

(public 2009)

Résumé : On étudie un modèle de battage de cartes. On cherche en particulier à déterminer le plus précisément possible le nombre de battages nécessaires pour obtenir un mélange satisfaisant du paquet.

Mots clés : Chaîne de Markov, convergence vers une loi stationnaire, groupe symétrique.

- *Il est rappelé que le jury n'exige pas une compréhension exhaustive du texte. Vous êtes laissé(e) libre d'organiser votre discussion comme vous l'entendez. Des suggestions de développement, largement indépendantes les unes des autres, vous sont proposées en fin de texte. Vous n'êtes pas tenu(e) de les suivre. Il vous est conseillé de mettre en lumière vos connaissances à partir du fil conducteur constitué par le texte. Le jury appréciera que la discussion soit accompagnée d'exemples traités sur ordinateur.*

1. Le battage par insertion

Considérons N cartes numérotées de C_1 à C_N et disposées en un paquet sur une table. On appelle première carte du paquet la carte située au sommet de la pile, deuxième carte celle qui se trouve immédiatement en-dessous, jusqu'à la N -ième carte qui est celle située au bas de la pile. On prendra garde de bien distinguer la position d'une carte dans le paquet du numéro qu'elle porte. Soit k un entier compris entre 1 et N . On appelle *insertion à la k -ième place* l'opération qui consiste à prendre la première carte du paquet et à l'insérer entre la k -ième et la $k + 1$ -ième carte. Une insertion à la première place ne change pas l'ordre des cartes. Une insertion à la N -ième place consiste à faire glisser la première carte sous le paquet.

Le *battage par insertion* du jeu de cartes consiste à effectuer une suite d'insertions aléatoires, en choisissant à chaque étape au hasard uniformément dans $\{1, \dots, N\}$ la place à laquelle l'insertion a lieu, indépendamment des insertions précédentes.

On représente à chaque étape la configuration du jeu de cartes par l'unique permutation σ de l'ensemble $\{1, \dots, N\}$ telle que pour tout i entre 1 et N , la carte C_i se trouve à la $\sigma(i)$ -ième position. Ainsi, une insertion à la k -ième place fait passer de la configuration σ à la configuration $(k, k-1, \dots, 2, 1)\sigma$, où $(k, k-1, \dots, 2, 1)$ désigne la permutation circulaire qui envoie k sur $k-1$, $k-1$ sur $k-2$, \dots , 2 sur 1 et 1 sur k .

Pour modéliser le battage par insertion du paquet, on considère la chaîne de Markov $(X_n)_{n \geq 0}$ dont l'espace d'états est le groupe symétrique \mathfrak{S}_N et dont les probabilités de transitions sont

données par

(1)

$$\mathbb{P}(X_{n+1} = \sigma' \mid X_n = \sigma) = \begin{cases} \frac{1}{N} & \text{s'il existe } k \in \{1, \dots, N\} \text{ tel que } \sigma' = (k, k-1, \dots, 2, 1)\sigma, \\ 0 & \text{sinon.} \end{cases}$$

Notons π la loi uniforme sur \mathfrak{S}_N . Elle est définie par le fait que pour toute partie A de \mathfrak{S}_N , $\pi(A) = \frac{1}{N!} \text{Card}(A)$.

Théorème 1. *La chaîne de Markov $(X_n)_{n \geq 0}$ est irréductible et apériodique. Elle possède une unique mesure de probabilité invariante sur \mathfrak{S}_N , qui est la mesure uniforme π et vers laquelle elle converge en loi.*

Démonstration. (Esquisse) Pour tous $\sigma, \sigma' \in \mathfrak{S}_N$, notons $p_{\sigma, \sigma'} = \mathbb{P}(X_{n+1} = \sigma' \mid X_n = \sigma)$.

Les permutations $(N, N-1, \dots, 2, 1)$ et $(2, 1)$ engendrent le groupe symétrique \mathfrak{S}_N , si bien qu'il existe pour toute paire de permutations un chemin de probabilité positive reliant l'une à l'autre. De plus, pour toute permutation σ , $p_{\sigma, \sigma} > 0$. Ainsi, la chaîne $(X_n)_{n \geq 0}$ est irréductible et apériodique.

