Projet 1

Permutations aléatoires et mesure d'Ewens

sujet proposé par Olivier Hénard

olivier.henard@universite-paris-saclay.fr

On décrit dans ce projet quelques propriétés des permutations aléatoires prises sous la mesure d'Ewens. Cette mesure a été introduite par Warren Ewens dans le domaine de la génétique des populations. Elle est notamment utilisée pour modéliser la distribution des allèles dans une population soumise à des mutations aléatoires, et donne lieu à diverses statistiques utiles pour estimer la diversité génétique. On propose dans ce projet une étude théorique des petits cycles sous la mesure d'Ewens, puis des cycles biaisés par la taille et des plus grands cycles pour une permutation uniforme.

1.1 Mesure d'Ewens et nombre total de cycles

Notons pour n entier ≥ 1 S_n l'ensemble des permutations de $[n] := \{1, \dots, n\}$. Une suite $u_2, u_3, \dots$ étant donnée avec $u_i \in [i]$ pour tout $i \geq 2$, on considère le procédé suivant de construction récursive d'une suite de permutations $(\sigma_n)_{n\geq 1}$ de S_n : σ_1 est la permutation identité: $\sigma_1(1) = 1$, puis on pose pour tout $n \geq 1$:

$$\sigma_{n+1} = \sigma_n \circ (u_{n+1}, n+1),$$

c'est-à-dire que σ_{n+1} est obtenue en composant σ_n par la droite par la transposition $(u_{n+1}, n+1)$ qui échange u_{n+1} et $n+1$.

De façon imagée, on imagine des clients $1, 2, 3...$ qui rentrent successivement dans un restaurant ou des clients sont déjà attablés; le client $n+1$ s'assoit à la table du client u_{n+1} à sa droite si $u_{n+1} \neq n+1$, ou s'assoit seul à une nouvelle table si $u_{n+1} = n+1$; les tables occupées par les clients à l'instant n correspondent aux cycles de la permutation σ_n, qui sont obtenus en parcourant les tables vers la droite.

Par exemple, si $n = 8$, les choix $u_2 = 1, u_3 = 3, u_4 = 1, u_5 = 4, u_6 = 2, u_7 = 7, u_8 = 3$ donnent lieu à la permutation σ_8 de S_8 suivante:

$$\sigma_8 = (14526)(38)(7) = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 4 & 6 & 8 & 5 & 2 & 1 & 7 & 3 \end{pmatrix} \tag{1.1}$$

où la première écriture donne la décomposition en cycles de σ_8, tandis que sur la seconde écriture, la seconde ligne donne les images par σ_8 des éléments de la première ligne.

Soit $\theta > 0$ un nombre réel. On injecte maintenant des probabilités en introduisant une suite $(U_n)_{n\geq 2}$ de variables indépendantes dont la loi est donnée, pour $n \geq 1$, par:

$$\begin{cases} U_{n+1} = k & \text{avec probabilité } \dfrac{1}{n+\theta}, \text{ si } k \in [n] \\ U_{n+1} = n+1 & \text{avec probabilité } \dfrac{\theta}{n+\theta}, \end{cases}$$

et on notera $(\Sigma_n)_{n\geq 1}$ la suite de permutations aléatoires obtenues; on dit alors que Σ_n possède la loi d'Ewens de paramètre θ sur S_n. Si σ est un élément de S_n, on notera $c_i(\sigma)$ le nombre de cycles de longueur i de σ, et $c(\sigma) = \sum_{i=1}^n c_i(\sigma)$ le nombre de cycles total. Sur l'exemple (1.1), $c(\sigma_8) = 3$ et $c_1(\sigma_8)$, $c_2(\sigma_8)$ et $c_5(\sigma_8)$ sont les trois valeurs non nulles, égales à 1.

T1. Montrer (par exemple par récurrence sur n) que, pour tout $n \geq 1$ entier et toute permutation $\sigma \in S_n$,

$$\mathbb{P}(\Sigma_n = \sigma) = \frac{\theta^{c(\sigma)}}{\theta(\theta+1)\dots(\theta+n-1)}, \tag{1.2}$$

et décrire qualitativement l'influence du paramètre θ. Quelle est la loi de la permutation aléatoire dans le cas particulier $\theta = 1$?

S1. Implémenter un algorithme qui prend en entrée n et θ et génère une permutation aléatoire Σ_n prise sous la loi d'Ewens de paramètre θ.

T2. Déduire de la construction précédente le calcul de la fonction génératrice suivante : $\sum_{\sigma\in S_n} s^{c(\sigma)}$.

