Combien de fois faut-il battre les cartes ?

Thomas Fernique Alexandra Ugolnikova

Dubna, 19-29 juillet 2014

1 Introduction

1.1 Chaînes de Markov

Définition 1 *Une* chaîne de Markov (X_t) *sur un espace* Ω *(ici fini) est un processus probabiliste qui se déplace sur* Ω *: quand* $X_t = x$*, alors* X_{t+1} *est choisi selon une loi de probabilité* $P(x, \cdot)$*.*

La position au temps $t + 1$ ne dépend ainsi que de celle au temps t : on dit qu'une chaîne de Markov n'a pas de "mémoire". Une chaîne de Markov se représente naturellement par un automate (Fig. 1).

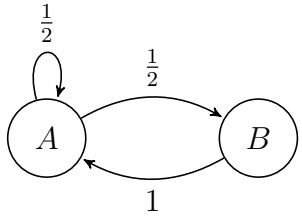

FIGURE 1 – Une chaîne de Markov simple.

Les chaînes de Markov sont fréquemment utilisées pour modéliser de façon simple un phénomène naturel (physique, biologique *etc*). Question typique : quelle est la probabilité qu'une chaîne se trouve en x à l'instant t ?

Une chaîne de Markov est complètement caractérisée par la *matrice* P telle que $P(x, y)$ donne la probabilité d'aller en y à partir de x. Chaque ligne est donc une loi $P(x, \cdot)$, donc somme à 1 : la matrice est dite *stochastique*. Si X

est le vecteur ligne dont la i-ème coordonnée donne la probabilité d'être dans le i-ème état, appliquer un pas de la chaîne revient à multiplier (à droite) par P.

Exemple 1 *La matrice correspondant à la chaîne de Markov de la Fig. 1 est*

$$P = \begin{pmatrix} 1/2 & 1/2 \\ 1 & 0 \end{pmatrix}.$$

Supposons qu'on lance la chaîne en A. On calcule la probabilité d'être en A ou B après k pas de la chaîne en multipliant k fois (à droite) par P (Fig. 2) :

$$(1,0) \xrightarrow{P} \left(\frac{1}{2}, \frac{1}{2}\right) \xrightarrow{P} \left(\frac{3}{4}, \frac{1}{4}\right) \xrightarrow{P} \left(\frac{5}{8}, \frac{3}{8}\right) \xrightarrow{P} \left(\frac{11}{16}, \frac{5}{16}\right) \xrightarrow{P} \left(\frac{21}{32}, \frac{11}{32}\right) \xrightarrow{P} \dots$$

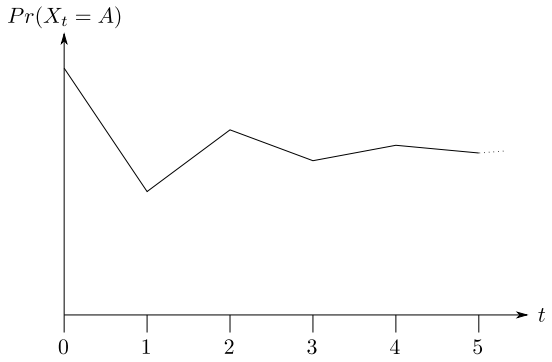

FIGURE 2 – Évolution d'une chaîne de Markov.

1.2 Distribution stationnaire

Définition 2 *Une distribution (*i.e.*, une loi de probabilité) μ sur Ω est dite* stationnaire *pour la chaîne de Markov P si $\mu = \mu P$.*

Une distribution stationnaire est donc un vecteur propre associé à la valeur propre 1. Pour en trouver une on résout le système formé des équations :

$$\pi(y) = \sum_{y \mid y \sim x} \pi(x) P(x, y).$$

Exemple 2 *On vérifie que la distribution μ définie par $\mu(A) = \frac{2}{3}$ et $\mu(B) = \frac{1}{3}$ est stationnaire pour la chaîne de Markov représentée Fig. 1.*

Une chaîne peut cependant avoir *plusieurs* distributions stationnaires.

Exemple 3 *On vérifie que, pour tout $x \in [0,1]$, la mesure μ_x définie par $\mu_x(A) = 0$, $\mu_x(B) = x$ et $\mu_x(C) = 1 - x$ est stationnaire pour la chaîne de Markov représentée Fig. 3.*

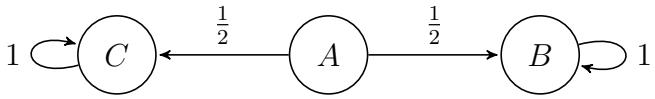

FIGURE 3 – Cette chaîne de Markov admet plusieurs distributions stationnaires.

Définition 3 *Une chaîne de Markov est dite* irréductible *si la probabilité d'aller de chaque état à tout autre est strictement positive :*

$$\forall x, y \in \Omega,\ \exists t,\ P^t(x,y) > 0.$$

On a alors la condition suffisante (conséquence du théorème 2) :

Proposition 1 *Une chaîne irréductible a une unique distribution stationnaire.*

Preuve. (sketch) On introduit le *temps de premier retour* en l'état x :

$$R(x) := \min\{t \geq 1 \mid x = X_0 = X_t\}.$$

On montre que $\mathbb{E}(R(x)) < \infty$ quand la chaîne est irréductible, puis que si π est stationnaire, alors $\pi(x) = 1/\mathbb{E}(R(x))$. □

On peut en fait être plus précis (exercice) :

Proposition 2 *Une chaîne de Markov sur Ω a une unique distribution stationnaire si et seulement s'il existe un unique* puits*, c'est-à-dire un sous-ensemble $S \subset \Omega$ tel que la chaîne est irréductible sur S et ne peut pas en sortir.*

On peut même caractériser les distributions stationnaires d'une chaîne de Markov : si $S_1,\ldots,S_k$ sont ses puits, alors elles sont de la forme

$$\mu = \sum_{i=1}^{k} a_i \mu_i,$$

où μ_i est l'unique distribution stationnaire de la restriction de la chaîne à S_i et les a_i sont des réels positifs de somme 1. C'est le cas de la chaîne de la Fig. 3.