Le fait que S_N soit un groupe entraîne pour tout $\tau \in \mathfrak{S}_N$ l'égalité $\text{Card}\{\sigma \in \mathfrak{S}_N \mid p_{\sigma, \tau} > 0\} = N$. Ainsi, la transposée de la matrice de transition de $(X_n)_{n \geq 0}$ est encore une matrice de transition et la mesure invariante de $(X_n)_{n \geq 0}$ est la loi uniforme. \square

Le problème que rencontre un joueur est de déterminer combien de battages sont nécessaires pour que le paquet soit convenablement mélangé. On peut mesurer la qualité du mélange à un instant donné n en calculant la distance en variation entre la loi de X_n et la loi uniforme π .

Définition 1. *Soient μ et ν deux mesures de probabilités sur \mathfrak{S}_N . On appelle distance en variation entre μ et ν le réel $d_V(\mu, \nu) \in [0, 1]$ défini par*

$$d_V(\mu, \nu) = \max\{|\mu(A) - \nu(A)| : A \subset \mathfrak{S}_N\}.$$

Avec ces notations, on vérifie aisément que $d_V(\mu, \nu) \leq \sum_{\sigma \in \mathfrak{S}_N} |\mu(\{\sigma\}) - \nu(\{\sigma\})|$. On a en fait mieux, puisque

$$(2) \quad d_V(\mu, \nu) = \frac{1}{2} \sum_{\sigma \in \mathfrak{S}_N} |\mu(\{\sigma\}) - \nu(\{\sigma\})|.$$

Pour chaque $n \geq 1$, notons μ_n la loi de X_n . La convergence de la chaîne vers sa mesure d'équilibre signifie que

$$(3) \quad \lim_{n \rightarrow \infty} d_V(\mu_n, \pi) = 0.$$

Le but des prochains paragraphes est d'examiner la vitesse à laquelle cette convergence a lieu afin d'apporter une réponse aussi simple et concrète que possible au problème du joueur de cartes.

2. La remontée des cartes

Supposons qu'à l'instant initial, toutes les cartes soient rangées dans l'ordre, c'est-à-dire que leur valeur coïncide avec leur position. Ceci se traduit par l'égalité $X_0 = \text{id}$. Observons alors, au cours de l'évolution de la chaîne, le mouvement de la carte C_N .

Elle reste tout d'abord au fond du paquet jusqu'à ce qu'à un certain instant T_1 , on glisse une carte sous elle. Elle se trouve alors en position $N - 1$. Un certain temps T_2 s'écoule ensuite jusqu'à ce qu'une autre carte soit glissée sous elle. La nouvelle carte a pu être glissée avec égale probabilité au-dessus ou en-dessous de la carte qui se trouvait en position N . À l'instant $T_1 + T_2$, la carte C_N se trouve donc en position $N - 2$ et les deux cartes qui se trouvent sous elle ont la même probabilité d'être rangées dans chacun des deux ordres possibles. De même, à l'instant $T_1 + T_2 + T_3$, une troisième carte est glissée sous la carte C_N , et les trois cartes qui se trouvent aux positions $N - 2$, $N - 1$ et N ont maintenant la même probabilité $1/3!$ d'être rangées dans chacun des $3!$ ordres possibles. On définit ainsi par récurrence, pour chaque i compris entre 0 et $N - 1$, un temps aléatoire T_i par

$$(4) \quad \begin{aligned} T_0 &= 0, \\ T_i &= \min\{n : X_n(N) = N - i\} - T_{i-1} - T_{i-2} - \dots - T_1, \quad i \geq 1. \end{aligned}$$

Dans cette définition, on a noté $X_n(N)$ l'image de l'entier N par la permutation X_n , qui est donc la position à l'instant n de la carte C_N . Posons

$$(5) \quad T = T_1 + T_2 + \dots + T_{N-1} + 1.$$

Le raisonnement que nous avons esquissé plus haut permet de démontrer le résultat suivant.

Proposition 1. *La permutation aléatoire X_T est indépendante de T et de loi uniforme sur \mathfrak{S}_N . Plus généralement, pour tout entier $m \geq 0$, la permutation aléatoire X_{T+m} est indépendante de T et de loi uniforme sur \mathfrak{S}_N .*

S'il était possible d'arrêter le battage au temps T , on obtiendrait un paquet parfaitement mélangé. C'est cependant impossible sans marquer la carte C_N . En revanche, on peut estimer l'écart entre la loi de X_n et la loi uniforme sur \mathfrak{S}_N en fonction de T .

