Soit (X_n) une chaîne de Markov.

On appelle *temps d'atteinte* (de i), et on note T_i, le premier instant $n \geqslant 0$ tel que $X_n = i$. On peut alors s'intéresser à la quantité $\mathbb{P}\left[T_i = n \,|\, X_0 = j\right]$.

On appelle *temps de premier retour* (à i), et on note S_i, le premier instant $n \geqslant 1$ tel que $X_n = i$ sachant que $X_0 = i$.

On notera que les T_i et les S_i sont finis pour une chaîne ergodique.

Ces notions seront approfondies en TD.

4.2.5 Un exemple : marche aléatoire sur $\mathbb{Z}$

On considère un automate qui effectue une marche aléatoire sur $\mathbb{Z}$, en partant de z_0. La variable X_n est la position (ou la fortune, etc.) au temps n. Soit $p \in]0,1[$. On définit un automate aléatoire comme suit :

$$\begin{aligned} \forall z \in \mathbb{Z}, \qquad & \mathbb{P}\left[X_{n+1} = z+1 \,|\, X_n = z\right] = p, \\ & \mathbb{P}\left[X_{n+1} = z-1 \,|\, X_n = z\right] = 1-p. \end{aligned}$$

(X_n) est un exemple de chaîne de Markov.

4.2.6 Un exemple plus détaillé : marche aléatoire sur un graphe[7]

Soit G un graphe non orienté à N sommets et A arêtes. On $d(g)$ le degré d'un sommet g, et V_g l'ensemble des sommets adjacents à g. On définit une chaîne de Markov $X_n : \Omega \to G$ par :

$$\mathbb{P}\left[X_{n+1} = h \,|\, X_n = g\right] = \begin{cases} \frac{1}{d(g)} & \text{si } h \in V_g, \\ 0 & \text{sinon.} \end{cases}$$

Soit Q la matrice de transition. C'est donc une matrice $N \times N$. Elle est ergodique dès que G est connexe.

[7] Cours du 13 décembre, rédigé par Stéphane Glondu.

Théorème 4.3. *La loi limite μ de X_n est la loi définie par :*

$$\forall g, \qquad \mu(g) = \frac{d(g)}{2A}.$$

Démonstration. En effet,

$$\begin{aligned}\sum_g \mu(g)Q(g,h) &= \sum_g \frac{d(g)}{2A} \times \mathbf{1}_{V_g}(h) \times \frac{1}{d(g)} \\ &= \frac{1}{2A} \left|\{g \,|\, h \in V_g\}\right| \\ &= \frac{1}{2A} \left|\{g \,|\, g \in V_h\}\right| = \mu(h)\end{aligned}$$ □

4.3 Application : mélange de cartes[8]

Le problème fut étudié par Donoho — qui était prestidigitateur de 14 à 21 ans — pour le compte de casinos.

4.3.1 Énoncé du problème

Le problème est de savoir si les croupiers mélangent correctement les cartes ou non. On formalise le problème de la façon suivante : G est le groupe des permutations de cartes, $X_1, X_2, \ldots, X_n$ sont des variables aléatoires $\Omega \to G = \mathfrak{S}_{52}$, et $g_0 \in G$ est l'ordre initial du paquet de cartes. Pour tout n, on note $g_{n+1} = X_{n+1} \circ g_n$.

Les X_i représentent les différentes permutations effectuées par le croupier. Elles sont toutes distribuées selon la même loi μ. On suppose les X_i indépendantes. Nous allons montrer que :

1. les g_i forment une chaîne de Markov ergodique ;
2. sa loi limite θ est la loi uniforme sur G (ce qui montrera que le mélange est correct, *i.e.* il ne favorise pas certaines permutations plutôt que d'autres).

En fait, à partir de $n = 5$, on est déjà très proche de la loi uniforme, *i.e.* mélanger 5 fois les cartes est suffisant pour bien les mélanger (ce résultat sera admis).

[8] Cours du 18 octobre, rédigé par Romain Bardou.

4.3.2 Résolution

Le premier point est généralisable et sera vu en TD. Nous n'allons traiter ici que le second point, *i.e.* nous allons montrer que la loi limite θ est la loi uniforme.

Démonstration. On a :

$$\begin{aligned} g_{n+1} = X_{n+1} \circ g_n &\implies X_{n+1} = g_{n+1} \circ g_n^{-1} \\ &\implies \mathbb{P}\left[g_{n+1} = j \,|\, g_n = i\right] = \mathbb{P}\left[X_{n+1} = j \circ i^{-1} \,\middle|\, g_n = i\right]. \end{aligned}$$

Étant donné que les $(g_k)_{k\geqslant 0}$ ne dépendent que des $(X_k)_{k\geqslant 1}$, et que les $(X_k)_{k\geqslant 1}$ sont indépendantes, X_{n+1} et g_n sont indépendantes et on a :

$$\begin{aligned} \mathbb{P}\left[g_{n+1} = j \,|\, g_n = i\right] &= \mathbb{P}\left[X_{n+1} = j \circ i^{-1}\right] \\ &= \mu\left(j \circ i^{-1}\right) \\ &= q_{i,j}, \end{aligned}$$

où $Q = (q_{i,j})$ désigne la matrice de transition de la chaîne. De plus, on peut montrer qu'il y a invariance à droite. En effet :

$$\begin{aligned} \forall h \in G, \qquad (j \circ h) \circ (i \circ h)^{-1} &= (j \circ h) \circ (h^{-1} \circ i^{-1}) \\ &= j \circ i^{-1}, \end{aligned}$$

d'où :

$$\forall h \in G, \qquad q_{i,j} = q_{i\circ h, j\circ h}.$$

Posons, pour $h \in G$, $\theta_h : i \mapsto \theta(i \circ h)$. θ_h est une loi de probabilité. D'après l'invariance à droite que l'on a remarqué précédemment, on a :

$$\theta \times Q = \theta \implies \theta_h \times Q = \theta_h.$$

On en déduit, de par l'unicité de la loi limite θ, que :

$$\forall h, i \in G, \qquad \theta_h(i) = \theta(i \circ h) = \theta(i),$$

d'où :

$$\forall i, j \in G, \qquad \theta(i) = \theta(j)$$

θ est donc bien la loi uniforme. □

5 Codages optimaux et entropie[9]

5.1 Généralités sur le codage

On considère des variables aléatoires $X_1, \ldots, X_n : \Omega \to K = \{x_1, \ldots, x_r\}$ de même loi de probabilité : $\forall j$, $\mathbb{P}[X = x_j] = p_j$. On veut coder le message $X_1 \cdots X_n$ afin de le transmettre. On utilise un alphabet A de cardinal a. Un *codage* est une fonction $C : K \to A^*$ (où A^* désigne l'ensemble des mots de longueur finie sur A). On note $C(x_j) = m_j = a_{1,j} \cdots a_{l_j,j}$, avec $l_j = |m_j|$. On pose alors naturellement $C(X_1 \cdots X_n) = C(X_1) \cdots C(X_n)$.

Définition 5.1 (Code injectif instantané)**.** Pour pouvoir déchiffrer le message, il faut que le code soit sans ambiguïté, c'est-à-dire injectif et sans préfixe, *i.e.* il n'existe pas de i et de j distincts tels que $m_i = m_j M$, avec $M \in A^*$. Un tel code est dit *injectif instantané*.

On veut aussi que le codage soit le « moins long » possible. On définit pour cela le *coût moyen de transmission* par :

$$\overline{L}_n = \frac{|C(X_1)| + \cdots + |C(X_n)|}{n}.$$

D'après la loi des grands nombres, on a $\overline{L}_n \xrightarrow[n\to+\infty]{} \mathbb{E}[|C(X)|]$. En effet, les X_i étant indépendants, les $|C(X_i)|$ le sont aussi. On note :

$$L = \mathbb{E}[|C(X)|] = \sum_{i=1}^{r} l_i \, p_i.$$

On veut trouver le minimum de L, C étant un codage injectif instantané.