Enfin, on peut montrer le théorème suivant, dit *ergodique* :

Théorème 1 *La proportion asymptotique du temps passé par une chaîne de Markov irréductible en chaque état est donnée par sa distribution stationnaire.*

Exemple 4 *Le célèbre* Google PageRank *n'est rien d'autre que la distribution stationnaire d'une chaîne de Markov. Les états sont les pages web et on quitte une page web donnée en choisissant équiprobablement une des pages vers lesquelles elle pointe. Le PageRank d'une page est donc la proportion du temps que passerait sur cette page un utilisateur se déplaçant sans fin aléatoirement sur le web. En fait, comme certaines pages n'ont aucun lien, on modifie la chaîne pour qu'elle soit irréductible en quittant chaque page web avec une petite probabilité vers l'une des quelques* $3,32$ *milliards de pages web (chiffre de juillet 2014). Le problème est alors de* calculer *la distribution stationnaire. Le vecteur propre dominant d'une matrice à* 10^{19} *coefficients est en effet incalculable. Sans compter que le web évolue en temps réel...*

1.3 Convergence

Exemple 5 *La matrice* P *de la chaîne de Markov représentée Fig. 1 vérifie :*

$$\underbrace{\begin{pmatrix} 1/2 & 1/2 \\ 1 & 0 \end{pmatrix}}_{P} = \underbrace{\begin{pmatrix} 1 & 1/3 \\ 1 & -2/3 \end{pmatrix}}_{B} \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & -1/2 \end{pmatrix}}_{D} \underbrace{\begin{pmatrix} 2/3 & 1/3 \\ 1 & -1 \end{pmatrix}}_{B^{-1}}.$$

Une distribution initiale $(x, 1-x)$ *devient donc après* t *pas de la chaîne :*

$$(x, 1-x)P^t = (x, 1-x)BD^tB^{-1} = \left(\frac{2}{3}, \frac{1}{3}\right) + \left(\frac{-1}{2}\right)^t \left(x - \frac{2}{3}\right)(1, -1).$$

L'écart avec la distribution stationnaire $(2/3, 1/3)$ *est donc divisé par* 2 *à chaque pas de la chaîne : on parle de* convergence exponentielle.

Pour formaliser, on introduit une distance classique sur les distributions :

Définition 4 *La* variation totale *entre deux distributions* μ *et* ν *sur* Ω *est*

$$||\mu - \nu|| := \max_{A \subset \Omega} |\mu(A) - \nu(A)|.$$

Proposition 3 *La variation totale s'écrit encore*

$$||\mu - \nu|| = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)| = \sum_{\mu(x) \geq \nu(x)} \mu(x) - \nu(x).$$

Preuve. Considérons deux distributions μ et ν (courbes rouge et bleu, Fig. 1.3). Leur graphe délimite trois zones :

— $M = \{(x, y) \mid \nu(x) \leq y \leq \mu(x)\}$ (en bleu) ;
— $N = \{(x, y) \mid \mu(x) \leq y \leq \nu(x)\}$ (en rouge) ;
— $I = \{(x, y) \mid y \leq \min(\mu(x), \mu(x))\}$ (en vert).

On a $|M| + |I| = |N| + |I| = 1$, d'où $|M| = |N|$ et donc

$$\frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)| = \frac{|M| + |N|}{2} = |M|.$$

Si $A \subset \Omega$, notons $A_+ := \{x \in A \mid \mu(x) \geq \nu(x)\}$ et $A_- := \{x \in A \mid \mu(x) \leq \nu(x)\}$.

$$\mu(A) - \nu(A) = \underbrace{\underbrace{\mu(A_+) - \nu(A_+)}_{0 \leq \cdot \leq |M|} + \underbrace{\mu(A_-) - \nu(A_-)}_{-|N| \leq \cdot \leq 0}}_{-|N| \leq \cdot \leq |M|}.$$

Donc $|\mu(A) - \nu(A)| \leq |M| = |N|$. La borne est atteinte pour $A = \Omega_+$. □

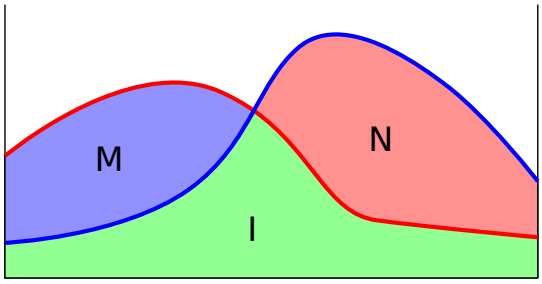

FIGURE 4 – Deux distributions

Soit P une chaîne de Markov irreductible de distribution stationnaire π. La distance entre π et la distribution obtenue à partir d'un état x quelconque après t pas de la chaîne est alors majorée par

$$d(t) := \max_{x \in \Omega} ||P^t(x, \cdot) - \pi||.$$

On dit que la chaîne *converge* si $d(t) \xrightarrow{t\infty} 0$. Les exemples représentés Fig. 5 montrent que ce n'est pas toujours le cas.