Proposition 2. *Pour tout entier $n \geq 0$, on a*

$$d_V(\mu_n, \pi) \leq \mathbb{P}(T > n).$$

Démonstration. Soit A une partie de \mathfrak{S}_N . Calculons $\mu_n(A) = \mathbb{P}(X_n \in A)$.

$$\begin{aligned} \mu_n(A) &= \mathbb{P}(X_n \in A, T \leq n) + \mathbb{P}(X_n \in A, T > n) \\ &\leq \sum_{k=0}^n \mathbb{P}(X_n \in A, T = k) + \mathbb{P}(T > n) \\ &= \sum_{k=0}^n \pi(A) \mathbb{P}(T = k) + \mathbb{P}(T > n) \\ &\leq \pi(A) + \mathbb{P}(T > n). \end{aligned}$$

Ainsi, $\mu_n(A) - \pi(A) \leq \mathbb{P}(T > n)$. En appliquant cette inégalité au complémentaire de A , on obtient $|\mu_n(A) - \pi(A)| \leq \mathbb{P}(T > n)$. \square Il faut maintenant étudier la loi de T .

Tout d'abord, les temps T_1, \dots, T_{N-1} sont tous indépendants et, pour chaque i entre 1 et $N - 1$, T_i suit la loi géométrique de paramètre i/N . Autrement dit, pour chaque $n \geq 1$, on a

$$(6) \quad \mathbb{P}(T_i = n) = \frac{i}{N} \left(\frac{N-i}{N} \right)^{n-1}.$$

On en déduit l'espérance de T :

$$(7) \quad \mathbb{E}[T] = N \left(1 + \frac{1}{2} + \dots + \frac{1}{N} \right) \simeq N \log N.$$

Des inégalités comme l'inégalité de Bienaymé-Tchebichev indiquent qu'une variable aléatoire a peu de chances de prendre des valeurs significativement plus grandes que son espérance, si bien que le paquet sera mélangé en un temps de l'ordre de $N \log N$. On peut quantifier cette affirmation en calculant la variance de T et obtenir, pour tout réel $c > 0$, une inégalité de la forme

$$(8) \quad \mathbb{P}(T > N \log N + cN) \leq \frac{K}{c^2},$$

où K est une constante connue. Cependant, des simulations numériques montrent une décroissance très rapide de $\mathbb{P}(T > n)$ lorsque n franchit la valeur $N \log N$. L'écart $d_V(\mu_n, \pi)$ doit donc décroître de la même façon.

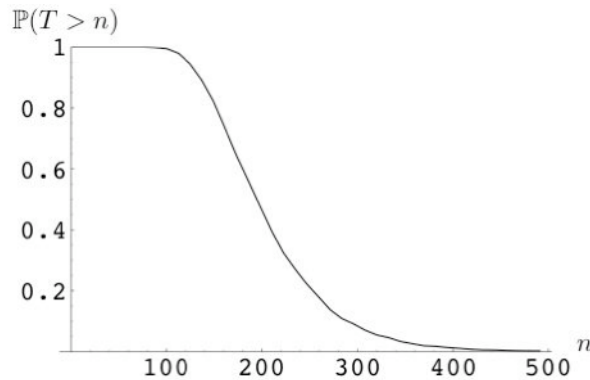


FIG. 1. Une simulation numérique de $\mathbb{P}(T > n)$ en fonction de n lorsque $N = 52$. On observe un saut brutal autour de $N \log N \simeq 205$.

Nous allons démontrer le résultat suivant, qui rend compte de ce comportement.

Théorème 2. Pour tout réel positif c ,

$$d_V(\mu_{N \log N + cN}, \pi) \leq e^{-c}.$$

3. Le collectionneur

Pour démontrer le théorème 2, nous allons interpréter différemment la loi de T .

Imaginons une personne qui collectionne les timbres. Elle reçoit chaque jour une lettre affranchie avec un timbre choisi au hasard uniformément parmi les N timbres en vigueur. Au bout de combien de temps cette personne aura-t-elle une collection complète, c'est-à-dire au moins un exemplaire de chacun des N timbres en vigueur ?