T3. Montrer que $c(\Sigma_n)$ le nombre de cycles d'une permutation prise sous la mesure d'Ewens peut être représenté sous la forme $c(\Sigma_n) = \sum_{i=1}^n 1_{\{U_i=i\}}$, si l'on adopte la convention $U_1 = 1$. En déduire espérance et variance de $c(\Sigma_n)$, puis, à l'aide de l'inégalité de Chebychev, montrer la convergence en probabilité:

$$\frac{c(\Sigma_n)}{\log(n)} \xrightarrow{\mathbb{P}} \theta, \text{ quand } n \to \infty.$$

S2. Représenter des trajectoires de $n \longmapsto c(\Sigma_n)$, pour $1 \leq n \leq 1000$, pour $\theta = 1, 4, 10$. Commenter l'influence de θ. Représenter des trajectoires de $n \longmapsto \frac{c(\Sigma_n)}{\log(n)}$, a-t-on convergence p.s.? (On prendra un intervalle de temps aussi grand que possible pour observer l'éventuelle convergence).

T4. On note Γ la fonction Gamma d'Euler [1] et on pose $u_n = \sqrt{\theta\log(n)}$. Montrer que la fonction génératrice $\varphi_n(t) = \mathbb{E}\big[e^{t\frac{c(\Sigma_n)-\theta\log(n)}{\sqrt{\theta\log(n)}}}\big]$ satisfait :

$$\varphi_n(t) = e^{-tu_n}\frac{\Gamma(\theta)\Gamma(n+e^{t/u_n}\theta)}{\Gamma(n+\theta)\Gamma(e^{t/u_n}\theta)}.$$

En passant à la limite dans l'expression précédente, on peut montrer [2] la convergence en loi vers une loi normale centrée réduite quand $n \to \infty$:

$$\frac{c(\Sigma_n)-\theta\log(n)}{\sqrt{\theta\log(n)}} \Rightarrow \mathcal{N}(0,1).$$

[1] on rappelle $\Gamma(x+1) = x\Gamma(x)$ et $\Gamma(1) = 1$ en particulier $\Gamma(n) = (n-1)!$ si n est un entier ≥ 1

[2] vous pouvez essayer, mais c'est tout à fait facultatif, à ne traiter que si tout est terminé; on rappelle la formule de Stirling $\Gamma(x) \sim \sqrt{2\pi}x^{x-1/2}e^{-x}$ (pas de faute de frappe ici...) quand $x \to \infty$

S3. Illustrer cette convergence en traçant l'histogramme associé à 500 réalisations de $\frac{c(\Sigma_n)-\theta\log(n)}{\sqrt{\theta\log(n)}}$ pour $n = 200$.

1.2 Formule d'échantillonnage d'Ewens et petits cycles

On appelle partition de $[n]$ une famille d'ensembles deux a deux disjoints de $[n]$ dont l'union est $[n]$. Noter que les cycles d'une permutation considérés chacun en tant qu'ensemble (sans l'ordre des éléments à l'intérieur des cycles) forment une partition de $[n]$. Ainsi sur l'exemple (1.1) la partition[3] de $[8]$ associée à la décomposition en cycles est $\{\{1,2,4,5,6\},\{3,8\},\{7\}\}$, elle a 3 blocs: 1 bloc de taille 1, 1 bloc de taille 2 et 1 bloc de taille 5.

T5. L'objectif de cette question est la démonstration de la formule d'échantillonnage d'Ewens qui décrit la loi jointe du vecteur $(c_1(\Sigma_n),\dots,c_n(\Sigma_n))$ pour Σ_n sous la mesure d'Ewens.

1. Pour une partition de $[n]$ donnée avec k_i blocs de taille i, $1 \le i \le n$, montrer que le nombre de permutations dont la décomposition en cycle qui se projette sur cette partition est $\prod_i (i-1)!^{k_i}$.
2. Montrer que le nombre de partitions de $[n]$ qui ont k_i blocs de taille i pour tout i entre 1 et n vaut: [4]
$$\frac{n!}{\prod_i k_i!\; i!^{k_i}}.$$
3. Quelle est la probabilité $\mathbb{P}(\Sigma_n = \sigma)$ pour σ une permutation de S_n telle que $c_i(\sigma_n) = k_i$ pour tout i entre 1 et n ?

En déduire la formule d'Ewens : pour $k_1,\dots,k_n \ge 0$ des entiers,

$$\mathbb{P}(c_1(\Sigma_n)=k_1, c_2(\Sigma_n)=k_2,\dots,c_n(\Sigma_n)=k_n) = \frac{n!}{\theta\dots(\theta+n-1)}\frac{1}{2^{k_2}\dots n^{k_n}k_1!\dots k_n!}\theta^{\sum_{i=1}^n k_i}\mathbf{1}_{\sum_i ik_i=n}$$

T6. Si $N_1, N_2, \dots, N_n$ sont des variables aléatoires indépendantes, et N_i suit la loi de Poisson de paramètre θ/i, calculer:
$$\mathbb{P}(N_1 = k_1, N_2 = k_2, \dots, N_n = k_n)$$
et en déduire que la loi jointe de $(c_1(\Sigma_n), c_2(\Sigma_n), \dots, c_n(\Sigma_n))$ est une loi conditionnelle simple.