5.2 Méthode des multiplicateurs de Lagrange

On veut minimiser une fonction $f : \mathbb{R}^r \to \mathbb{R}$ du vecteur v avec une contrainte. Pour que f atteigne un extremum en v_0, il faut que toutes les dérivées partielles de f par rapport à v soient nulles en v_0 (mais cette condition n'est pas suffisante !). On met la contrainte sous la forme $g(v) = 0$. f et g sont supposées différentiables.

On calcule $f - \lambda g$, λ étant une constante à déterminer, et on détermine les extrema de $f - \lambda g$, sans contrainte. On résout donc le système :

$$\frac{\partial f}{\partial v_j} - \lambda \frac{\partial g}{\partial v_j} = 0,$$

[9] Cours du 8 novembre, rédigé par Jean Daligault (et Stéphane Glondu pour la version LaTeX).

ce qui conduit à des solutions $v = \varphi(\lambda)$. On peut alors déterminer λ à partir de la contrainte $g(\varphi(\lambda)) = 0$.

Cette méthode est facilement généralisable à un nombre quelconque (fini) de contraintes.

5.3 Une application au codage : l'entropie

Revenons au problème du codage.

Définition 5.2 (Entropie)**.** On définit l'*entropie* $H(p)$ de la loi de probabilité p par :

$$H(p) = -\sum_{i=1}^{r} p_i \log p_i.$$

On va montrer que l'entropie caractérise L. Ainsi, si $|A| = 2$, alors $H(p)$ représente le nombre minimum de bits nécessaires au codage.

Remarquons tout d'abord un lemme utile :

Lemme 5.1. *H est concave.*

Démonstration. On rappelle que, par définition, A est *définie négative* si $\forall u \neq 0$, $u^{\mathrm{T}} A u < 0$ (où u^{T} désigne la transposée du vecteur colonne u), et que le *hessien* d'une fonction f du vecteur v est défini par :

$$\mathrm{Hessien}(f) = \left[\frac{\partial^2 f}{\partial v_i \partial v_j}\right]_{1 \leqslant i,j \leqslant r}.$$

Dans le cas de l'entropie, $\mathrm{Hessien}(H)$ est diagonale car :

$$\frac{\partial^2 H}{\partial p_i \partial p_j} = \begin{cases} -\dfrac{1}{p_i} & \text{si } i = j, \\ 0 & \text{sinon.} \end{cases}$$

$\mathrm{Hessien}(H)$ est donc une matrice diagonale à coefficients strictement négatifs, donc c'est une matrice définie négative, donc H est concave. □

Montrons maintenant le résultat suivant, qui donne un encadrement de l'entropie :

Proposition 5.2. *$H(p)$ est maximale — et vaut $\log r$ — lorsque p est la loi uniforme. Ainsi, on a toujours :*

$$0 \leqslant H(p) \leqslant \log r.$$

Démonstration. On va appliquer la méthode des multiplicateurs de Lagrange. Ici, $g = \sum_i p_i - 1$. On a :

$$\frac{\partial H(p)}{\partial p_i} = \frac{\partial(-p_i \log p_i)}{\partial p_i} = -1 - \log p_i$$
$$-\lambda \frac{\partial g}{\partial p_i} = -\lambda.$$

À i fixé, le cas intéressant est donc $-\log p_i = \lambda + 1$, soit $p_i = e^{-(\lambda+1)}$. Cette expression est indépendante de i, donc p est la loi uniforme. On en déduit que la loi de probabilité qui rend H extrémale est la loi uniforme, et c'est un maximum d'après le lemme. □

Ainsi, intuitivement, l'entropie représente le « désordre » dans le message :
- si p est concentrée, alors $H(p) \to 0$;
- si p est dispersée, alors $H(p) \to \log r$.

5.4 Codage minimal

On imagine bien que l'on ne peut pas trop compresser sans perte de données. Le théorème suivant formalise cette intuition :

Théorème 5.3 (Kraft)**.** *Soit $C : K \to A^*$ un codage injectif instantané. On note $a = |A|$. Pour tout $x_i \in K = \{x_1, \ldots, x_r\}$, on note :*

$$m_i = C(x_i), \qquad \text{et} \qquad l_i = |m_i| .$$

On a :

$$\sum_{i=1}^{r} a^{-l_i} \leqslant 1.$$

Démonstration. On construit l'arbre de tous les mots possibles sur A de longueur $l_{\max} = \max_i l_i$. Pour chaque mot m_i, on note P_i le paquet de mots dont m_i est préfixe, *i.e.* le sous-arbre enraciné en m_i. Ces paquets sont disjoints et de cardinal $a^{l_{\max} - l_i}$ (on ne s'intéresse qu'aux feuilles). L'union de tous ces paquets P_i est plus petite que l'arbre total, donc :

$$\sum_{i=1}^{r} a^{l_{\max} - l_i} \leqslant a^{l_{\max}},$$

d'où le résultat. □

Le codage qui nous intéresse donc consiste à diminuer le plus possible les l_i tout en gardant la condition $\sum_{i=1}^{r} a^{-l_i} \leqslant 1$. De plus, si $l_1, \ldots, l_r$ vérifient Kraft, alors il existe un codage injectif instantané qui utilise ces longueurs $l_1, \ldots, l_r$.

On veut donc minimiser $L = \sum_{i=1}^{r} p_i\, l_i$ sous la contrainte $\sum_{i=1}^{r} a^{-l_i} \leqslant 1$. On peut s'imposer $\sum_{i=1}^{r} a^{-l_i} = 1$ (en effet, on a tout intérêt à utiliser tout l'espace que laisse notre contrainte). On applique encore la méthode des multiplicateurs de Lagrange :

$$\begin{aligned} L - \lambda g &= \sum_{i=1}^{r} p_i\, l_i - \lambda \left(\sum_{i=1}^{r} a^{-l_i} - 1 \right), \\ \frac{\partial (L - \lambda g)}{\partial l_i} &= p_i + \lambda (\log a) e^{-l_i \log a}. \end{aligned}$$

Cette dérivée partielle s'annule lorsque :

$$-l_i \log a = \log p_i + \mu,$$

où μ est une constante (dépendant de λ) à déterminer. On a alors :

$$l_i = \frac{1}{\log a} \log \frac{1}{p_i} + \mu'.$$

Il faut vérifier la contrainte :

$$\begin{aligned} 1 = \sum_{i=1}^{r} a^{-l_i} &= \sum_{i=1}^{r} e^{-l_i \log a} \\ &= e^{\mu} \sum_{i=1}^{r} p_i = e^{\mu}. \end{aligned}$$

On a donc $\mu = 0$, et le codage optimal devrait donc vérifier :

$$\begin{aligned} l_i &= -\log_a p_i \\ L &= \frac{H(p)}{\log a}. \end{aligned}$$

Le codage optimal $\frac{H(p)}{\log a}$ est réalisé à un bit près. En effet, $-\log_a p_i$ n'est en général pas entier. En prenant les longueurs $l_i = \lceil -\log_a p_i \rceil$, on obtient :

$$L - \frac{H(p)}{\log a} = \sum_{i=1}^{r} \Big(\lceil -\log_a p_i \rceil + \log_a p_i \Big) p_i \leqslant \sum_{i=1}^{r} p_i \leqslant 1,$$

donc $\frac{H(p)}{\log a} \leqslant L \leqslant \frac{H(p)}{\log a} + 1$.

5.5 Résumé

Étant donné un alphabet, des événements à coder par des mots, et leur probabilité d'apparition, on sait maintenant que :

- pour obtenir le meilleur codage, il faut que $l_i = -\log_a p_i$;
- le meilleur codage sera le plus long lorsque tous les p_i sont égaux (c'est donc la situation la plus embêtante).

Pour plus d'informations, se référer à *Entropy and Information Theory* (de Cover et Thomas).

6 Graphes décisionnels et réseaux bayésiens[10]

6.1 Graphes décisionnels

6.1.1 Qu'est-ce et à quoi cela sert-il ?

Les graphes décisionnels servent à représenter des règles décisionnelles probabilistes (raisonnement flou). Sachant que l'on est dans l'état x, le graphe donne la probabilité de passer dans l'état y.

