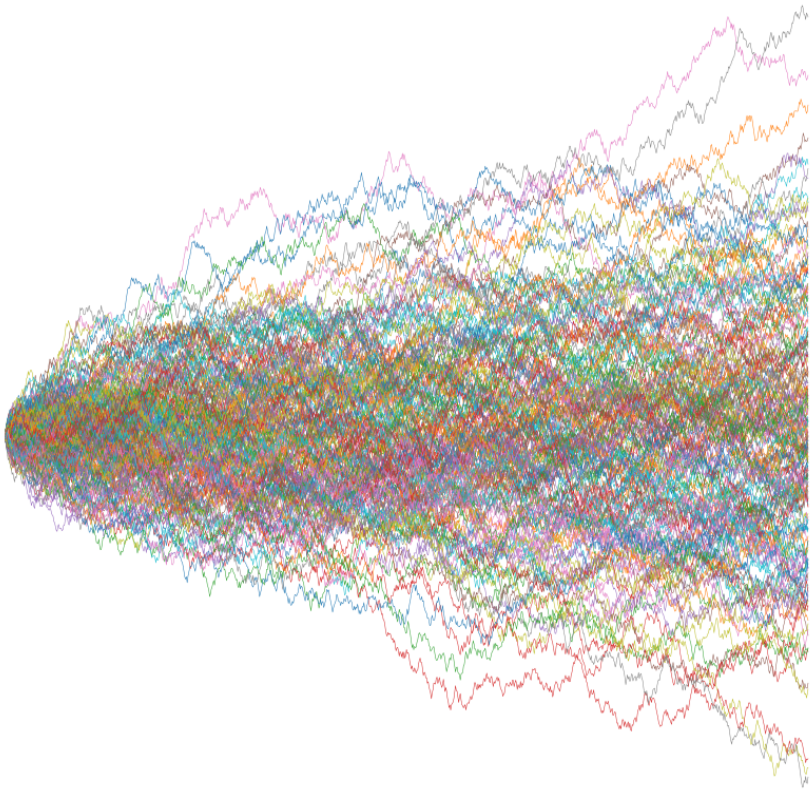
Projets de simulation Python

MAP361P - *Aléatoire*



2020-2021

Projets de simulation – Instructions générales

Chaque élève doit travailler sur un projet en **binôme**. La formation des binômes est libre, mais ceux-ci ne doivent pas être constitués de plus de **deux** élèves.

Ce projet, qui donne lieu à une note comptant pour 30% de la note du cours MAP361, sera choisi dans la liste ci-jointe, chaque projet devant être choisi par 13 binômes au maximum. Bien entendu, si plusieurs binômes qui ont choisi le même sujet rendent des copies ou des programmes “trop proches", nous nous verrons dans l'obligation de réduire la note de chacun.

1 Calendrier

► Présentation de la liste des projets : le fascicule sera disponible à partir du **3 mai 2021**.

► Choix des projets : les élèves devront exprimer leurs vœux à l'adresse

`https://de.polytechnique.fr`

au plus tard le **vendredi 14 mai 2021, 23h59**.

Dans les jours qui suivent, un algorithme répartira les sujets de façon automatique, en fonction des vœux émis, et les élèves seront ensuite informés du sujet qui leur est attribué. Les élèves n'ayant pas informé la scolarité de leurs vœux dans les temps se verront affecter un sujet parmi ceux qui restent disponibles.

► Remise des rapports de projets (avec les programmes) par mail à l'enseignant ayant proposé le sujet dont l'adresse électronique figure dans la présentation de chaque projet. La date limite de retour des projets est fixée au

vendredi 2 juillet 2021 à midi.

Tout projet rendu en retard verra sa note réduite.

2 Cours de Python

► Un cours en ligne d'introduction à Python sera disponible le jeudi 22 avril sur la page moodle du MAP361P.

► Des séances de TP dédiées à Python à l'utilisation des notebooks Jupyter auront lieu les **5 et 12 mai** via zoom. Vérifier au préalable votre numéro de groupe afin de pour connaitre la date et l'heure de votre TP.

► Des liens vers de la documentation concernant le langage Python sont accessibles sur la page moodle.

► En cas de difficultés liées au sujet, le binôme devra s'adresser par mail à l'enseignant ayant proposé le sujet pour des explications complémentaires.