Longueurs des cycles et restaurant de Feller On veut illustrer les résultats précédents. Si on s'intéresse aux seules longueurs des cycles, du point de vue de la simulation, il peut être utile de considérer le restaurant de Feller; en un mot l'idée est la suivante; l'entier n étant fixé, on cherche à donner directement la composition des tables, en énumérant dans l'ordre les clients de la première table (en partant de 1), puis les clients de la seconde table en commençant par le plus petit entier de cette table,...

Précisément, soit $(B_k^n)_{1\le k\le n-1}$ une suite finie de variables aléatoires indépendantes de loi de Bernoulli avec
$$\mathbb{P}(B_k^n = 1) = 1 - \mathbb{P}(B_k^n = 0) = \frac{\theta}{\theta+n-k}.$$
Ensuite, on pose $A_1 = \{2,\dots,n\}$ et $U_0 = 1$. À l'étape $k = 1,\dots,n-1$:

[3] Noter que l'ordre des blocs n'a pas d'importance, seule compte la relation d'equivalence "être dans le même bloc" associée a la partition

[4] on pourra commencer par regarder la définition du coefficient multinomial, qui généralise le coefficient binomial

1. Si $B_k^n = 0$, le client U_k, avec U_k variable de loi uniforme sur A_k se place à droite du client U_{k-1} à sa table.

2. Si $B_k^n = 1$, on ferme la table en cours et le client $U_k = \min A_k$ commence une nouvelle table.

Enfin, $A_{k+1} := A_k \setminus U_k$. Après l'étape $k = n-1$, on obtient une permutation aléatoire, notée Θ_n, via sa décomposition en cycles.

Noter que A_k comprend les indices des clients non encore placés avant l'étape k, A_k possède $n-k$ éléments, et les indices $B_1^n, B_2^n, \ldots$ décident de la fermeture des tables. Si on reprend l'exemple 1.1, alors on a que la suite des $(U_0, U_1, U_2, \ldots)$ est (14526387) et $(B_1^8, B_2^8, \ldots)$ est (0000101)

T7. Montrer que la permutation Θ_n ainsi construite suit encore la loi d'Ewens de paramètre θ (on pourra vérifier que l'équation (1.2) vaut encore pour Θ_n).

Pour σ une permutation, on note $\ell_1(\sigma)$ la longueur du cycle qui contient 1, et $\ell_2(\sigma)$ la longueur du cycle qui contient le plus petit entier non compris dans le cycle qui contient 1, et ainsi de suite : c'est-à-dire qu'on énumère les cycles dans leur ordre d'apparition. Pour l'exemple (1.1), $\ell_1(\sigma_8) = 5$, $\ell_2(\sigma_8) = 2$, $\ell_3(\sigma_8) = 1$.

T8. Donner une formule pour les longueurs $(\ell_i(\Theta_n), i \geq 1)$ des cycles de Θ_n en fonction des espacements entre les éléments de l'ensemble $\{1 \leq i \leq n-1 : B_i^n = 1\} =: \{J_1 < J_2 < \ldots < J_k\}$.

S4. Pour $\theta = 1, 2, 5$, implémenter un algorithme qui prend n et θ en arguments et retourne une réalisation de $(c_1(\Theta_n), c_2(\Theta_n), \ldots, c_n(\Theta_n))$[5]. Tracer l'histogramme de $c_1(\Theta_n)$ ainsi que les valeurs de la fonction de masse $k \mapsto \mathbb{P}(N_1 = k)$ de la loi de Poisson de paramètre 1 sur le même schéma pour $k = 0, \ldots, 10$.

1.3 Grands cycles d'une permutation uniforme

Dans toute la suite du projet, on considère le cas $\theta = 1$ où la loi de Σ_n est uniforme.

T9. Montrer que $\ell_1(\Sigma_n)$ la longueur du cycle qui contient 1 dans Σ_n est de loi uniforme sur $[n]$, c'est-à-dire:

$$\mathbb{P}\left(\ell_1(\Sigma_n) = k_1\right) = \frac{1}{n}$$

pour $1 \leq k_1 \leq n$. En déduire la loi jointe de $(\ell_1(\Sigma_n), \ldots, \ell_r(\Sigma_n))$:

$$\begin{aligned}\mathbb{P}(\ell_1(\Sigma_n) = k_1, \ell_2(\Sigma_n) = k_2, \ldots, \ell_r(\Sigma_n) = k_r) &= \frac{1}{n}\frac{1}{n-k_1}\frac{1}{n-k_1-k_2}\cdots\frac{1}{n-k_1-\ldots-k_{r-1}} \\ &= \prod_{\ell=1}^{r}\frac{1}{n-\sum_{j=0}^{\ell-1}k_j}\end{aligned}$$

pour des entiers non nuls $k_1, \ldots, k_r$ tels que $k_1 + k_2 + \ldots + k_r \leq n$.