Cela sert, par exemple, à rechercher les causes d'un accident ou à faire du diagnostic médical (étant donné les symptômes, on assigne une probabilité aux maladies).

Chaque nœud peut prendre plusieurs états.

Notons S l'ensemble des nœuds du graphe, et x_s l'état du noeud $s \in S$. Les états de certains nœuds peuvent être des fonctions — déterministes — de l'état d'autres noeuds.

Le graphe est la donnée de l'ensemble des sommets S et de l'ensemble des transitions entre les sommets $\{s \to t \,|\, s, t \in S\}$. On va supposer dans la suite que le graphe est acyclique.

L'état global du graphe est donc donné par le vecteur aléatoire $X = (X_s)_{s \in S}$.

Dans toute la suite, nous adopterons les notations suivantes : X désigne une variable ou un vecteur aléatoire correspondant aux états des sommets du graphe ; x désigne un état particulier du graphe, d'un sous-graphe ou d'un sommet (suivant l'indice).

Ce qui va nous intéresser plus particulièrement est la probabilité qu'un sommet soit dans un certain état, soit :

$$\mathbb{P}\left[X_s = x_s \,|\, X_{S-s} = x_{S-s}\right].$$

[10] Cours du 15 novembre, rédigé par Charles Bouillaguet.

6.1.2 Hypothèse des spécifications locales

Le *voisinage immédiat* de s est :

$$V_s = \{t \in S \,|\, \exists t \to s\}\,.$$

On va faire l'hypothèse que :

$$\mathbb{P}\left[X_s = x_s \,|\, X_{S-s} = x_{S-s}\right] = \mathbb{P}\left[X_s = x_s \,|\, X_{V_s} = x_{V_s}\right],$$

i.e. que la probabilité qu'un sommet s soit dans un certain état ne dépend que des états des sommets qui ont une transition vers s.

On va aussi supposer que le graphe se divise en « couches » ayant des liens les unes vers les autres, dans un seul sens.

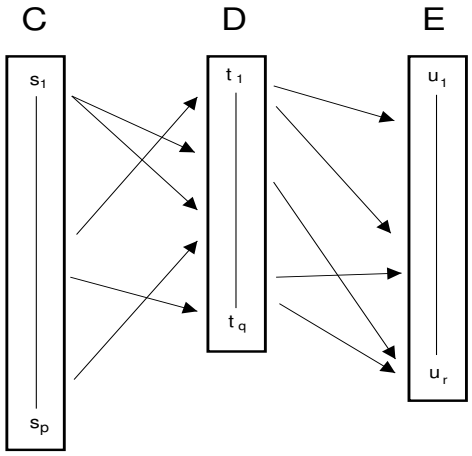

FIG. 2 – Graphe à trois couches

Par exemple, sur la figure 2, on a :

$$\begin{array}{rcl} S & = & C \cup D \cup E \\ X_S & = & (X_C, X_D, X_E)\,. \end{array}$$

L'hypothèse des spécifications locales consiste à dire que l'état d'un sommet d'une certaine couche ne dépend que des sommets de la couche immédiatement à gauche faisant une transition vers lui. Pour un sommet t de la couche D, cela s'écrit :

$$\begin{aligned} \mathbb{P}\left[X_t = x_t \,|\, X_C = x_C\right] &= \mathbb{P}\left[X_t = x_t \,|\, X_{V_t} = x_{V_t}\right] \\ &= \phi_t(x_t, x_{V_t}). \end{aligned}$$

Si on considère la couche D dans sa globalité, on voit bien qu'elle ne dépend que de la couche C :

$$\begin{aligned}\mathbb{P}\left[X_D = (x_{t_1}, x_{t_2}, \ldots, x_{t_q}) \,|\, X_C = x_C\right] &= \prod_{j=1}^{q} \mathbb{P}\left[X_{t_j} = x_{t_j} \,|\, X_C = x_C\right] \\ &= \prod_{j=1}^{q} \mathbb{P}\left[X_{t_j} = x_{t_j} \,|\, X_{V_{t_j}} = x_{V_{t_j}}\right].\end{aligned}$$

De même,

$$\mathbb{P}\left[X_u = x_u \,|\, X_C = x_C, X_D = x_D\right] = \mathbb{P}\left[X_u = x_u \,|\, X_{V_u} = x_{V_u}\right]$$

et

$$\begin{aligned}\mathbb{P}\left[X_E = x_E \,|\, X_D = x_D, X_C = x_C\right] &= \mathbb{P}\left[X_E = x_E \,|\, X_D = x_D\right] \\ &= \prod_{j=1}^{r} \mathbb{P}\left[X_{u_j} = x_{u_j} \,|\, X_{V_{u_j}} = x_{V_{u_j}}\right],\end{aligned}$$

donc (en allégeant les notations) :

$$\begin{aligned}\mathbb{P}\left[(x_E, x_D, x_C)\right] &= \mathbb{P}\left[x_E \,|\, x_D, x_C\right] \mathbb{P}\left[x_D, x_C\right] \\ &= \mathbb{P}\left[x_E \,|\, x_D\right] \mathbb{P}\left[x_D \,|\, x_C\right] \mathbb{P}\left[x_C\right].\end{aligned}$$

6.1.3 Problème de l'inférence

Il s'agit de partir de la fin (de l'état d'un sommet de la couche la plus à droite) et de déterminer quelles peuvent être les « causes » de cet état.

Par exemple, on peut supposer :

- x_C connu (variables d'ambiance — pression, température, …) ;
- x_E connu (nature du problème — crevaison, explosion, arrêt des moteurs, …) ;
- x_D inconnu (cause du problème).

Le but du jeu est donc de trouver un *estimateur* g, qui donne une valeur probable de x_D en fonction de x_C et x_E :

$$\widehat{x_D} = g(x_C, x_E).$$

Il s'agit ensuite d'évaluer la « performance » de g. Pour cela, on se donne une *mesure de performance*.

Par exemple, la quantité d'erreurs commises par g :

$$\begin{aligned}
\operatorname{Perf} \widehat{x_D} &= \mathbb{P}\left[X_D \neq g(X_C, X_E)\right] \\
&= 1 - \mathbb{P}\left[X_D = g(X_C, X_E)\right] \\
&= 1 - \sum_{x_C, x_E} \mathbb{P}\left[X_D = g(X_C, X_E), X_C = x_C, X_E = x_E\right] \\
&= 1 - \sum_{x_C, x_E} \mathbb{P}\left[X_D = g(x_C, x_E), X_C = x_C, X_E = x_E\right] \\
&= 1 - \sum_{x_C, x_E} \mathbb{P}\left[X_D = g(x_C, x_E) \,|\, X_C = x_C, X_E = x_E\right] \mathbb{P}\left[X_C = x_C, X_E = x_E\right].
\end{aligned}$$

On cherche à minimiser cette grandeur. Pour cela, on va tenter de maximiser chaque terme de la somme. Dans chacun des termes de la somme, x_E et x_C sont fixés. On cherche à maximiser :

$$\mathbb{P}\left[X_D = \lambda \,|\, X_C = x_C, X_E = x_E\right] = \phi(\lambda, x_C, x_E).$$

En fait,

$$\max_{\lambda} \mathbb{P}\left[X_D = \lambda \,|\, X_C = x_C, X_E = x_E\right]$$

est atteint pour un certain λ_0. On pose simplement $g(x_C, x_E) = \lambda_0$.

C'est le principe du *maximum de vraisemblance* : on associe à $g(x_C, x_E)$ la valeur de x_D la plus probable.

6.2 Formule de Bayes

Prenons le graphe à deux couches X et Y de la figure 3.