3 Rapport de projet

Chaque projet contient une composante théorique portant sur le contenu du cours MAP361 ainsi qu'une partie programmation/simulations qui doit être réalisée à l'aide du langage Python. En particulier, les questions de simulation aboutiront à la génération de courbes et de graphiques qui doivent être rendus avec le projet, tout comme les codes utilisés pour leur génération. Courbes et codes doivent donner lieu à des commentaires, qui seront pris en compte dans la note du projet.

Les initiatives personnelles sont fortement encouragées.

Dans les énoncés des projets, les questions marquées **(T)** sont théoriques, les questions **(S)** relèvent de la programmation/simulation. Chaque projet mélange donc deux aspects:

Partie théorique. Elle doit contenir les réponses aux questions théoriques, les réponses aux questions expérimentales (pouvant prendre la forme de résultats numériques ou de graphiques), des commentaires et des explications sur les résultats expérimentaux.

Partie programmation. Elle doit contenir tous les codes. Le correcteur doit pouvoir facilement exécuter les programmes et les tester en modifiant les paramètres essentiels : il faut donc que ces paramètres soient représentés par des lettres dans les programmes, dont les valeurs sont définies au début du programme. Par exemple, l'utilisation d'un paramètre n se fera en commençant le programme par $n = 1000$ (par exemple) puis en se référant à n tout au long du programme plutôt que de recopier 1000 à chaque utilisation du paramètre.

4 Consignes pour le rapport

Un seul rapport par binôme est exigé. Il doit rendu sous la forme

(a) d'un **notebook Jupyter, faisant office de code+rapport**. On veillera à la bonne utilisation de Markdown et Latex (pour les formules mathématiques) pour une mise en forme lisible. On veillera à respecter les consignes suivantes:

- Ne pas utiliser le module d'extension *some Latex environements for Jupyter*[1] car il pose problème sur certains ordinateurs. Ceci ne vous empêche en rien d'utiliser les commandes Latex/Markdown disponibles par défaut avec Jupyter notebook..
- **Ne pas utiliser de bibliothèques Python non standards**, *i.e.* autres que les modules de base, `Numpy`, `Scipy` et `Matplotlib`, sauf si c'est expressément spécifié dans l'énoncé du projet.
- **Envoyer votre notebook (le fichier .ipynb) déjà exécuté**: le correcteur doit pouvoir visualiser les résultats de vos programmes même sans exécuter les cellules de code.

(b) accompagné éventuellement d'un rapport en format `pdf` (écrit avec latex ou un autre traitement de texte). Ce ficher est optionnel. Il peut s'avérer pratique dans la cas d'un projet comprenant une importante partie théorique mais **il ne pourra en aucun cas se substituer au notebook lui-même**.

On veillera à joindre les fichiers du rapport nommés sous la forme:

`NomBinome1_NomBinome2_NumeroProjet`

[1] C'est le module que la video du cours en ligne sur moodle vous propose d'installer. Mais c'est, en fin de compte, inutile.

Liste des projets

PROJET 1

Un modèle épidémiologique: S.I.R.

sujet proposé par Ch. Bertucci

charles.bertucci@polytechnique.edu

Le but de ce projet est d'étudier quelques aspects d'un modèle très classique (et simpliste) en épidémiologie, le modèle *SIR* (*Sane, Infected, Recovered*).

Dans ce modèle, on se donne une population qui est répartie en trois catégories S, I et R désignant respectivement les individus sains, infectés et guéris (donc immunisés). Les individus sains peuvent être infectés et les individus infectés peuvent guérir. Nous allons étudier deux variantes de ce modèle : une déterministe et une stochastique.

1 Un modèle SIR déterministe

Théorie

Dans ce modèle, les quantités $S(t)$, $I(t)$ et $R(t)$ représentent les quantités d'individus de chacune de ces catégories à l'instant $t \geq 0$. La quantité totale d'individus est donnée par $M_0 = S(0) + I(0) + R(0)$. On suppose que l'évolution des ces trois quantités est régie par le système d'équations différentielles ordinaires (EDO) :

$$\begin{aligned} \frac{dS(t)}{dt} &= -\beta \frac{I(t)S(t)}{M_0}, \\ \frac{dI(t)}{dt} &= \beta \frac{I(t)S(t)}{M_0} - \gamma I(t), \\ \frac{dR(t)}{dt} &= \gamma I(t), \end{aligned} \tag{1.1}$$

où γ et β sont deux constantes strictement positives qui représentent respectivement des taux d'infection et de guérison. Pour information, le (désormais) fameux coefficient $\mathcal{R}_0$ associé à ce modèle est simplement égal à $\gamma^{-1}\beta$.