Définition 5 *Une chaine de Markov P est dite* apériodique *si*

$$\forall x, y \in \Omega, \quad pgcd(t \mid P^t(x, y) > 0\} = 1.$$

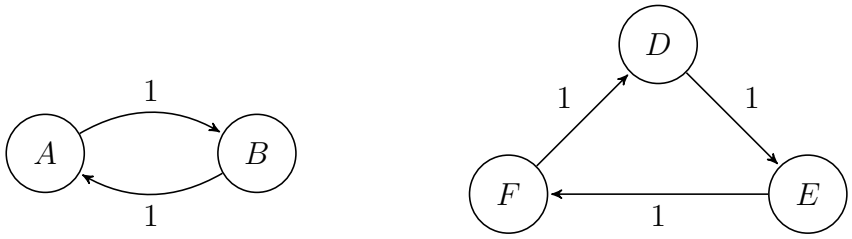

FIGURE 5 – Chaînes non convergentes. Et en combinant via un sommet ?

Un moyen facile de rendre apériodique une chaîne de Markov sans changer sa distribution stationnaire est de rester sur place en chaque état avec probabilité donnée (on parle de chaîne *fainéante*). Un autre moyen est celui de Google : aller dans un état quelconque au hasard avec une petit probabilité. Combinée à l'irréductibilité, l'apériodicité garantit une convergence exponentielle :

Théorème 2 *Si P est une chaîne de Markov irréductible et apériodique, alors il existe $\alpha \in [0,1)$ et $C \geq 0$ tels que, pour tout $t \geq 0$*

$$d(t) \leq C\alpha^t.$$

Dans l'exemple en début de paragraphe, on a $C = 2/3$, $\alpha = 1/2$.

Il y a plusieurs façons de prouver ce théorème :
- Th 4.9 LPW : direct, besoin de rien, mais technique et pas super éclairant ;
- exo 5.1 LPW : couplage (on peut le faire après leçon couplage) ;
- Eq. 4.35 LPW : via résultat sur temps de mélange (couplage caché) ;
- via Perron-Frobenius : résultat non trivial mais classique d'algébre linéaire et preuve qui fait le lien avec le trou spectral.

Théorème 3 (Perron-Frobenius, 1907-1912) *Toute matrice positive irreductible admet un vecteur propre strictement positif associé à une valeur propre simple strictement plus grande (en module) que toutes les autres :*

$$\left\{ \begin{array}{l} A \geq 0 \\ \forall i,j \exists r \mid (A^r)_{ij} > 0 \end{array} \right. \Rightarrow \exists \left\{ \begin{array}{l} \vec{u} > 0 \\ \lambda > 0 \end{array} \right. \mid \left\{ \begin{array}{l} A\vec{u} = \lambda\vec{u} \\ A\vec{v} = \lambda'\vec{v} \Rightarrow |\lambda'| < \lambda \text{ ou } \vec{v} \in \mathbb{R}\vec{u}. \end{array} \right.$$

Preuve. (du Théorème 2, sketch)
- On suppose P diagonalisable (c'est par exemple le cas si la chaîne est symétrique). Soit alors $(\vec{p}_1, \ldots, \vec{p}_n)$ une base formée de vecteurs propres associés aux valeurs propres $\lambda_1 > |\lambda_2| \geq \ldots \geq |\lambda_n|$, avec $\vec{p}_1 > 0$.
- On a $\lambda_1 = 1$ car le rayon spectral d'une matrice stochastique est 1.

— Soit $\vec{p} = \sum_i \alpha_i \vec{p_i}$ un vecteur (une distribution). On calcule $P^t\vec{p} = \sum_i \alpha_i \lambda_i^t \vec{p_i}$, qui tend vers $\alpha_1\vec{p_1}$ car $|\lambda_i| < 1$ pour $i \geq 2$. Donc $\alpha_1 = 1$ sinon on ça ne serait plus une distribution.
— Plus précisément, comme λ_2 est la seconde plus grande valeur propre, il existe $C > 0$ tel que $|P^t\vec{p} - \vec{p_1}| \leq C|\lambda_2|^t$: la borne est liée trou spectral.

□

On a donc toujours une convergence exponentielle vers la distribution stationnaire. C'est bien, mais à Tchernobyl aussi la radioactivité converge exponentiellement vers 0... En outre, la constante α peut être de plus en plus proche de 1 quand la taille du système grandit (on s'intéresse aux grands systèmes en général). Peut-on donc être plus précis ?

1.4 Temps de mélange

Définition 6 *Le* temps de mélange *à ε près est défini par*

$$\tau_{mix}(\varepsilon) := \min\{t \mid d(t) \leq \varepsilon\}.$$

On introduit aussi (de façon arbitraire mais, on verra, sans importance)

$$\tau_{\text{mix}} := \tau_{\text{mix}}\left(\frac{1}{2\text{e}}\right).$$

Le temps de mélange s'avère paramètrer la vitesse de convergence vers la distribution stationnaire (comme un temps de demi-vie d'un élément radioactif) :

Théorème 4

$$d(t) \leq \exp(-t/\tau_{mix}).$$

Preuve. On introduit

$$\overline{d}(t) := \max_{x,y\in\Omega} ||P^t(x,\cdot) - P^t(y,\cdot)||.$$

Montrons

$$d(t) \leq \overline{d}(t) \leq 2d(t).$$

La deuxième inégalité découle directement de l'inégalité triangulaire pour la variation totale. Montrons la première. Pour tout $A \subset \Omega$, on a

$$\pi(A) = \sum_{y\in\Omega} \pi(y)P^t(y,A).$$

On calcule alors

$$\begin{aligned} d(t) = ||P^t(x,\cdot) - \pi|| &= \max_{A\subset\Omega} |P^t(x,A) - \pi(A)| \\ &= \max_{A\subset\Omega} \left| \sum_{y\in\Omega} \pi(y) \left(P^t(x,A) - P^t(y,A)\right) \right| \\ &\leq \max_{A\subset\Omega} \sum_{y\in\Omega} \pi(y) \left|P^t(x,A) - P^t(y,A)\right| \\ &\leq \sum_{y\in\Omega} \pi(y) \max_{A\subset\Omega} \left|P^t(x,A) - P^t(y,A)\right| \\ &= \sum_{y\in\Omega} \pi(y) ||P^t(x,\cdot) - P^t(y,\cdot)|| \\ &\leq \max_{y\in\Omega} \pi(y) ||P^t(x,\cdot) - P^t(y,\cdot)|| = \overline{d}(t). \end{aligned}$$