[5] on rappelle que $c(\sigma)$ pour σ permutation a été définie juste avant **T1**

On note $\ell_{(1)}(\sigma) \geq \ell_{(2)}(\sigma) \geq \dots$ le réordonnancement par taille décroissante des longueurs de cycles $\ell_1(\sigma), \ell_2(\sigma), \dots$. On admet la convergence en loi de la longueur renormalisée par n du plus grand cycle pour une permutation uniforme de $[n]$, et on note L la loi limite obtenue, c'est-à-dire:

$$\frac{\ell_{(1)}(\Sigma_n)}{n} \Rightarrow L. \tag{1.3}$$

T10. Soit $(U_i)_{i\geq 1}$ des variables aléatoires uniformes sur $[0,1]$ i.i.d. Montrer pour tout $r \geq 1$ la convergence en loi suivante quand $n \to \infty$,

$$\left(\frac{\ell_1(\Sigma_n)}{n}, \frac{\ell_2(\Sigma_n)}{n}, \dots, \frac{\ell_r(\Sigma_n)}{n}\right) \Rightarrow \big(U_1, (1-U_1)U_2, , \dots, (\prod_{i=1}^{r-1}(1-U_i))U_r\big).$$

Proposer [6] une formule pour la loi limite L qu'on exprimera comme un certain maximum d'une fonction des variables aléatoires $(U_i)_{i\geq 1}$.

Dans la suite du projet on étudie la loi de L, également connue sous le nom de loi de max-Dickman.

T11. Soit k entier tel que $n/2 < k \leq n$. Montrer que le nombre de permutations qui possèdent un cycle de longueur k satisfait :

$$\{\sigma \in S_n : \sigma \text{ possède un cycle de longueur } k\} = \frac{n!}{k}$$

En déduire, pour ces mêmes valeurs de k, la valeur de $\mathbb{P}\left(\ell_{(1)}(\Sigma_n) = k\right)$, puis le calcul de la limite suivante pour t nombre réel tel que $1/2 < t \leq 1$:

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\ell_{(1)}(\Sigma_n)}{n} \leq t\right),$$

et enfin la fonction de répartition de L sur l'intervalle $[1/2, 1]$.

T12. Montrer que L (définie par (1.3)) satisfait l'équation en loi suivante, avec U variable uniforme sur l'intervalle $[0,1]$ indépendante de L dans le membre de droite:

$$L \overset{\text{loi}}{=} \max\{U, (1-U)L\}$$

T13. Montrer que la fonction $\rho : [1,\infty) \to \mathbb{R}, u \mapsto \rho(u) := \mathbb{P}(L \leq 1/u)$ satisfait:

$$\rho(u) = \frac{1}{u}\int_{u-1}^{u} \rho(t)dt$$

puis en déduire que ρ satisfait l'équation différentielle à retard suivante:

$$\begin{cases} \rho(u) = 1 \text{ si } u \in [0,1], \\ u\rho'(u) + \rho(u-1) = 0 \text{ si } u \geq 1. \end{cases}$$

[6] on ne demande pas ici une preuve rigoureuse!

S5. Représenter la fonction $u \mapsto \rho(u)$ sur $[0,3]$. S'il vous reste de l'énergie, comparer à la fonction de répartition empirique évaluée en $\frac{1}{u}$ de la variable $\frac{\ell_{(1)}(\Sigma_n)}{n}$, soit à la fonction:

$$u \longmapsto \frac{1}{N}\text{Card}\left\{1 \leq i \leq N : \frac{\ell_{(1)}(\Sigma_n^i)}{n} \leq \frac{1}{u}\right\}$$

avec $(\Sigma_n^i)_{1\leq i\leq N}$ des copies indépendantes de Σ_n.

Notes supplémentaires : Ramanajuan a prouvé que la fonction ρ peut être représentée sous la forme intégrale suivante:

$$\rho(u) = 1 + \sum_{k\geq 1} \frac{(-1)^k}{k!} \int_{I_k(u)} \frac{1}{y_1 \dots y_k} dy_1 \dots dy_k$$

avec le domaine d'intégration $I_k(u) = \{y_1 > 1/u, \dots, y_k > 1/u, y_1 + \dots + y_k < 1\}$.