T1. On admet qu'il existe une unique solution $\mathcal{C}^1$ $(S(t), I(t), R(t))$ sur $[0, \infty[$ caractérisée par le système (1.1) et les conditions initiales $(S(0), I(0), R(0))$. Montrer que pour tout temps $t \geq 0$, on a $R(t) + I(t) + S(t) = M_0$. En déduire que pour tout temps $t \geq 0$ les quantités $S(t), I(t)$ et $R(t)$ sont dans $[0, M_0]$.

T2. En admettant toujours qu'il existe une solution au problème, montrer que cette solution converge nécessairement vers un point limite $(S(\infty), 0, R(\infty))$ lorsque t tend vers l'infini.

T3. Sous quelle condition sur les paramètres $S(0), \beta$ et γ la fonction I est-elle décroissante avec le temps ? Quelles sont les variations de I lorsqu'elle n'est pas décroissante sur $[0, \infty[$? En déduire que si $\mathcal{R}_0 \leq 1$, alors le nombre d'infectés n'augmente jamais au cours du temps. Montrer au contraire que, lorsque $\mathcal{R}_0 > 1$, il y a forcement une phase de croissance des infections lorsque l'on part initialement de $R(0) = 0$ (personne n'est immunisé) et que le nombre initial d'infectés $I(0)$ est suffisamment petit.

T4. **(facultatif)** En utilisant le théorème de Cauchy-Lipschitz, justifier qu'il existe une unique solution au système d'EDO étant données des conditions initiales.

Simulation

Pour simuler le modèle précédent, on discrétise les EDO de la façon suivantes :

- On ne cherche plus des fonctions du temps mais 3 suites $(S_n)_{n\geq 0}, (I_n)_{n\geq 0}, (R_n)_{n\geq 0}$.
- Le système d'EDO est remplacé par la relation de récurrence

$$\begin{aligned}
\frac{S_{n+1} - S_n}{\Delta t} &= -\beta \frac{I_n S_n}{M_0}, \\
\frac{I_{n+1} - I_n}{\Delta t} &= \beta \frac{I_n S_n}{M_0} - \gamma I_n, \\
\frac{R_{n+1} - R_n}{\Delta t} &= \gamma I_n,
\end{aligned}$$

  où $\Delta t > 0$ est un paramètre du modèle.

S1. Ecrire un programme qui donne la valeur des suites $(S_n)_{n\geq 0}, (I_n)_{n\geq 0}, (R_n)_{n\geq 0}$ jusqu'à un rang N en fonction des valeurs des différents paramètres.

S2. Tracer l'évolution des suites $(S_n)_{n\geq 0}, (I_n)_{n\geq 0}, (R_n)_{n\geq 0}$ sur un même graphique (on tracera l'évolution des suites jusqu'à ce qu'on est atteint ce qui semble être un état stationnaire) pour les valeurs des paramètres suivantes :$M_{tot} = 1$, $R(0) = 0, S(0) \in \{0.99; 0.8\}$, $(\beta, \gamma) \in \{(0.9, 1.2); (1.2, 0.9); (1, 1)\}$, $\Delta t = 0.01$.

S3. Que se passe t'il lorsque Δt est trop grand ?

S4. Implémenter directement la résolution du système (1.1) à l'aide de la méthode `odeint` issue du package `scipy.integrate` et comparer avec les résultats obtenus manuellement.

2 Un modèle SIR stochastique

Le modèle déterministe précédent suppose que la population considerée est très grande (assimilée comme étant infinie) et ne prend donc pas en compte l'aléa inhérent à la dynamique locale de propagation de l'épidémie. On s'intéresse maintenant à un modèle stochastique pour des populations de taille finie. On notera respectivement S_t, I_t et R_t le nombre d'individus dans chacune de ces catégories à l'instant t. La population est initialement dans l'état (S_0, I_0, R_0) et on note $M_0 = S_0 + I_0 + R_0$. Les quantités S_t, I_t et R_t sont des variables aléatoires définies de la façon suivante :

- Si à l'instant t, la quantité d'individus infectés est I_t, tous les individus sains ont des horloges exponentielles indépendantes de paramètre $\beta \frac{I_t}{M_0}$ qui leur sont associées.