Appelons S ce temps aléatoire. Il s'écrit naturellement comme une somme de temps indépendants $S = S_1 + S_2 + \dots + S_N$, où S_i est le temps que le collectionneur doit attendre pour que le nombre de timbres différents qu'il possède passe de $i - 1$ à i .

Proposition 3. *Les variables $S_i, i = 1 \dots N$ sont indépendantes et pour chaque i entre 1 et N , S_i suit la loi géométrique de paramètre $\frac{N-i+1}{N}$. Ainsi, S a la même loi que T .*

L'avantage de cette description de la loi de T est qu'il est maintenant plus aisé de démontrer le résultat suivant.

Proposition 4. *Pour tout entier $m \geq 1$,*

$$\mathbb{P}(T > m) = \mathbb{P}(S > m) \leq N \left(1 - \frac{1}{N}\right)^m \leq N e^{-\frac{m}{N}}.$$

Démonstration. Supposons que les N timbres en vigueur soient numérotés de 1 à N . Pour tout j compris entre 1 et N , appelons B_j l'événement "Le jour m , le collectionneur n'a toujours pas reçu de lettre affranchie avec le timbre numéro j ". Alors

$$\mathbb{P}(S > m) = \mathbb{P}\left(\bigcup_{j=1}^N B_j\right) \leq \sum_{j=1}^N \mathbb{P}(B_j).$$

Comme $\mathbb{P}(B_j) = \left(1 - \frac{1}{N}\right)^m$ pour tout j , on a le résultat. □

Le théorème 2 découle maintenant de la proposition 2 et de la proposition 4 appliquée avec $m = N \log N + cN$.

Nous avons finalement abouti à une estimation simple qui permet de répondre à la question initiale de façon quantitative. Si par exemple on estime qu'une distance en variation de 0.2 avec la loi uniforme est acceptable et si $N = 52$, on sait qu'il suffit de battre le jeu 290 fois selon le procédé qu'on a considéré ici.

4. Minoration du nombre de battages nécessaires

Le résultat que nous avons obtenu ne répond cependant que partiellement à la question que se pose un joueur. Il veut en effet être sûr que son jeu soit convenablement mélangé, mais il souhaite également passer le moins de temps possible à le battre. La simulation numérique semble indiquer que $\mathbb{P}(T > n)$ reste très proche de 1 jusqu'à un nombre de battages proche de $N \log N$. Nous allons donner un énoncé rigoureux qui rend compte de ce phénomène.

Pour tout entier j entre 2 et N , définissons une partie A_j de \mathfrak{S}_N comme suit :

$$A_j = \{\sigma \in \mathfrak{S}_N : \sigma(N-j+1) < \sigma(N-j+2) < \dots < \sigma(N)\}.$$

Autrement dit, A_j est l'ensemble des configurations du paquet où les j cartes C_{N-j+1}, \dots, C_N , qui étaient initialement les j dernières cartes du paquet, sont restées dans leur ordre relatif initial.

Si pour un n donné la probabilité $\mathbb{P}(X_n \in A_j)$ est proche de 1, le paquet n'est pas convenablement mélangé au temps n , et ce d'autant plus que j est grand. En effet, on a pour tout $n \geq 0$ l'inégalité

$$(9) \quad d_V(\mu_n, \pi) \geq |\mathbb{P}(X_n \in A_j) - \pi(A_j)| \geq \mathbb{P}(X_n \in A_j) - \frac{1}{j!}.$$

Pour montrer, ce qui est notre but, qu'à un temps n voisin mais légèrement inférieur à $N \log N$, le paquet est encore mal mélangé, il suffit de minorer la probabilité $\mathbb{P}(X_n \in A_j)$ par une probabilité assez élevée.

Pour cela, observons, de façon similaire à ce que nous avons fait au paragraphe 3, le mouvement de la carte C_{N-j+1} au cours du battage. Notons R_j le premier instant auquel cette carte se trouve au sommet du paquet, c'est-à-dire que $R_j = \min\{n \geq 0 : X_n(N-j+1) = 1\}$. Alors, au moins jusqu'à l'instant R_j , l'ordre relatif des j cartes qui se trouvaient initialement au fond du paquet n'a pas été modifié et le paquet se trouve dans une configuration qui appartient à A_j . La proposition suivante formalise cette observation ainsi qu'une autre qui va nous permettre de conclure. Notons que le temps T défini au paragraphe 3 n'est autre que $R_1 + 1$.