Par ailleurs, on montre (argument non-trivial de couplage optimal, en exercice ou admis) que $\overline{d}$ est sous-multiplicative :

$$\overline{d}(s+t) \leq \overline{d}(s)\overline{d}(t).$$

On en déduit

$$d(k\tau_{\text{mix}}) \leq \overline{d}(k\tau_{\text{mix}}) \leq \overline{d}(\tau_{\text{mix}})^k \leq (2d(\tau_{\text{mix}}))^k \leq e^{-k}.$$

Le résulat en découle en prenant $k = t/\tau_{\text{mix}}$. □

On retrouve le théorème de convergence exponentielle (Th. 2). On en déduit aussi que $d(t) \leq \varepsilon$ dès que $t \geq \tau_{\text{mix}} \log(\varepsilon)$, d'où

$$\tau_{\text{mix}}(\varepsilon) \leq \tau_{\text{mix}} \lceil -\log(\varepsilon) \rceil,$$

ce qui montre que choix de ε pour définir τ_{mix} n'a donc que peu d'importance.

Le problème devient alors de trouver un asymptotique de τ_{mix} quand la taille du système tend vers l'infini. En particulier, on est content si on peut montrer l'existence d'un *cutoff* :

Définition 7 *Une chaîne de Markov admet un* cutoff *si*

$$\lim_{|\Omega|\to\infty} d(c\tau_{mix}) = \begin{cases} 1 & si\ c < 1, \\ 0 & si\ c > 1. \end{cases}$$

Une chaîne de Markov qui a un cutoff admet donc une transition de phase : le mélange passe rapidement de quasi-inexistant à quasi-parfait autour du temps de mélange. Ce n'est cependant pas le cas de toutes les chaînes.

2 Mélange américain (Riffle shuffle)

2.1 Expérimentation

Répéter la procédure suivante en incrémentant à chaque fois t à partir de $t = 1$:

1. trier le jeu de carte par ordre croissant ou décroissant (non divulgué) ;
2. faire t passes de mélange américain ;
3. laisser les élèves examiner le jeu obtenu ;
4. les sonder sur l'ordre initial du paquet (croissant ou décroissant).

Jusqu'à ce que les avis soient à peu près partagés (ou que personne n'ait d'avis). L'évolution du pourcentage d'avis majoritaire est censée refléter celle de $d(t)\ldots$

2.2 Modélisation

Pour pouvoir montrer quelque chose, il faut un modèle mathématique. On modélise le mélange américain par la chaîne de Markov définie comme suit.

Définition 8 (Mélange américain) *Tirer M tel que $\mathbb{P}(M = k) = C_n^k/2^n$ (nombre de "pile" sur n lancers*[1]*). Couper en deux paquets (les M premières et les $n-M$ dernières cartes). Entrelacer en faisant tomber les cartes, la probabilité qu'une carte tombe d'un paquet étant proportionnelle à la taille de ce paquet.*

Proposition 4 *Si la coupe est en M, alors le mélange américain choisit uniformément au hasard un des C_n^M entrelacements possibles.*

Preuve. Considérons un entrelacement donné. Si p est la taille du premier paquet (et donc $k-p$ celle du deuxième paquet) quand on lâche la k-ème carte, alors c'est la bonne carte avec probabilité $(M-p)/(n-k)$ si elle doit venir du premier paquet, $(n-M-(k-p))/(n-k)$ sinon. Au final, la probabilité d'obtenir le bon entrelacement est donc la fraction dont

— le dénominateur est le produit de $n, n-1, \ldots, 1$ (on lâche toutes les cartes) ;
— le numérateur est le produit de $M, M-1, \ldots, 1$ (on lâche toutes les cartes du premier paquet) et de $n-M, n-M-1, \ldots, 1$ (on lâche toutes les cartes du second paquet).

C'est donc $M!(n-M)!/n! = 1/C_n^M$: les entrelacements sont équiprobables. □

On appelle *séquence montante* d'une permutation σ une suite maximale $i_1, \ldots, i_k$ d'indices tels que $\sigma(i_{j+1}) = \sigma(i_j) + 1$ pour $1 \leq j < k$.