- De même, au temps t, tous les individus infectés ont des horloges exponentielles indépendantes de paramètre γ (qui sont également indépendantes des horloges associées aux individus sains).
- Lorsque la première de toutes ces horloges sonne, l'individu associé change d'état : il devient infecté s'il était sain et devient guéri s'il était infecté. Puis la dynamique continue...

On remarquera qu'à chaque changement d'état, la valeur de I_t change, et donc aussi les paramètres des horloges des individus sains.

Premier regard sur le modèle stochastique

T5. En utilisant la propriété de perte de mémoire des horloge exponentielles ainsi que leur indépendance, justifier rapidement que presque sûrement le modèle introduit est bien défini. Expliquer en particulier pourquoi on peut, au choix, retirer ou conserver les horloges de guérison restantes après chaque évènement (guérison ou infection) alors qu'il faut retirer les horloges d'infections.

T6. Montrer que, presque sûrement, la population est, au bout d'un temps fini, dans un état du type $(S(\infty), 0, M_0 - S(\infty))$ pour $S(\infty)$ un entier inférieur à M_0.

S5. Ecrire un programme qui simule, en fonction des paramètres, une évolution de la population de l'état (S_0, I_0, R_0) jusqu'à un état du type $(S(\infty), 0, M_0 - S(\infty))$ pour $S(\infty)$ un entier inférieur à M_0. On pourra commencer par écrire un code qui simule une seule transition de la population.

S6. Représenter 10 évolutions indépendantes de la population dans les cas suivants $M_0 \in \{10; 50\}$, $\beta = 1.2$, $\gamma = 0.9$, $R_0 = 0$, $I_0 = 0.1 * M_0$ puis pour M_0 le plus grand possible (avec un programme qui tourne en moins de 30s disons...). Choisissez ensuite vous-même d'autres jeux de paramètres pour tester les différents comportements possibles du système.

Simplification du modèle

T7. Soit $X_1, ...X_n$ des variables aléatoires exponentielles de paramètres respectifs $\lambda_1, ..., \lambda_n$. Montrer que $\min\{X_1, ..., X_n\}$ est une variable aléatoire exponentielle, dont on donnera le paramètre, qui est indépendante des évènements $\{X_i = \min\{X_1, ..., X_n\}\}$ pour tout $1 \leq i \leq n$.

On introduit un autre modèle (S'_t, I'_t, R'_t) défini par :

- $(S'_0, I'_0, R'_0) = (S_0, I_0, R_0)$.
- Si à l'instant t la population est dans l'état (S'_t, I'_t, R'_t) avec $S'_t, I'_t \geq 1$, alors elle peut passer dans l'état $(S'_t - 1, I'_t + 1, R'_t)$. C'est une transition Infection.
- Si à l'instant t la population est dans l'état (S'_t, I'_t, R'_t) avec $I'_t \geq 1$, alors elle peut passer dans l'état $(S'_t, I'_t - 1, R'_t + 1)$. C'est une transition Guérison.
- Les transitions Infection et Guérison sont associées à des horloges exponentielles indépendantes de paramètres respectifs $\beta \frac{I'_t S'_t}{M_0}$ et $\gamma I'_t$ et, comme dans le modèle précédent, lorsque la première horloge sonne, la transition associée a lieu.

T8. Montrer que (S'_t, I'_t, R'_t) et (S_t, I_t, R_t) ont même loi.

S7. Même question que S6 pour ce nouveau modèle. Comparez l'efficacité des deux implémentations.

Limite du modèle pour M_0 grand

T9. On regarde le premier modèle (S_t, I_t, R_t) en partant d'un état initial $(0, N, 0)$.
On note $(S_{N,t}, I_{N,t}, R_{N,t})_{t\geq 0}$ la solution associée à cette condition initiale. Pour tout temps $t \geq 0$, donner

$$\lim_{N\to\infty} \frac{I_{N,t}}{N}.$$

S8. En déduire une conjecture pour l'évolution de $(M_0)^{-1}(S_t, I_t, R_t)_{t\geq 0}$ lorsque M_0 tend vers l'infini. Illustrer numériquement (à l'aide du deuxième modèle) l'évolution de $(M_0)^{-1}(S_t, I_t, R_t)_{t\geq 0}$ pour M_0 grand.