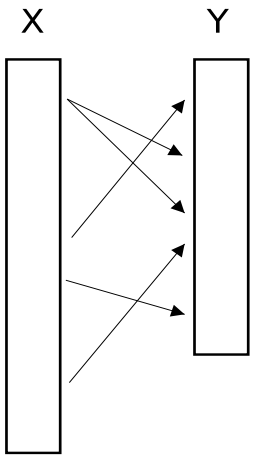

FIG. 3 – Graphe à deux couches

Supposons que l'on observe $Y = y$. Quelle sont les valeurs de X les plus probables ? On a :

$$\begin{aligned}\mathbb{P}[Y = y \,|\, X = x] &= \frac{\mathbb{P}[Y = y \text{ et } X = x]}{\mathbb{P}[Y = y]} \\ &= \frac{\mathbb{P}[Y = y \text{ et } X = x]}{\sum_{x'} \mathbb{P}[Y = y \text{ et } X = x']}.\end{aligned}$$

Cette dernière formule est connue sous le nom de *formule de Bayes*. Si on veut maximiser cette grandeur par rapport à x, on peut toujours appliquer le principe du maximum de vraisemblance :

$$\widehat{x}(y) = \arg\max_x \mathbb{P}[Y = y \,|\, X = x].$$

6.3 Loi de Gibbs

On va désormais se placer dans un cadre un peu plus général. En particulier, on supposait auparavant que le graphe était découpé en couches, et donc ne contenait pas de cycles. On ne fera plus cette hypothèse dorénavant. On va introduire maintenant une famille de lois appelées *lois de Gibbs*. Ces lois ont été introduites dans le cadre du ferro-magnétisme. Voyons donc un exemple de telle loi.

Considérons un réseau (une « grille » en 2 dimensions) de particules $s \in S$ dont le spin vaut $x_s = \pm 1$. Un état du réseau est $x_S = (x_s)_{s \in S}$. Le *modèle d'Ising* nous donne le niveau d'énergie d'une certaine configuration :

$$U(x) = \alpha \sum_{s \in S} x_s + \beta \sum_{\langle s,t \rangle \in S} x_s x_t,$$

où $\langle s, t \rangle$ veut dire que s et t sont voisins.

D'autres raisonnements physiques — honteusement admis et passés sous silence ici — nous permettent d'affirmer que l'énergie d'une telle distribution est une fonction de la température, ce que nous exprimerons par :

$$\mathbb{E}[U(X)] = C(T).$$

On voudrait, en faisant le moins d'hypothèses possible, trouver une loi de probabilité $p(x) = \mathbb{P}[X = x]$ pour X (à valeurs dans $\Omega = \{-1, +1\}^S$) vérifiant cette dernière équation. Faire le moins d'hypothèses possible signifie — d'après les physiciens — ne pas supposer que la matière est dans un ordre particulier, et donc que l'entropie de la loi est maximale.

On cherche donc à maximiser l'entropie, notée dans toute la suite H :

$$H(p) = -\sum_{x\in\Omega} p(x)\log p(x)$$

avec les contraintes :

$$\mathbb{E}\left[U(X)\right] = \sum_x p(x)U(x) = C(T) \qquad \text{et} \qquad \sum_x p(x) = 1.$$

On applique la bonne vieille méthode des multiplicateurs de Lagrange :

$$L(p) = H(p) - \lambda\sum_x p(x)U(x) - \mu\sum_x p(x).$$

On cherche les valeurs de $p(x)$ (pour tout $x \in \Omega$) qui maximisent $L(p)$. On doit donc avoir :

$$\forall x \in \Omega, \qquad \frac{\partial L}{\partial p(x)} = 0.$$

Or :

$$\begin{aligned}\frac{\partial H(p)}{\partial p(x)} &= -\frac{\partial p(x)\log p(x)}{\partial p(x)}\\ &= -1 - \log p(x),\end{aligned}$$

donc :

$$\frac{\partial L}{\partial p(x)} = -1 - \log p(x) - \lambda U(x) - \mu = 0,$$

soit :

$$\begin{aligned}\log p(x) &= -\lambda U(x) - 1 - \mu\\ p(x) &= \gamma e^{-\lambda U(x)}\end{aligned}$$

où $\gamma = e^{-(1+\mu)}$. Comme $\sum_x p(x) = 1$, on a nécessairement :

$$\gamma = \frac{1}{\sum_{x\in\Omega} e^{-\lambda U(x)}},$$

ce qui fixe μ. Et comme $\sum_x p(x)U(x) = C(T)$, on a également :

$$\gamma\sum_x e^{-\lambda U(x)}U(x) = C(T),$$

ce qui fixe λ (on admet qu'une solution en $\lambda > 0$ existe).

Ainsi, $p(x) = \gamma e^{-\lambda U(x)}$ définit la loi de Gibbs associée à l'énergie U (on admet que le point d'annulation trouvé précédemment est un maximum de L).

7 Champs markoviens[11]

7.1 Notations et définitions utilisées

Considérons un réseau fini, et S l'ensemble de ses sommets. On note s le singleton $\{s\}$. Le voisinage V_s de s est défini par $V_s = \{t \in S \,|\, t \leftrightarrow s\}$. On suppose que la relation de voisinage est symétrique ; aussi, $t \in V_s \iff s \in V_t$.

Chaque sommet a un état qui est soit 0, soit 1. L'ensemble de tous les états possibles du réseau est $\Omega = \{0,1\}^S$. Si $x \in \Omega$ et $E \subset S$, on note $x_E = (x_i)_{i \in E}$. On appelle X la variable aléatoire — à valeurs dans Ω — attachée à l'état du réseau.

On cherche des lois de probabilité P sur Ω ($P(x) = \mathbb{P}[X = x]$) vérifiant la *propriété de Markov par rapport aux voisinages* $(V_s)_{s \in S}$, *i.e.* pour tous $s \in S$, $x_s \in \{0,1\}$ et $x_{S-s} \in \{0,1\}^{S-s}$, on a :

$$\mathbb{P}[X_s = x_s \,|\, X_{S-s} = x_{S-s}] = \mathbb{P}[X_s = x_s \,|\, X_{V_s} = x_{V_s}].$$

On rappelle la loi de Gibbs :

$$P(x) = \frac{1}{Z} e^{-U(x)} \qquad \text{avec} \qquad Z = \sum_{x \in \Omega} e^{-U(x)},$$

où $U : \Omega \to \mathbb{R}$ est une fonction qui à tout x associe une *énergie* $U(x)$. On sait que c'est la loi la plus dispersée possible avec la contrainte $\mathbb{E}[U(X)] = T$, T étant une constante donnée (appelée *température*).

Définition 7.1. Si $K \subset S$, on dit que K est une *clique* si, pour tous s et t distincts dans K, s et t sont voisins. On note $\mathcal{C}$ l'ensemble des cliques.

Par exemple, si $S = \mathbb{Z}^2$ et si $s \leftrightarrow t \iff |s - t| = 1$, alors les singletons $\{s\}$ sont des cliques, ainsi que les paires $\{s,t\}$, où $|s - t| = 1$.

Définition 7.2. On définit une *énergie de voisinage* en posant :

$$U(x) = \sum_{K \in \mathcal{C}} W_K(x_K),$$

où les W_K sont des fonctions $\{0,1\}^K \to \mathbb{R}$.

[11] Cours du 22 novembre, rédigé par Stéphane Glondu.

Le modèle d'Ising est un exemple d'énergie de voisinage :

$$W_{\{s\}}(x_{\{s\}}) = \alpha\, x_s,$$
$$W_{\{s,t\}}(x_{\{s,t\}}) = \beta\, x_s\, x_t,$$

ce qui donne l'expression suivante :

$$U(x) = \alpha \sum_{s\in S} x_s + \beta \sum_{\substack{s,t\in S \\ |s-t|=1}} x_s s_t.$$

Remarque 7.1. On peut définir un réseau symétrique indifféremment par ses cliques $\mathcal{C}$ ou par sa relation de voisinage $\leftrightarrow \subset S^2$.

7.2 Loi de Gibbs et propriété de Markov

Théorème 7.1. *Si :*

$$P(x) = \frac{1}{Z} e^{-U(x)} \qquad \textit{avec} \qquad Z = \sum_{x\in\Omega} e^{-U(x)},$$

où l'énergie U est définie par :

$$U(x) = \sum_{K\in\mathcal{C}} W_K(x_K),$$

alors P vérifie la propriété de Markov par rapport aux voisinages définis par les cliques $\mathcal{C}$.