1. Loi binômiale $B(n, 1/2)$. Rappel : $M \sim B(n,p)$ quand $\mathbb{P}(M = k) = C_n^k p^k (1-p)^k$.

Proposition 5 *Le mélange américain a pour matrice*

$$P(\sigma,\sigma') = \begin{cases} (n+1)/2^n & \text{si } \sigma' = \sigma, \\ 1/2^n & \text{si } \sigma' \circ \sigma^{-1} \text{ a exactement deux séquences montantes,} \\ 0 & \text{sinon.} \end{cases}$$

Preuve. Sans restriction de généralité, σ est la permutation identité.
Si σ' a exactement deux séquences montantes, chacune provient d'un paquet et la longueur k de la coupe est donc caractérisée. Pour aller en σ' il a donc fallu couper en $M = k$ (probabilité $C_n^k/2^n$) et choisir le bon entrelacement (probabilité $1/C_n^k$), d'où une probabilité totale de $1/2^n$.
Si σ' est l'identité, c'est que les deux paquets ont juste été remis l'un sur l'autre après la coupe. On a donc pu couper pour n'importe quel k dans $\{0,\dots,n\}$ (probabilité $C_n^k/2^n$ à chaque fois) puis il a fallu choisir le bon "entrelacement", *i.e.* celui qui n'entrelace rien du tout (probabilité $1/C_n^k$), d'où une probabilité totale de $(n+1)/2^n$.
Enfin, σ' ne peut clairement pas avoir plus de deux séquences montantes. □

Proposition 6 *Le mélange américain est irréductible, apériodique et sa distribution stationnaire est uniforme.*

Preuve. La chaîne est apériodique car la probabilité de ne rien changer est non nulle. La distribution stationnaire est uniforme car $P(\sigma,\sigma') = P(\sigma',\sigma)$ (vérification sur la formule). Pour montrer l'irréductibilité, on montre par induction qu'on peut trier une permutation σ quelconque. Supposons qu'au dessus du paquet on ait, dans l'ordre, les cartes 1 à k. On repère alors la carte $k+1$ dans le paquet, on coupe juste au dessus et on choisit un entrelacement qui garde les cartes 1 à k (haut du premier paquet), met la carte $k+1$ (haut du deuxième paquet) puis met le reste dans n'importe quel ordre. On a alors les cartes 1 à $k+1$ au dessus du paquet, qui est donc trié quand $k+1 = n$. □

Question : combien au plus de passes du mélange américain suffisent à passer de n'importe quel jeu donné à n'importe quel autre donné (diamètre du graphe) ? Indication : montrer qu'on peut diviser par deux ($k \to \lceil k/2 \rceil$) le nombre de *séquences croissantes* (séquence de cartes consécutives de valeur croissante).

2.3 Temps stationnaire fort

Un *temps d'arrêt* pour une chaîne de Markov (X_t) est une variable aléatoire T telle que l'évènement $\{T = t\}$ ne dépende que de $X_0,\dots,X_t$ (et éventuellement

des tirages aléatoires qui ont servi à aller de X_0 à X_t).

Définition 9 *Un* temps stationnaire fort *pour une chaîne de Markov (X_t) de distribution stationnaire π est un temps d'arrêt T tel que*

$$\forall x, \quad \mathbb{P}(X_t = x | t \geq T) = \pi(x).$$

Exemple 6 *Considérons la chaîne de Markov sur $\Omega = \{0,1\}^n$ qui, à chaque pas, tire uniformément au hasard une coordonnée et la met à 0 ou 1 selon un lancer de pièce (marche aléatoire sur l'hypercube). Alors le temps T pour que toutes les coordonnées aient été mises à jour est un temps stationnaire fort.*

Proposition 7 *Si T est un temps stationnaire fort, alors*

$$d(t) \leq \mathbb{P}(t < T).$$

Preuve. Soit T_x le temps stationnaire fort de la chaîne initialement en x.

$$\begin{aligned} P^t(x,A) &= \mathbb{P}(X_t \in A) \\ &= \mathbb{P}(X_t \in A, T_x > t) + \mathbb{P}(X_t \in A, T_x \leq t) \\ &= \mathbb{P}(X_t \in A | T_x > t)\mathbb{P}(T_x > t) + \mathbb{P}(X_t \in A | T_x \leq t)\mathbb{P}(T_x \leq t) \\ &= \mathbb{P}(X_t \in A | T_x > t)\mathbb{P}(T_x > t) + \pi(A)(1 - \mathbb{P}(T_x > t)) \\ &= \pi(A) + \mathbb{P}(T_x > t)\left(\mathbb{P}(X_t \in A | T_x > t) - \pi(A)\right). \end{aligned}$$

Comme la différence de deux probabilité est toujours dans $[-1,1]$, on en déduit

$$|P^t(x,A) - \pi(A)| = \mathbb{P}(T_x > t)\,|\mathbb{P}(X_t \in A | T_x > t) - \pi(A)| \leq \mathbb{P}(T_x > t).$$

Le résultat en découle en prenant le maximum sur x puis sur $A \subset \Omega$. □

Exemple 7 *On reprend l'exemple de la marche aléatoire sur l'hypercube. On veut borner $\mathbb{P}(t < T)$. C'est en fait le problème dit du* collectionneur de coupon. *Soit T_k le temps pour avoir mis à jour k coordonnées. On a*

$$T = T_n = T_1 + (T_2 - T_1) + \ldots + (T_n - T_{n-1}).$$

Or $T_{k+1} - T_k$ suit une loi géométrique[2] *de paramètre $(n-k)/n$. Son espérance est donc $n/(n-k)$. Celle de T_1 est trivialement 1. Par linéarité de l'espérance :*

$$\mathbb{E}(T) = 1 + \sum_{k=1}^{n-1} \frac{n}{n-k} = n \sum_{k=1}^{n} \frac{1}{k} \leq n \ln(n).$$

2. Rappel : $X \sim G(p)$ quand $\mathbb{P}(X = k) = (1-p)^{k-1}p$, *i.e.*, succès au k-ème essai. On a $\mathbb{E}(X) = 1/p$.