PROJET 2

Descente de gradient stochastique

sujet proposé par Ch. Bertucci

charles.bertucci@polytechnique.edu

Le but de ce projet est de se familiariser avec différents algorithmes d'optimisation stochastique.

1 Descente de gradient dans le cas déterministe

On se donne une fonction $f : \Omega \to \mathbb{R}$ où Ω est un ouvert convexe de $\mathbb{R}^d$ pour un entier $d \geq 1$. On note $\|\cdot\|$ la norme euclidienne de $\mathbb{R}^d$. On considère le problème de trouver $x^* \in \Omega$ tel que

$$f(x^*) = \min_{x \in \Omega} f(x). \tag{2.1}$$

On suppose ici que f est une fonction convexe et différentiable et on propose d'approcher x^* (s'il existe) à l'aide d'un algorithme dit de descente de gradient expliqué ci-dessous.

$$\begin{cases} x_0 \in \Omega, \\ \forall n \geq 0, \text{ choisir } \epsilon_n > 0 \text{ et poser } x_{n+1} := x_n - \epsilon_n \nabla f(x_n). \end{cases}$$

Dans l'algorithme précédent le choix de la suite de pas $(\epsilon_n)_{n \geq 0}$ est d'une importance capitale sur la (vitesse de) convergence, cependant nous ne rentrerons pas ici dans ce genre de détails et nous étudierons principalement des suites de pas constantes. Nous nous concentrons ici sur un cas simple.

Un peu de théorie

T1. (Question préliminaire) Soit $f : \Omega \to \mathbb{R}$ une fonction α-convexe, c'est à dire telle que

$$f(ta + (1-t)b) + \frac{\alpha}{2} t(1-t)\|a-b\|^2 \leq tf(a) + (1-t)f(b).$$

Montrer que pour tout $x, y \in \Omega$:

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \alpha \|x - y\|^2.$$

T2. On suppose ici que : la solution x^* de (2.1) existe, f est C^2, que ses dérivées secondes sont bornées sur Ω (en particuliers ∇f est donc une application Lipschitzienne pour une certaine constante $L > 0$), et que f est α-convexe pour $\alpha > 0$. Montrer que pour $\eta > 0$ suffisamment petit, pour tout $x_0 \in \Omega$ si pour tout $n \geq 0, \epsilon_n = \eta$, alors il existe $\delta \in (0, 1)$ tel que pour tout $n \geq 1$:

$$\|x_{n+1} - x_n\|^2 \leq \delta \|x_n - x_{n-1}\|^2.$$

En déduire que dans ce cas l'algorithme de descente de gradient converge.

T3. Que peut-on espérer lorsque la fonction f n'est pas α convexe mais juste convexe ? Lorsqu'elle n'est pas convexe ?

Mise en pratique

On veut ici mettre en oeuvre l'algorithme précédent dans le cas suivant : $\Omega = \mathbb{R}^{100}$, f définie sur Ω par

$$f(x) := x^T A x - \langle g, x \rangle,$$

où $A \in \mathcal{M}_{100}(\mathbb{R})$ est la matrice de coefficients $(a_{ij})_{1 \leq i,j \leq n}$ donnés par

$$\begin{cases} a_{ij} = 4 \text{ si } i = j, \\ a_{ij} = -1 \text{ si } |i - j| = 1, \\ a_{ij} = 0 \text{ sinon}, \end{cases} \tag{2.2}$$

et g est le vecteur donné par

$$g_i = \cos(\frac{2i\pi}{20}). \tag{2.3}$$

T4. Montrer que la fonction f est α-convexe pour un certain $\alpha > 0$ en remarquant que: i) la fonction $x \to x^T x$ est 1-convexe, ii) la somme d'une fonction α-convexe et d'une fonction convexe est α-convexe.

T5. En remarquant que la matrice A est symétrique, montrer que la solution du problème est donnée dans ce cas par

$$x^* = \frac{1}{2} A^{-1} g.$$

S1. Ecrire un programme qui réalise les n premières étapes de la méthode de descente de gradient pour une suite de pas $(\epsilon_n)_{n \geq 0}$ choisis constants.