Démonstration. Soient $s \in S$ et $x_{S-s} \in \{0,1\}^{S-s}$. On note $x^{[i]}$ le vecteur où la coordonnée s vaut i, et où les autres coordonnées valent x_{S-s}. On a :

$$\begin{aligned}
\mathbb{P}\left[X_s = 1 \,|\, X_{S-s} = x_{S-s}\right] &= \frac{\mathbb{P}\left[X_s = 1 \text{ et } X_{S-s} = x_{S-s}\right]}{\mathbb{P}\left[X_{S-s} = x_{S-s}\right]} \\
&= \frac{P\left(x^{[1]}\right)}{P\left(x^{[0]}\right) + P\left(x^{[1]}\right)} \\
&= \frac{\frac{1}{Z} e^{-U\left(x^{[1]}\right)}}{\frac{1}{Z} e^{-U\left(x^{[0]}\right)} + \frac{1}{Z} e^{-U\left(x^{[1]}\right)}} \\
&= \frac{e^{-U\left(x^{[1]}\right)}}{e^{-U\left(x^{[0]}\right)} + e^{-U\left(x^{[1]}\right)}}.
\end{aligned}$$

Notons $\mathcal{D}$ l'ensemble des cliques incluses dans $s \cup V_s$. On a :

$$U(x) = \underbrace{\sum_{K \in \mathcal{D}} W_K(x_K)}_{Q(x)} + \underbrace{\sum_{K \in \mathcal{C}-\mathcal{D}} W_K(x_K)}_{R(x)},$$

où $Q(x)$ ne dépend que des $(x_t)_{t \in s \cup V_s}$, et $R(x)$ ne dépend que des $(x_t)_{t \notin s \cup V_s}$. On a alors $R\left(x^{[0]}\right) = R\left(x^{[1]}\right) = R$, et :

$$\begin{aligned}\mathbb{P}\left[X_s = 1 \,|\, X_{S-s} = x_{S-s}\right] &= \frac{e^{-Q\left(x^{[1]}\right)-R}}{e^{-Q\left(x^{[0]}\right)-R} + e^{-Q\left(x^{[1]}\right)-R}} \\ &= \frac{e^{-Q\left(x^{[1]}\right)}}{e^{-Q\left(x^{[0]}\right)} + e^{-Q\left(x^{[1]}\right)}},\end{aligned}$$

et cette dernière expression ne dépend que des $(x_t)_{t \in V_s}$. D'où :

$$\mathbb{P}\left[X_s = 1 \,|\, X_{S-s} = x_{S-s}\right] = \mathbb{P}\left[X_s = 1 \,|\, X_{V_s} = x_{V_s}\right],$$

ce qui est la propriété de Markov par rapport aux voisinages $(V_s)_{s \in S}$. □

Définition 7.3. Un tel réseau, vérifiant la propriété de Markov relative aux voisinages $(V_s)_{s \in S}$, est appelé *champ markovien*.

7.3 Quelques applications

7.3.1 Transmission d'une image

Si on met une loi de Gibbs avec énergie de voisinage sur l'ensemble des pixels d'une image, on peut corriger une image localement dégradée.

7.3.2 Résultat d'un référendum

On considère un ensemble S de votants, et on suppose qu'ils s'influencent de manière symétrique. Pour $x \in \Omega$, et $s \in S$, on pose $x_s = 1$ si s a voté *oui*, et $x_s = -1$ si s a voté *non*. On modélise la situation par un modèle d'Ising :

$$U(x) = \beta \sum_{s \in S} x_s - \alpha \sum_{\{s,t\} \in \mathcal{C}} x_s s_t, \qquad \alpha > 0.$$

Si $\beta > 0$, la tendance générale est au *non* ; si $\beta < 0$, la tendance générale est au *oui* ; si $\beta = 0$, il n'y a pas de tendance générale. On suppose connaître α,

β, ainsi que les votes des relations de s. On peut alors appliquer le principe du maximum de vraisemblance pour deviner le vote de s :

$$\begin{aligned}\mathbb{P}\left[X_s = 1 \,|\, X_{S-s} = x_{S-s}\right] &= \mathbb{P}\left[X_s = 1 \,|\, X_{V_s} = x_{V_s}\right] \\ &= P\left(x^{[1]}\right) \\ &= \frac{1}{Z} e^{-\beta + \alpha \sum_{t \in V_s} x_t}.\end{aligned}$$

Par conséquent,

$$\frac{\mathbb{P}\left[X_s = 1 \,|\, X_{S-s} = x_{S-s}\right]}{\mathbb{P}\left[X_s = -1 \,|\, X_{S-s} = x_{S-s}\right]} = e^{-2\beta + 2\alpha \sum_{t \in V_s} x_t}.$$

Si $\displaystyle\sum_{t \in V_s} x_t > \frac{\beta}{\alpha}$, on peut supposer que s a voté *oui* ;

si $\displaystyle\sum_{t \in V_s} x_t < \frac{\beta}{\alpha}$, on peut supposer que s a voté *non*.

8 Recuit simulé[12]

8.1 Utilité du recuit simulé

On considère ici un réseau où l'ensemble des sites S — toujours fini — est très grand. Chaque site est dans un certain état ; on note F l'ensemble des états possibles. Par exemple, pour une image, on peut associer à chaque pixel sa couleur. On se donne une *fonction de coût* $U : \Omega = F^S \to \mathbb{R}$ que l'on souhaiterait minimiser. Bien sûr, on pourrait énumérer toutes les valeurs prises par U (car Ω est fini), mais cela peut prendre beaucoup de temps si Ω est très grand. De plus, le calcul de U en un point peut lui-même être très long. On propose ici une méthode probabiliste pour minimiser U qui fait appel à des calculs certes longs, mais beaucoup moins que l'exploration exhaustive. C'est utile, par exemple, pour apporter des solutions satisfaisantes aux problèmes NP-complets.

8.2 Exploration probabiliste

On définit une loi de probabilité (de Gibbs) sur Ω par :

$$Q_T(x) = \frac{1}{Z_T} e^{-\frac{U(x)}{T}} \qquad \text{avec} \qquad Z_T = \sum_{x \in \Omega} e^{-\frac{U(x)}{T}},$$

[12]Cours du 29 novembre, rédigé par Stéphane Glondu.

où $T > 0$ est un paramètre que l'on appellera *température*. $Z_T > 0$ est là juste pour normaliser la loi ; il n'est jamais calculé explicitement et n'interviendra jamais dans les calculs.

Théorème 8.1. *On note* $\Omega_{\min} = \{x \in \Omega \,|\, \forall x' \in \Omega, U(x) \leqslant U(x')\}$. Q_T *se concentre sur* $\Omega_{\min}$ *quand* $T \to 0$, i.e. *:*

$$\lim_{T \to 0} Q_T(\Omega_{\min}) = 1.$$

Démonstration. Si $\Omega_{\min} = \Omega$, le résultat est évident. Plaçons-nous donc dans le cas où $\Omega - \Omega_{\min} \neq \emptyset$. Posons $a = \min_{\Omega - \Omega_{\min}} U$. a est bien défini car Ω est fini. On a alors $\forall x \in \Omega - \Omega_{\min}, e^{-\frac{U(x)}{T}} \leqslant e^{-\frac{a}{T}}$, et en sommant, on obtient :

$$Q_T(\Omega - \Omega_{\min}) \leqslant \frac{|\Omega - \Omega_{\min}|}{Z_T} e^{-\frac{a}{T}}.$$

De plus, $\forall x \in \Omega_{\min}, U(x) < a$, donc il existe $\varepsilon > 0$ tel que $\forall x \in \Omega_{\min}, U(x) \leqslant a - \varepsilon$. On peut donc écrire :

$$Z_T \geqslant \sum_{x \in \Omega_{\min}} e^{-\frac{U(x)}{T}} \geqslant \sum_{x \in \Omega_{\min}} e^{-\frac{a-\varepsilon}{T}} \geqslant |\Omega_{\min}|\, e^{-\frac{a-\varepsilon}{T}}.$$

On en déduit :

$$Q_T(\Omega - \Omega_{\min}) \leqslant \frac{|\Omega - \Omega_{\min}|}{|\Omega_{\min}|} e^{-\frac{\varepsilon}{T}} \xrightarrow[T \to 0]{} 0.$$

On trouve alors le résultat en passant au complémentaire. □

On voit alors que pour trouver un minimum de U, une idée serait de tirer $x \in \Omega$ au hasard avec la loi Q_T, T étant très petit. Mais comment procéder ?