On borne alors $\mathbb{P}(t < T)$ via l'inégalité de Markov :

$$\mathbb{P}(t < T) \leq \frac{\mathbb{E}(T)}{t}.$$

La proposition 7 permet alors de borner le temps de mélange :

$$\tau_{mix} \leq 2en\ln(n).$$

2.4 Mélange inversé

Revenons au mélange américain. On va en fait étudier son *inverse* :

Définition 10 (Mélange américain inversé) *Associer à chacune des cartes un lancer de pièce. Faire un paquet des cartes "pile", un des cartes "faces". Rassembler en un seul paquet en mettant le paquet "pile" sous le paquet "face".*

C'est la chaîne inverse de celle du mélange américain : elle commence par "désentrelacer" en deux paquets, puis à "dé-couper" en superposant ces paquets. Sa matrice est

$$\widehat{P}(\sigma, \sigma') = P(\sigma^{-1}, \sigma'^{-1})$$

Irréductibilité, apériodicité, distribution stationnaire et temps de mélange sont préservés car tout est symétrique via $\sigma \to \sigma^{-1}$.

Définition 11 *On appelle* histoire *d'une carte dans la séquence de pile ou face qui ont déterminé dans quel paquet elle allait à chaque étape.*

Exemple 8 *On lance la chaîne sur une séquence "abcdefgh" de huit cartes. Chaque ligne représente un pas de la chaîne, avec en majuscule les cartes labellées "pile" pour le pas suivant.*

A	*b*	*C*	*D*	*e*	*f*	*g*	*h*
a	*c*	*d*	*B*	*E*	*f*	*G*	*H*
B	*E*	*g*	*h*	*A*	*c*	*D*	*f*
b	*E*	*A*	*d*	*g*	*H*	*C*	*F*

Après ce quatrième pas, les cartes ont toutes des histoires différentes :

AaAA bBBb CccC DdDd eEEE fffF gGgg hHhH

La permutation obtenue, "eahcfbdg", se retrouve en remontant les histoires :

1. *La dernière lettre montre que a,c,e,f,h ont été placées au dessus de b,g,d. On note acefh/bgd.*

2. *La troisième lettre donne ae/cfh et bd/g. Donc ae/cfh/bd/g.*
3. *La deuxième lettre donne e/a, h/cf et b/d. Donc e/a/h/cf/b/d/g.*
4. *La première lettre donne enfin c/f.*

Proposition 8 *Le temps T auquel toutes les cartes du paquet ont une histoire différente est un temps stationnaire fort.*

Preuve. Au temps T, la permutation est caractérisée par les histoires des cartes. Comm les histoires sont équiprobables, les permutations au temps T aussi. □

Proposition 9 *Le mélange américain vérifie pour n assez grand*

$$\tau_{mix} \leq 2\log_2(5n/3).$$

Preuve. Pour $t \geq T$, chaque carte a sa propre histoire parmi les 2^t histoires :

$$\mathbb{P}(t \geq T) = \prod_{k=0}^{n-1} \left(1 - \frac{k}{2^t}\right).$$

Soit c tel que $2^t = n^2/c^2$. On utilise $\log(1+x) = x + O(x^2)$ pour calculer

$$\begin{aligned}
\log \mathbb{P}(t \geq T) &= \sum_{k=0}^{n-1} \log\left(1 - \frac{k}{2^t}\right) \\
&= -\sum_{k=0}^{n-1} \left(\frac{c^2 k}{n^2} + O\left(\frac{c^4 k^2}{n^4}\right)\right) \\
&= -\frac{c^2 n(n-1)}{2n^2} + O\left(\frac{c^4 n(n+1)(2n+1)}{6n^4}\right). \\
&= -\frac{c^2}{2} + O\left(\frac{1}{n}\right).
\end{aligned}$$

Or, d'après la proposition 7 :

$$d(t) \leq \mathbb{P}(t < T) = 1 - \mathbb{P}(t \geq T).$$

Donc pour c tel que $1 - \exp(-c^2/2) < 1/2\mathrm{e}$, τ_{mix} est borné par $t = 2\log_2(n/c)$ pour n assez grand. Avec $c = 3/5$ on obtient la borne annoncée. □

On peut par ailleurs assez facilement obtenir une borne inférieure :

Proposition 10 *Soit $\delta > 0$. Le mélange américain vérifie pour n assez grand*

$$\tau_{mix} \geq (1-\delta)\log_2(n).$$

Preuve. Soit Ω_x^t l'ensemble des permutations atteignables à partir de x par au plus t pas de la chaîne. Comme à chaque pas de la chaîne on peut atteindre au plus 2^n nouvelles permutations (selon les n lancers de pièce), on a $|\Omega_x^t| \leq 2^{nt}$. D'où :

$$d(t) = \max_{x\in\Omega} ||P^t(x,\cdot) - \pi|| \geq P^t(x,\Omega_x^t) - \pi(\Omega_x^t) = 1 - \frac{|\Omega_x^t|}{|\Omega|} \geq 1 - \frac{2^{nt}}{n!}.$$

Donc $d(t) > \frac{1}{2e}$, *i.e.*, $t < \tau_{\text{mix}}$, tant que $1 - \frac{2^{nt}}{n!} > \frac{1}{2e}$. On en déduit la borne annoncée en utilisant la formule de Stirling, qui donne $\ln(n!) = (1+o(1))\ln(n)$. □

On peut en fait être plus précis :

Théorème 5 (Bayer-Diaconis, 1992) *Il y a cutoff à $\tau_{mix} = \Theta(\frac{3}{2}\log_2(n))$.*

Bayer et Diaconis ont également mené à bout le calcul explicite de $d(t)$ pour $n = 52$, ce qui est non trivial vu la taille de l'espace (8×10^{67} jeux différents). Ils obtiennent (voir aussi Fig. 2.4)

$t \leq 4$	5	6	7	8	9	10	11	12
1.000	.9237	.6135	.3341	.1672	.0854	.0429	.0215	.0108