S2. Tracer l'évolution de l'erreur de convergence pour différentes valeurs de pas (on prendra $\epsilon_n \in \{10; 2; 1; 0.1\}$) en prenant $x_0 = 0$.

2 Le problème stochastique

Le problème auquel on s'intéresse est ici de trouver x^* tel que

$$x^* = \underset{x \in \Omega}{\text{argmin}} \mathbb{E}[f(x, \xi)], \tag{2.4}$$

où ξ est une variable aléatoire p dimensionnelle sur un espace probabilisé $(\mathcal{O}, \mathcal{A}, \mathbb{P})$, et $f : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$ une fonction donnée.

On pose $F(x) = \mathbb{E}[f(x, \xi)]$.

T6. Montrer que f convexe en x pour tout ξ (resp. α convexe) implique F convexe (resp. α-convexe).

T7. Sous quelles hypothèses sur f peut on avoir

$$\nabla F(x) = \mathbb{E}[\nabla_x f(x, \xi)], \tag{2.5}$$

où $\nabla_x f(x, p)$ est le gradient de $x \to f(x, p)$ au point x.

En pratique le calcul de l'espérance suivant la loi de ξ peut être très couteux voire carrément impossible. C'est pourquoi on n'applique directement l'algorithme de la descente de gradient à l'aide de la relation (2.5) mais on cherche à remplacer le gradient de F par un estimateur. On s'intéresse maintenant à deux cas particuliers de tels estimateurs.

Estimateur de type Batch

On va ici estimer le gradient de F de la façon suivante. On se donne N réalisations $(\xi_1, ..., \xi_N)$ de la variable aléatoire ξ. Lorsque l'on se trouve au point x_n, au lieu d'écrire

$$x_{n+1} = x_n - \epsilon_n \nabla F(x_n),$$

on construit

$$x_{n+1} = x_n - \epsilon_n \frac{1}{N} \sum_{i=1}^{N} \nabla_x f(x_n, \xi_i).$$

On va mettre en oeuvre cet algorithme dans le cas suivant :

- $\xi = ((\alpha_{ij})_{1 \leq i,j \leq 100}, (\beta_i)_{1 \leq i \leq 100})$ est à valeurs dans $\mathcal{M}_{100}(\mathbb{R}) \times \mathbb{R}^{100}$. Toutes les composantes de ξ sont indépendantes les unes des autres, de loi Gaussienne d'écart type 0.1 et de moyenne a_{ij} et g_i pour respectivement α_{ij} et β_i (ou a et g sont donnés par (2.2) et (2.3)).
- $f(x, (M, v)) = x^T M x - \langle v, x \rangle$.

T8. Que dire de la convexité de F dans ce cas ?

T9. Calculer la solution x^* du problème (2.4).

S3. Écrire un programme qui simule N réalisations indépendantes de ξ et qui réalise les n premières étapes de la descente stochastique de type Batch pour une suite de pas de temps constant.

S4. Tracer l'évolution de l'erreur de convergence dans pour différentes valeurs de pas (on prendra $\epsilon_n \in \{10; 2; 1; 0.1\}$) en fonction de n et de N. On pourra partir de $x_0 = 0$. Commenter.

Descente de gradient stochastique classique

On considère ici un autre estimateur du gradient de F. Lorsqu'on se trouve au point x_n au lieu de calculer x_{n+1} à l'aide de $\nabla F(x)$ on va tirer une réalisation de ξ_n indépendante de x_n (qui est en général aléatoire) et construire

$$x_{n+1} = x_n - \epsilon_n \nabla_x f(x_n, \xi_n). \tag{2.6}$$

La suite d'itérations $(x_n)_{n \geq 0}$ ainsi obtenue est donc aléatoire.

S5. Même question que S3 et S4 pour ce nouvel algorithme (i.e. dans le même cas). Comparer les résultats ainsi obtenus dans les deux méthodes. Commenter.