Si, par exemple, on définit une loi de probabilité sur $\Omega = \{\alpha_1, \alpha_2, \alpha_3\}$ par $Q(\alpha_i) = q_i$ $(q_1 + q_2 + q_3 = 1)$, alors on peut procéder comme suit. On tire uniformément un réel x dans $[0, 1[$, et :

- si $x \in [0, q_1[$, on renvoie α_1 ;
- si $x \in [q_1, q_1 + q_2[$, on renvoie α_2 ;
- si $x \in [q_1 + q_2, 1[$, on renvoie α_3.

Néanmoins, cette méthode n'est pas applicable en pratique dans le cas qui nous intéresse car $|\Omega|$ est trop grand. Il faut donc une autre méthode de simulation de la loi Q_T.

8.3 Dynamique de Metropolis

Le problème fut étudié par le chimiste Metropolis. Son idée est de tirer au hasard $X_1, X_2, \ldots, X_n, \ldots$ dans Ω en s'arrangeant pour la suite obtenue forme une chaîne de Markov ergodique dont la loi limite est Q_T. On pose alors :

$$\mathbb{P}\left[X_{n+1} = y \,|\, X_n = x\right] = R(x,y) \qquad \text{avec} \qquad \forall x \in \Omega, \quad \sum_{y\in\Omega} R(x,y) = 1.$$

La loi de X_n va converger vers Q — unique — vérifiant :

$$\forall y \in \Omega, \quad \sum_{x\in\Omega} Q(x)R(x,y) = Q(y).$$

On aimerait bien que $Q(x) = Q_T(x)$! Des raisonnements spécifiques à la chimie nous poussent à introduire le :

Définition 8.1 (Principe de microéquilibre)**.** Lorsque l'on a :

$$\forall x, y \in \Omega, \quad Q(x)R(x,y) = Q(y)R(y,x),$$

on dira que Q et R vérifient le *principe de microéquilibre* (en anglais *micro-balance*).

Ce principe a une propriété remarquable :

Théorème 8.2 (Microéquilibre)**.** *Si Q et R vérifient le principe de microéquilibre, alors Q est invariante par R,* i.e. *:*

$$Q \times R = Q.$$

Démonstration.

$$\begin{aligned}\sum_{x\in\Omega} Q(x)R(x,y) &= \sum_{x\in\Omega} Q(y)R(y,x) \\ &= Q(y) \underbrace{\sum_{x\in\Omega} R(y,x)}_{\substack{=1 \\ \text{car } R \text{ stochastique}}} \\ &= Q(y)\end{aligned}$$

□

Remarquons que si $U(y) < U(x)$, alors $Q_T(y) > Q_T(x)$. C'est en essayant de vérifier le principe de microéquilibre que Metropolis a suggéré le processus suivant :

Définition 8.2 (Dynamique de Metropolis)**.** On suppose que $X_n = x$. On choisit $y \in \Omega$ au hasard avec une loi uniforme. On définit X_{n+1} comme suit :
- si $Q_T(y) \geqslant Q_T(x)$, alors $X_{n+1} = y$;
- si $Q_T(y) < Q_T(x)$, alors :

$$X_{n+1} = \begin{cases} y & \text{avec la probabilité } \dfrac{Q_T(y)}{Q_T(x)}, \\ x & \text{avec la probabilité } 1 - \dfrac{Q_T(y)}{Q_T(x)}. \end{cases}$$

Théorème 8.3 (Metropolis)**.** *Q_T est la loi invariante de la chaîne de Markov définie ci-dessus.*

Démonstration. Soient R la matrice de transition de la chaîne de Markov définie par la dynamique de Metropolis, et x et y deux éléments de Ω. Supposons que $Q_T(x) < Q_T(y)$ (les autres cas se traitent de la même manière). On a alors :

$$R(x,y) = \frac{1}{|\Omega|} \qquad \text{et} \qquad R(y,x) = \frac{1}{|\Omega|}\frac{Q_T(x)}{Q_T(y)}.$$

On vérifie alors facilement que $Q_T(x)R(x,y) = Q_T(y)R(y,x)$, et on conclut grâce au théorème 8.2. □

On a donc trouvé une chaîne de Markov dont la loi limite est Q_T. Mais combien d'étapes faut-il pour atteindre Q_T ? Bien entendu, *there is no free lunch* : plus T est petite, plus la convergence est lente. Si on note μ_n la loi de X_n, on a un schéma du type :

$$d\big(\mu_n(T), Q_T\big) \leqslant \rho(T)^n \qquad \text{avec} \qquad \rho(T) \xrightarrow[T\to 0]{} 1.$$

Pour pallier ce problème, Geman et Hayek ont repris une idée introduite originellement par le physicien Kirkpatrick, qui consiste à faire évoluer lentement la température vers 0 (schéma de refroidissement). Ainsi, pour passer de X_n à X_{n+1}, on se sert de la température T_n. Il ont établi une expression optimale de T_n :

Théorème 8.4 (Geman & Hayek)**.** *Si $T_n = \frac{C}{\log n}$, alors la loi de X_n tend vers la loi uniforme sur $\Omega_{\min}$, où C est une constante (liée à ε tel que défini dans la démonstration du théorème 8.1).*

En pratique, on utilise plutôt une expression du type $T_n = a^n$, où $a < 1$ est un réel positif très proche de 1.

8.4 Quelques applications

8.4.1 Le voyageur de commerce (*traveling salesman*)

Un voyageur de commerce doit visiter K villes, en passant une seule fois par chacune d'entre elles, et en revenant (à la fin) à son point de départ. Quel est le chemin de longueur minimale ? Ici, $\Omega = \mathfrak{S}_K$, groupe des permutations sur $\{1, \ldots, K\}$. Si $\sigma \in \Omega$, $\sigma(1)$ désigne la première ville visitée, $\sigma(2)$ la deuxième, et ainsi de suite. On a $|\Omega| = K!$. On définit $U(\sigma)$ comme étant la longueur totale de la permutation $\sigma \in \Omega$.

On applique une modification élémentaire — choisie aléatoirement — qui transforme σ en σ'. On garde σ' si $U(\sigma') \leqslant U(\sigma)$. Sinon, on garde σ' avec une probabilité de $e^{-\frac{U(\sigma')-U(\sigma)}{T_n}}$. Si on choisit comme modifications élémentaires les transpositions, $U(\sigma') - U(\sigma)$ peut se calculer très rapidement. On peut aussi tabuler la fonction exponentielle pour aller encore plus vite.

8.4.2 Centralisation des communications

Ce problème fut étudié par le CNET. On veut placer de manière optimale 70 centralisateurs sur le territoire français. Chaque centralisateur C_i est relié à des nœuds terminaux $V_{1,i}, \ldots, V_{n_i,i}$. On définit une fonction de coût par :

$$U(x) = \sum_{i,j} d\big(C_i, V_{j,i}\big).$$

Le problème a été traité par Lutton grâce au recuit simulé.

8.4.3 Attribution d'emplois du temps

Ce problème fut étudié à Rennes (par PSA). Il s'agit de régler les emplois du temps pour la formation continue (environ 8 000 utilisateurs). On veut trouver des assignations compatibles avec des contraintes. Cela revient à minimiser une fonction de coût $U(x)$ avec des contraintes $C_i(x) > 0$. La résolution de ce problème a mené à l'élaboration d'un logiciel utilisant le recuit simulé et la méthode des multiplicateurs de Lagrange.

9 Algorithmes génétiques[13]

9.1 Aspect qualitatif

Certains problèmes ont une complexité trop importante pour être résolus de façon exacte. On a vu que le recuit simulé permettait d'obtenir assez rapidement une approximation du résultat. Il existe un autre type d'algorithme générique, inspiré cette fois non pas de la chimie mais de la biologie. Il consiste à simuler l'évolution naturelle d'une population, en faisant évoluer un groupe d'individus aléatoirement et en gardant les meilleurs à chaque étape.