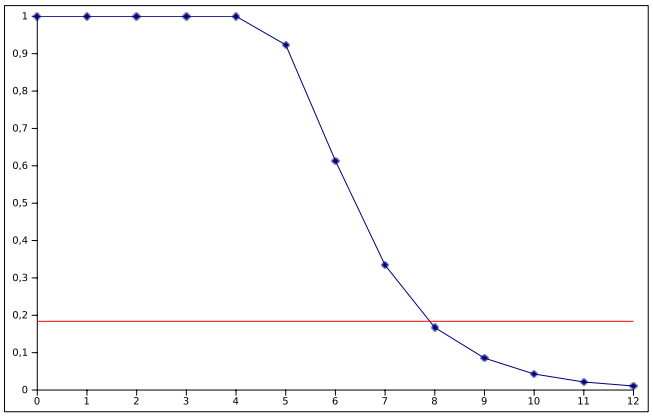

FIGURE 6 – Mélange américain de 52 de cartes : graphe de $t \to d(t)$.

3 Couplage

3.1 Couplage avant

Un point obtenu en faisant évoluer un certain temps une chaîne de Markov (trajectoire) est distribué selon la distribution stationnaire si cette trajectoire a "oublié" son point de départ, *i.e.*, le biais initial a disparu. L'idée fondamentale du *couplage* est qu'une façon d'être sûr d'avoir "oublié" le point de départ d'une trajectoire est de considérer en fait *deux* trajectoires : au moment où elles se recontrent, on ne peut plus savoir laquelle est partie de quel point. Formalisons.

Définition 12 *Un* couplage *d'une chaîne de Markov P est une paire (X_t, Y_t) de processus aléatoires suivant chacun la loi P.*

Ici encore, on peut corréler à dessein X_t et Y_t.

Définition 13 *Le* temps de coalescence T *d'un couplage (X_t, Y_t) d'une chaîne de Markov est la variable aléatoire définie par*

$$T = \max_{x,y\in\Omega} \min\{t \ : \ X_t = Y_t \mid (X_0, Y_0) = (x, y)\}.$$

Proposition 11 *Si T est un temps de coalescence, alors*

$$d(t) \leq \mathbb{P}(t < T).$$

Preuve. Si (X, Y) est un couplage de μ et ν, alors

$$||\mu - \nu|| \leq \mathbb{P}(X \neq Y)$$

Soit en effet $A \subset \Omega$.

$$\begin{aligned} \mu(A) - \nu(A) &= \mathbb{P}(X \in A) - \mathbb{P}(Y \in A) \\ &= \mathbb{P}(X \in A,\ Y \notin A) + \mathbb{P}(X \in A,\ Y \in A) - \mathbb{P}(Y \in A) \\ &\leq \mathbb{P}(X \in A,\ Y \notin A) \\ &\leq \mathbb{P}(X \neq Y). \end{aligned}$$

En particulier

$$||P^t(x, \cdot) - P^t(y, \cdot)|| \leq \mathbb{P}(X_t \neq Y_t \mid X_0 = x,\ Y_0 = y) \leq \mathbb{P}(t < T)$$

En prenant le maximum sur x et y on majore $\overline{d}(t)$, puis $d(t) \leq \overline{d}(t)$. □

Exemple 9 *Considérons encore la marche aléatoire sur l'hypercube* $\{0,1\}^n$. *Un exemple de couplage* (X_t, Y_t) *est le suivant : à chaque étape, une coordonnée est choisie uniformément au hasard, puis on lance une pièce pour mettre à* 0 *ou* 1 *cette coordonnée dans* X_t *et* Y_t *(même valeur dans les deux). Quand les coordonnées ont été mises à jour au moins une fois, alors* $X_t = Y_t$, i.e., *les deux trajectoires ont coalescé. Ce temps d'arrêt majore donc le temps de coalescence, et la proposition précédente permet de retrouver* $\tau_{mix} \leq 2e\log(n)$.

On a vu deux deux bornes sur $d(t)$ du type $\mathbb{P}(t < T)$ où T est un temps d'arrêt : dans un cas T est un temps stationnaire fort (Prop. 7), dans l'autre c'est un temps de coalescence (Prop. 11). De plus, l'exemple de la marche aléatoire sur l'hypercube exhibe un temps d'arrêt qui est à la fois un temps stationnaire fort et un temps de coalescence. D'où la question naturelle : un temps de coalescence est-il toujours un temps stationnaire fort ? L'exemple de la chaîne de Markov représentée Fig. 1 montre qu'il n'en est rien (la coalescence a toujours lieu en A, or la distributions stationnaire donne un poids positive à B). On va voir cependant comment remédier à ce biais via l'algorithme de *couplage arrière*.

3.2 Couplage arrière

Principe général (Propp-Wilson, 1996) :

- fixer un pas de temps t ;
- faire évoluer tous les états en parallèle du temps $-t$ au temps 0 (t pas) ;
- si deux états coalescent, les faire évoluer identiquement ensuite ;
- si il y a eu coalescence de tous les états, renvoyer l'état obtenu ;
- sinon, doubler t et recommencer *avec le même hasard de* $-t$ *à* 0.

Reprendre l'exemple de la chaîne de Markov représentée Fig. 1 pour bien faire comprendre cette histoire de "même hasard".

A priori pas très intéressant du point de vue algorithmique (Ω est grand !), sauf si on a un moyen de déterminer un petit nombre de points dont la coalescence garantit la coalescence de tous les points. Le cas le plus simple est celui d'un couplage *monotone*.

Définition 14 *Un couplage* (X_t, Y_t) *est dit* monotone *pour l'ordre partiel* $\preceq$ *si*

$$X_t \preceq Y_t \ \Rightarrow X_{t+1} \preceq Y_{t+1}.$$

La coalescence des maxima et minima assure alors la coalescence générale.