T10. (Question difficile/facultative, des ébauches de réponses seront suffisantes...) On suppose que f et F sont aussi réguliers que nécessaire. On suppose qu'il existe $\alpha > 0$ tel que pour tout p, $f(\cdot, p)$ est α-convexe. On suppose également qu'il existe une solution x^* au problème (2.4). Montrer que si on prend une suite $(\epsilon_n)_{n\geq 0}$ constante égale à η, avec $\eta > 0$ suffisamment petit, alors la suite d'itérations $(x_n)_{n\geq 0}$ de l'algorithme de descente de gradient stochastique satisfait

$$\limsup_{n\to\infty} \mathbb{E}[F(x_n)] \leq F(x^*) + \frac{\eta L M}{2\alpha}, \tag{2.7}$$

où L et M sont des constantes telles que

$$\|\nabla F(x) - F(y)\| \leq L\|x - y\|,$$

$$\mathbb{E}[\|\nabla_x f(x, \xi)\|^2] \leq M + 2\|\nabla F(x)\|^2.$$

On pourra en particuliers pour s'aider majorer $(\mathbb{E}[F(x_n)] - F(x^*))_{n\geq 0}$ par une suite décroissante en utilisant éventuellement la relation

$$2\alpha(F(x) - F(x^*)) \leq \|\nabla F(x)\|^2.$$

PROJET 3

Modélisation et simulation d'un aimant permanent à température ambiante

sujet proposé par A. de Bouard

anne.debouard@polytechnique.edu

Une particule ferromagnétique de taille nanométrique (assimilée à une particule ponctuelle) est décrite par son aimantation, qui est un vecteur m de $\mathbb{R}^3$, de norme constante M_s (appelée aimantation à saturation). A très basse température, l'évolution du vecteur m au cours du temps est décrite par l'équation de Landau-Lifshitz-Gilbert (qui est donc une équation différentielle à valeurs vectorielles dans ce cas) :

$$\frac{dm}{dt} = m \wedge h_{\text{eff}}(m) - \alpha m \wedge (m \wedge h_{\text{eff}}(m)). \tag{3.1}$$

On suppose dans toute la suite que $M_s = 1$, pour simplifier. Ici, α est une constante positive (constante d'amortissement), le symbole $\wedge$ désigne le produit vectoriel sur $\mathbb{R}^3$, et $h_{\text{eff}}(m)$ est le "champ effectif", qui peut s'écrire :

$$h_{\text{eff}}(m) = h_{\text{ext}} - K\nabla G(m), \tag{3.2}$$

où h_{ext} correspond au champ extérieur (un vecteur donné de $\mathbb{R}^3$), $K = \pm 1$ et $G(m)$ est l'énergie d'anisotropie de la particule.

Dans la suite on prendra $K = 1$ et $G(m) = 1 - (m \cdot u)^2$, où u est un vecteur fixé de $\mathbb{R}^3$, de norme 1, de sorte que $\nabla G(m) = -2(m \cdot u)u$. Cette énergie correspond à une anisotropie uniaxe.

Dans tout l'énoncé, on notera $\mathbb{S}^2$ la sphère unité de $\mathbb{R}^3$, c'est-à-dire $\mathbb{S}^2 = \{v \in \mathbb{R}^3,\ |v| = 1\}$, et si u et v sont deux vecteurs de $\mathbb{R}^3$, on notera $u \cdot v$ leur produit scalaire. La notation I_3 désigne la matrice identité de taille 3×3.

Préliminaire : On rappelle que si a, b et c sont des vecteurs de $\mathbb{R}^3$, alors

(i) $|a \wedge b| \leq |a|\,|b|$,

(ii) $a \wedge b = -b \wedge a$, (antisymétrie)

(iii) $(a \wedge b) \cdot a = (a \wedge b) \cdot b = 0$, (orthogonalité du produit vectoriel)

(iv) $(a \wedge b) \wedge c = (a \cdot c)b - (b \cdot c)a$. (formule du double produit).

Soit $v = (v_1, v_2, v_3)^t \in \mathbb{S}^2$. Calculer, en fonction de v_1, v_2, v_3, les coefficients de la matrice M_v de l'endomorphisme de $\mathbb{R}^3$ donné par $x \mapsto v \wedge x$. Montrer que M_v est une matrice antisymétrique, et que $M_v M_v^t$ est la matrice de la projection orthogonale sur le plan orthogonal à v.

T1. Montrer que si $m \in C^1(\mathbb{R}^+;\mathbb{R}^3)$ est solution de (3.1), alors $|m(t)| = |m(0)|$, pour tout $t \geq 0$ (on pourra calculer $\frac{d|m(t)|^2}{dt}$).