On notera donc Pop_n la population, c'est-à-dire l'ensemble des individus à la génération n. Si $x \in Pop_n$, on notera $f(x)$ sa « qualité ». f est en réalité une fonction à valeur dans $\mathbb{R}^+$, et l'on cherche l'individu x qui maximise $f(x)$.

Pour cela, on va simuler le processus d'évolution naturelle. On a besoin de coder chaque individu par son patrimoine génétique. Cela correspond à l'ADN en biologie. On fera évoluer cet ADN aléatoirement pour la population en suivant des règles ressemblant à celles qui régissent l'évolution naturelle, puis on simulera la sélection naturelle en gardant les individus qui nous plaisent le plus.

9.2 Règles régissant l'évolution naturelle

9.2.1 Formalisation

Chaque individu est codé par son patrimoine génétique. Il s'agit ici d'une suite finie à valeurs dans $\{0, 1\}$, chaque suite ayant la même longueur L. Dans la suite de cette section, on considérera les individus $x = (x_s)_{1 \leqslant s \leqslant L}$ et $y = (y_s)_{1 \leqslant s \leqslant L}$, qui seront les parents de deux fils x' et y'.

9.2.2 Mutations

Les mutations consistent à changer aléatoirement un bit du patrimoine génétique en son opposé. Il y a deux étapes :

1. on tire au hasard un booléen b avec une probabilité μ très faible qu'il soit vrai ;
2. si b est vrai, on tire au hasard s dans $\{1, \ldots, L\}$ avec la loi uniforme sur $\{1, \ldots, L\}$. Si le bit s de l'individu vaut 1, on le change en 0, sinon on le change en 1.

[13] Cours du 6 décembre, rédigé par Romain Bardou.

9.2.3 Crossing-over

En réalité, il ne s'agit pas du même crossing-over qu'en biologie. Tout ce que l'on fait est couper en deux les suites (x_s) et (y_s) au même endroit, tiré aléatoirement de façon uniforme, puis échanger les parties coupées. Cela donne deux nouveaux individus x' et y'. Formellement, on tire s aléatoirement dans $\{1, \dots, L-1\}$, puis on affecte les valeurs suivantes aux fils :

$$x' = x_1 \cdots x_s y_{s+1} \cdots y_L,$$
$$y' = y_1 \cdots y_s x_{s+1} \cdots x_L.$$

9.3 Règles régissant la sélection naturelle

9.3.1 Choix des individus à conserver

Le but est de conserver les meilleurs individus tout en gardant une certaine diversité des gènes disponibles dans la population. On peut donc trier les individus en fonction de leur *fitness* (donnée par la fonction f). Ensuite, on peut appliquer par exemple les règles suivantes :

1. On garde les $x\,\%$ meilleurs individus, où x n'est pas trop élevé (par exemple $10\,\%$). On notera M l'ensemble de ces individus.
2. On complète ensuite selon le nombre d'individus que l'on veut conserver au total, en choisissant au hasard parmi les individus disponibles. On notera $Dispos$ l'ensemble $Pop_n \setminus M$, et $Choix$ l'ensemble des individus choisis. On choisit avec une probabilité plus importante les individus les plus en forme ($f(x)$ élevé). Pour cela, on peut par exemple utiliser une loi du type :

$$P(x) = \frac{f(x)}{\displaystyle\sum_{z \in Dispos} f(z)} \qquad \text{pour } x \in Dispos.$$

   Il faudra choisir au hasard un individu avec une loi uniforme, puis utiliser la loi définie ci-dessus pour savoir si on le garde ou non, et réitérer le processus jusqu'à obtenir le nombre voulu d'individus à la génération $n+1$.

9.3.2 Passage d'une génération à une autre

Supposons que l'on maintienne une population de taille N à chaque génération, N étant pair. On peut alors passer d'une génération à une autre en choisissant $\frac{N}{2}$ couples au hasard avec une loi donnée. Pour chacun de

ces couples (x, y) on applique alors un crossing-over, suivi d'une mutation pour chacun des deux fils. On obtient alors $2N$ individus. On applique la règle précédente pour choisir quels individus conserver pour la génération suivante.

Bien sûr, cet algorithme est à répéter de nombreuses fois pour s'approcher de plus en plus du résultat voulu.

Cerf a montré que l'on était sûr — lorsque μ est petit — que les meilleurs individus pour f allaient prendre le dessus dans la population au bout d'un certain temps.

9.4 Exemple d'algorithme génétique

On suppose qu'on a une population de 50 individus, avec $L = 1\,000$. On suppose de plus que f vérifie :

$$\exists s_0, \quad f(x) = \begin{cases} 1 & \text{si } x \text{ a le gène } s_0, \\ \frac{1}{10} & \text{sinon.} \end{cases}$$

De plus, la probabilité de mutation est $\mu = \frac{1}{1\,000}$, on garde les 10 % meilleurs individus à chaque génération, et on utilise la loi donnée dans la section précédente pour choisir les survivants parmi ceux qui restent.

On a donc un gène s_0 qui rend un individu bien meilleur que les autres. On va déterminer dans quelle mesure il va envahir rapidement la population. On suppose qu'au début, personne ne possède ce gène.

On voit d'abord que la probabilité que le gène s_0 vaille 1 par mutation est $1 - \left(1 - \frac{1}{100\,000}\right)^{50}$, soit environ $\frac{50}{100\,000} = \frac{1}{2\,000}$. Le temps d'attente moyen pour voir apparaître ce gène est donc de 2 000 générations, et la variance de ce temps d'attente est aussi d'environ 2 000 générations (voir la proposition 1.4). Donc au bout de $2\,000 + 2 \times 2\,000 = 6\,000$, la probabilité que le gène spécial soit apparu est très forte.

Ensuite, une fois le gène apparu, il subsiste. En effet, comme cet individu est bien meilleur que les autres, il fait partie des 10 % d'individus choisis de manière certaine pour la génération suivante.

Une fois que cet individu est apparu, la quantité de super-individus (individus possédant le gène s_0) augmente beaucoup plus vite. En effet, il est transmis à exactement 1 fils la plupart du temps, pour chaque super-individu. Donc le nombre de super-individus double à chaque génération jusqu'à saturer les 10 % meilleurs individus. En fait, ceci est à tempérer par le fait qu'il y a une chance non nulle mais très faible que le fils d'un super-individu perde le super-gène par mutation, et que d'autres fils n'ayant pas le gène l'obtiennent eux-même par mutation. Mais ces phénomènes sont désormais négligables.

Après saturation des 10 %, la progression sera plus lente, mais va continuer.

9.5 Conclusion sur les algorithmes génétiques

Ce type d'algorithme est applicable dans de nombreux problèmes compliqués où l'on cherche une réponse approchée rapidement, sans avoir à dénombrer tous les cas. On peut montrer qu'il y a une équivalence mathématique entre le recuit simulé et les algorithmes génétiques. Intuitivement, dans le recuit simulé, on cherche à minimiser une énergie, alors que, dans les algorithmes génétiques, on cherche à maximiser une fonction de fitness.

La difficulté des algorithmes génétiques réside dans le choix des gènes codant les individus, dans le choix de la fonction f, et dans le choix de toutes les autres constantes du problèmes (μ, nombre d'individus par génération, méthode de choix des individus les meilleurs...).

10 Algorithmes randomisés[14]

10.1 Tri rapide (*quick sort*)

10.1.1 Présentation

On veut trier une liste $L = [x_1, \ldots, x_N]$ de N nombres en évitant de tester toutes les $O(n^2)$ comparaisons. Pour cela, on connaît l'algorithme de *quick sort* qui utilise le paradigme « diviser pour régner ». On décrit ici une version randomisée de cet algorithme, qui a de bonnes performances en moyenne :

1. on tire un élément x_j $(1 \leqslant j \leqslant N)$ au hasard dans L ;
2. on met dans une liste G tous les éléments de L plus petits que x_j ;
3. on met dans une liste D tous les éléments de L plus grands que x_j ;
4. on trie récursivement G et D ;
5. on fusionne le tout en renvoyant $G \cdot [x_j] \cdot D$.