Exemple 10 *Dyck paths*

Exemple 11 *dimer tilings*

On peut imaginer que la coalescence générale soit garantie par d'autre propriétés. Par exemple, si on a une métrique sur Ω et un couplage qui préserve la convexité, alors la coalescence des points extrémaux entraîne la coalescence générale.

3.3 Contractions

comment borner le temps de coalescence ? un principe : contraction (Wilson et/ou Randall sur Dyck path)

Exercices

après premier cours : donner quelques chaines (graphes), demander : irréductible ? apériodique ? distrib statio ?
après deuxième : strong stationnary time for Top-in-random shuffle
après troisième : coupling for Top-in-random shuffle
Rubik's cube (pb ouvert a priori)

3.4 Couplage de distributions (SAUTER)

Définition 15 *Un* couplage *de deux distributions μ et ν est une paire (X, Y) de variables aléatoires suivant respectivement les lois μ et ν.*

L'intérêt est que ces variables aléatoires peuvent être corrélées.

Exemple 12 *Soit $\mu = \nu$ la distribution d'un lancer de pièce équilibré. Voici deux couplages (X, Y) de μ et ν :*

1. *X et Y sont tirées indépendamment avec une pièce équilibrée ;*
2. *on tire X avec une pièce équilibrée puis on pose $Y = X$.*

Pour le premier couplage, $\mathbb{P}(X = x,\ Y = y) = 1/4$. Pour le second couplage, $\mathbb{P}(X = Y = \textit{pile}) = \mathbb{P}(X = Y = \textit{face}) = 1/2$ et $\mathbb{P}(X \neq Y) = 0$.

Proposition 12 (Coupling lemma) *Pour deux distributions μ et ν sur Ω :*

$$||\mu - \nu|| = \min\{\mathbb{P}(X \neq Y) \mid (X, Y) \text{ est un couplage de } \mu \text{ et } \nu\}.$$

Preuve. La majoration est facile. Soit en effet $A \subset \Omega$.

$$\begin{aligned}
\mu(A) - \nu(A) &= \mathbb{P}(X \in A) - \mathbb{P}(Y \in A) \\
&= \mathbb{P}(X \in A,\ Y \notin A) + \mathbb{P}(X \in A,\ Y \in A) - \mathbb{P}(Y \in A) \\
&\leq \mathbb{P}(X \in A,\ Y \notin A) \\
&\leq \mathbb{P}(X \neq Y).
\end{aligned}$$

Montrons maintenant que la borne est atteinte. Considérons à nouveau la figure 1.3 (page 5). Tirer X (resp. Y) selon la loi μ, c'est tirer uniformément un point dans la zone $M \cup I$ (resp. $N \cup I$). On construit un couplage alors un couple (X, Y) de (μ, ν) tel que $X = Y$ aussi souvent que possible :

- on choisit d'abord avec probabilité $|I|$ si on va tirer X et Y dans I ;
- si ce n'est pas le cas, alors on tire X dans M et Y dans N (donc $X \neq Y$) ;
- si c'est le cas, alors plutôt que de tirer X et Y indépendamment dans I, on tire X et on pose $Y = X$.

Définis ainsi, X et Y suivent bien respectivement les lois μ et ν. Mais surtout :

$$\mathbb{P}(X \neq Y) = 1 - |I| = |M| = ||\mu - \nu||.$$

□

On a introduit la définition suivante dans la preuve du théorème 4 (page 7) :

$$\overline{d}(t) := \max_{x,y \in \Omega} ||P^t(x, \cdot) - P^t(y, \cdot)||.$$

Montrons ici le résultat qui avait alors été admis :

Proposition 13 *La fonction $\overline{d}$ est sous-multiplicative :*

$$\overline{d}(s+t) \leq \overline{d}(s)\overline{d}(t).$$

Preuve. Soit $s > t > 0$. Soit x et y dans Ω. La proposition 12 assure qu'il existe un couplage (X_t, Y_t) de $P^t(x, \cdot\,)$ et $P^t(y, \cdot\,)$ tel que

$$||P^t(x, \cdot\,) - P^t(y, \cdot\,)|| = \mathbb{P}(X_t \neq Y_t).$$

On construit un couplage (X_{t+s}, Y_{t+s}) de $P^{t+s}(x, \cdot\,)$ et $P^{t+s}(y, \cdot\,)$ comme suit :

- si $X_t = Y_t = x'$, alors on pose $X_{t+s} = Y_{t+s} = x'$;
- si $X_t = x'$ et $Y_t = y' \neq x'$, alors la proposition 12 assure qu'il existe un couplage tel que

$$\mathbb{P}(X_{t+s} \neq Y_{t+s} | X_t = x',\ Y_t = y') = ||P^s(x', \cdot\,) - P^s(y', \cdot\,)||.$$

On calcule alors

$$\begin{aligned} ||P^{t+s}(x, \cdot\,) - P^{t+s}(y, \cdot\,)|| &\leq \mathbb{P}(X_{t+s} \neq Y_{t+s}) \\ &= \mathbb{P}(X_{t+s} \neq Y_{t+s} | X_t = x',\ Y_t = y') \times \mathbb{P}(X_t \neq Y_t) \\ &\leq ||P^s(x', \cdot\,) - P^s(y', \cdot\,)|| \times ||P^t(x, \cdot\,) - P^t(y, \cdot\,)||. \end{aligned}$$

Le résultat en découle en prenant le maximum sur x, y, x' et y'. □

Preuve de Sinclair. Preuve de LPW plus convaincante