Proposition 5. 1. R_j a la même loi que $T_j + T_{j+1} + \dots + T_{N-1}$.

2. On a l'inclusion suivante d'événements :

$$\{n \leq R_j\} \subset \{X_n \in A_j\}.$$

En particulier, $\mathbb{P}(X_n \in A_j) \geq \mathbb{P}(R_j \geq n)$.

Nous pouvons maintenant démontrer le résultat suivant.

Théorème 3. Soit $(c_N)_{N \geq 1}$ une suite de réels positifs telle que $\lim_{N \rightarrow \infty} c_N = +\infty$ et telle que pour tout $N \geq 1$, $c_N N$ soit inférieur à $N \log N$. Alors

$$\lim_{N \rightarrow \infty} d_V(\mu_{N \log N - c_N N}, \pi) = 1.$$

Démonstration. Soit j un entier fixé. La première partie de la proposition 5 permet d'établir, lorsque N tend vers l'infini, les équivalents suivants :

$$\mathbb{E}[R_j] \sim N \log N, \quad \text{Var}[R_j] \sim C(j)N^2,$$

où $C(j)$ est une constante qui ne dépend que de j .

Lorsque N est assez grand, on peut appliquer l'inégalité de Bienaymé-Tchebichev pour obtenir

$$\mathbb{P}(R_j \leq N \log N - c_N N) \leq \frac{C(j)N^2}{c_N^2 N^2} = \frac{C(j)}{c_N^2}.$$

Lorsque N tend vers l'infini, cette dernière quantité tend vers 0, si bien que

$$\lim_{N \rightarrow \infty} \mathbb{P}(R_j \geq N \log N - c_N N) = 1.$$

D'après la seconde partie de la proposition 5, on en déduit que $\mathbb{P}(X_{N \log N - c_N N} \in A_j)$ tend vers 1 lorsque N tend vers l'infini. Enfin, l'équation (9) entraîne

$$\lim_{N \rightarrow \infty} d_V(\mu_{N \log N - c_N N}, \pi) \geq 1 - \frac{1}{j!}.$$

On conclut maintenant en faisant tendre j vers l'infini. □

Suggestions pour le développement

- ▶ *Soulignons qu'il s'agit d'un menu à la carte et que vous pouvez choisir d'étudier certains points, pas tous, pas nécessairement dans l'ordre, et de façon plus ou moins fouillée. Vous pouvez aussi vous poser d'autres questions que celles indiquées plus bas. Il est très vivement souhaité que vos investigations comportent une partie traitée sur ordinateur et, si possible, des représentations graphiques de vos résultats.*
- *Développements mathématiques*
 - Pour vous familiariser avec le modèle, vous pouvez écrire complètement la matrice de transition de la chaîne (X_n) lorsque $N = 3$ et l'étudier.
 - Vous pouvez détailler la preuve du théorème 1.
 - Vous pouvez démontrer certains des résultats qui sont admis dans le texte, comme l'équation (2) et les propositions 1, 3 et 5.
 - Vous pouvez chercher une forme quantitative de l'inégalité (8) et comparer la qualité du résultat à celui de la proposition 2.
 - Vous pouvez étudier la variante suivante du battage par insertion : à l'étape n , au lieu d'insérer la première carte à une position uniforme dans le paquet, on l'insère à une place choisie uniformément au hasard dans $\{N - n + 1, \dots, N\}$.
- *Modélisation*
 - Le modèle de battage examiné vous paraît-il satisfaisant ? Pouvez-vous en proposer un ou plusieurs autres, plus réalistes ?
 - Pensez-vous que la distance en variation entre la loi de X_n et la loi uniforme soit une mesure pertinente de la qualité du battage au temps n ? Pourriez-vous en proposer d'autres mesures ?
- *Étude numérique*
 - Vous pouvez simuler pour de petites valeurs de N , de l'ordre de 10, l'évolution de l'état du jeu de cartes au cours du temps.
 - Vous pouvez illustrer par une simulation numérique la plupart des résultats du texte lorsque N est plus grand, par exemple $N = 52$. En particulier, vous pouvez vérifier l'allure du graphe de la figure 1.