On supposera pour simplifier que les éléments de L sont distincts deux à deux.

10.1.2 Analyse

Soit $T(N)$ le nombre de comparaisons effectuées lors du tri de la liste L. $T(N)$ est une variable aléatoire. On cherche à déterminer le nombre moyen de comparaisons, autrement dit $f(N) = \mathbb{E}[T(N)]$.

[14]Cours du 13 décembre, rédigé par Stéphane Glondu.

Pour cela, soit X le *rang* de x_j dans L et $m \in [1, \dots, N]$. Clairement,

$$\mathbb{P}[X = m] = \frac{1}{N}.$$

De plus, si $X = m$, il faut :
- $N - 1$ comparaisons pour construire G et D ;
- $T(m-1)$ comparaisons pour trier G ;
- $T(N-m)$ comparaisons pour trier D,

on a donc :

$$T(N) = T(m-1) + T(N-m) + N - 1,$$

soit encore :

$$\begin{aligned}\mathbb{E}[T(N) \mid X = m] &= \mathbb{E}[T(m-1)] + \mathbb{E}[T(N-m)] + N - 1 \\ &= f(m-1) + f(N-m) + N - 1.\end{aligned}$$

En sommant sur toutes les valeurs possibles de X, on obtient :

$$\begin{aligned}f(N) &= \sum_{m=1}^{N} \mathbb{E}[T(N) \mid X = m]\, \mathbb{P}[X = m] \\ &= \frac{1}{N} \sum_{m=1}^{N} \mathbb{E}[T(N) \mid X = m] \\ &= \frac{1}{N} \sum_{m=1}^{N} \Big(f(m-1) + f(N-m) + N - 1 \Big) \\ &= N - 1 + \frac{2}{N} \sum_{m=1}^{N-1} f(m).\end{aligned}$$

On va montrer par récurrence sur N que $f(N) \leqslant c\, N \log N$. Supposons que ce soit vrai jusqu'au rang $N-1$. Alors :

$$\sum_{m=1}^{N-1} f(m) \leqslant c \sum_{m=1}^{N-1} m \log m.$$

Or la croissance de la fonction $x \mapsto x \log x$ nous permet d'écrire :

$$m \log m \leqslant \int_{m}^{m+1} x \log x \, \mathrm{d}x,$$

d'où :

$$\sum_{m=1}^{N-1} f(m) \leqslant c \int_1^N x \log x \, \mathrm{d}x = c \left[\frac{x^2}{2} \log x - \frac{x^2}{4} \right]_1^N$$

$$\frac{1}{N} \sum_{m=1}^{N-1} f(m) \leqslant c \left(\frac{N}{2} \log N - \frac{N}{4} + \frac{1}{4N} \right)$$

$$f(N) \leqslant c\, N \log N + \underbrace{N \left(1 - \frac{c}{2}\right)}_{\leqslant 0} \underbrace{-1 + \frac{c}{2N}}_{\leqslant 0}$$

pour $c \geqslant 2$. On a donc finalement :

$$f(N) \leqslant 2N \log N,$$

inégalité qui est évidemment vérifiée pour $N = 1$.

10.2 Déconnexion d'un graphe connexe : *minimal cut*

10.2.1 Présentation

Le problème de la déconnexion d'un graphe connexe non orienté à N sommets et A arêtes consiste à trouver des arêtes $a_1, \ldots, a_m$ telles que si elles sont coupées, le graphe se sépare en deux composantes connexes (par exemple, sur la figure 4, on déconnecte le graphe en coupant les trois arêtes marquées). On voudrait que m soit minimal.

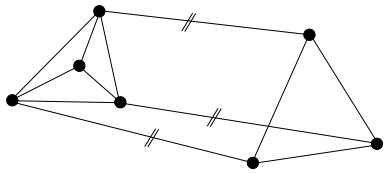

FIG. 4 – Illustration de la déconnexion de graphe

10.2.2 Méthode de condensation

On choisit (au hasard) une arête et on en fusionne les extrémités tout en éliminant les boucles. On réitère le processus jusqu'à n'avoir plus que deux sommets. Enfin, on choisit pour arêtes à couper les arêtes restantes. Par exemple, sur la figure 5, les arêtes condensées sont marquées d'un trait épais et les arêtes à couper sont marquées de deux traits fins.

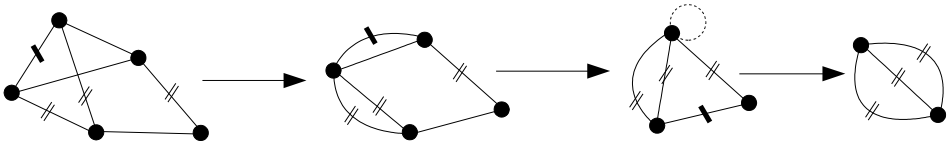

FIG. 5 – Illustration de la méthode de condensation

10.2.3 Analyse

Soit $K = \{a_1, \ldots, a_m\}$ une coupure minimale. Quelle est la probabilité de découvrir K par la méthode de condensation ?

On notera $d(s)$ le *degré* du sommet s, *i.e.* le nombre d'arêtes adjacentes à s. En posant $d_0 = \min_s d(s)$, on a :

$$\begin{aligned} 2A &= \sum_s d(s) \geqslant d_0 N \\ A &\geqslant d_0 \frac{N}{2} \end{aligned}$$

De plus, comme on peut déconnecter le graphe en coupant d_0 arêtes, la minimalité de m entraîne $d_0 \geqslant m$, ce qui donne finalement :

$$A \geqslant m\frac{N}{2}.$$

La probabilité de tirer une certaine arête est $\frac{1}{A}$. En notant X_n la n-ième arête tirée, on a :

$$\begin{aligned} \mathbb{P}\left[X_1 \notin K\right] &= 1 - \frac{m}{A} \\ &\geqslant 1 - \frac{2}{N} \\ \mathbb{P}\left[X_2 \notin K \text{ et } X_1 \notin K\right] &= \mathbb{P}\left[X_2 \notin K \,|\, X_1 \notin K\right] \mathbb{P}\left[X_1 \notin K\right] \\ &\geqslant \left(1 - \frac{2}{N-1}\right)\left(1 - \frac{2}{N}\right) \\ &\vdots \end{aligned}$$

Finalement, l'événement $E = \{\text{trouver } K\}$ peut encore s'écrire :

$$E = \left\{X_1 \notin K \text{ et } \ldots \text{ et } X_{N-2} \notin K\right\},$$

et on a :

$$\begin{aligned}\mathbb{P}[E] &\geqslant \left(1-\frac{2}{N}\right)\left(1-\frac{2}{N-1}\right)\left(1-\frac{2}{N-2}\right)\cdots\left(1-\frac{2}{3}\right)\\ &= \frac{N-2}{N}\times\frac{N-3}{N-1}\times\frac{N-4}{N-2}\times\cdots\times\frac{3}{5}\times\frac{2}{4}\times\frac{1}{3}\\ &= \frac{2}{N(N-1)}.\end{aligned}$$

Si on répète l'algorithme r fois, on a donc :

$$\begin{aligned}\mathbb{P}[\text{ne pas trouver } K] &\leqslant \left(1-\frac{2}{N(N-1)}\right)^r\\ &\leqslant \left(1-\frac{2}{N^2}\right)^r.\end{aligned}$$

On sait par ailleurs que, lorsque $r \to +\infty$, on a l'équivalent :

$$\left(1-\frac{x}{r}\right)^r \sim e^{-x},$$

et si on prend $r = a\,N^2$, on a donc pour N assez grand :

$$\mathbb{P}[\text{ne pas trouver } K] \leqslant e^{-2a}.$$

En prenant $a = 5$, on a $\mathbb{P}[\text{ne pas trouver } K] \leqslant 10^{-4}$ en moins de $5N^3$ opérations de condensation !