On discrétise l'équation (3.1) au moyen d'un schéma d'Euler explicite de la manière suivante : soit $T > 0$, on fixe $N \in \mathbb{N}$ (grand) et on pose $\delta t = \frac{T}{N}$; étant donné $m_0 = m(0) \in \mathbb{S}^2$, on construit de manière récurrente une approximation m_k de $m(t_k)$, où $t_k = k\delta t$, $k = 0, \ldots, N$, en posant

$$\frac{m_{k+1} - m_k}{\delta t} = m_k \wedge h_{\text{eff}}(m_k) - \alpha m_k \wedge (m_k \wedge h_{\text{eff}}(m_k)), \tag{3.3}$$

où $h_{\text{eff}}(m_k)$ est donné par l'équation (3.2) avec $m = m_k$.

S1. Mettre en oeuvre le schéma (3.3) à l'aide d'un programme en Python pour simuler une approximation de $(m(t))_{t\in[0,T]}$. On visualisera la trajectoire de l'extrémité du vecteur $m(t)$, et, sur le même graphique, la sphère unité de $\mathbb{R}^3$ (voir indication ci-dessous). On pourra prendre différentes données initiales m_0. On prendra $\alpha = 0.1$, $T = 20$, $h_{\text{ext}} = (1, 0, 1)^t$, $u = (1, 0, 0)^t$. On fixera N assez grand pour que la contrainte $|m_k| = 1$ soit approximativement respectée, avec un ordre d'approximation raisonnable. Commenter. *Indication :* pour la visualisation, on pourra utiliser le module 3D de Matplotlib (voir `https://matplotlib.org/2.0.2/mpl_toolkits/mplot3d/tutorial.html`) et en particulier les tracés “wireframe" (pour tracer la sphère en fil de fer) ou “surface" (pour la tracer en surface pleine), ainsi que “scatter" pour visualiser la trajectoire de l'extrémité du vecteur $m(t)$ au cours du temps.

On s'intéresse maintenant au cas où la température n'est plus proche du zero absolu. Dans ce cas, il est nécessaire de prendre en compte, pour l'évolution de l'aimantation $m(t)$, des fluctuations thermiques, qui peuvent être modélisées en rajoutant au champ effectif h_{eff} un terme de bruit blanc $\varepsilon\frac{d\xi}{dt}$, où $\varepsilon > 0$ est une constante. La définition précise de ce terme dépasse le cadre du projet, mais la propriété importante est que pour tous $0 = t_0 < t_1 < \cdots < t_n$, la famille de vecteurs aléatoires $(\xi(t_k) - \xi(t_{k-1}))_{1\leq k\leq n}$ est une famille indépendante, et que $\xi(t_k) - \xi(t_{k-1})$ est un vecteur aléatoire gaussien centré à valeurs dans $\mathbb{R}^3$, de matrice de covariance $(t_k - t_{k-1})I_3$.

Ainsi, il est naturel d'approcher $\frac{d\xi}{dt}$ au temps t_k par $\frac{\xi((k+1)\delta t)-\xi(k\delta t)}{\delta t} = \frac{\chi_{k+1}}{\sqrt{\delta t}}$, où $(\chi_k)_{0\leq k\leq N}$ est une famille indépendante identiquement distribuée de vecteurs de loi $\mathcal{N}(0, I_3)$, et de rajouter ce terme au champ effectif donné par l'équation (3.2).

La question suivante vise à montrer qu'il suffit de modifier h_{eff} uniquement dans le premier terme du membre de droite de (3.3).

T2. Soit χ un vecteur aléatoire à valeurs dans $\mathbb{R}^3$ de loi $\mathcal{N}(0, I_3)$. Soit m un vecteur de $\mathbb{S}^2$, et α un réel strictement positif. Montrer que le vecteur aléatoire $m \wedge \chi$ a la même loi que le vecteur aléatoire $\frac{1}{\sqrt{1+\alpha^2}}(m \wedge \chi - \alpha m \wedge (m \wedge \chi))$. Qu'en déduit-on sur le modèle (discrétisé) à considérer?

Dans toute la suite, on considère une famille $(\chi_k)_{k\in\mathbb{N}^*}$ indépendante identiquement distribuée de vecteurs de loi $\mathcal{N}(0, I_3)$.

S2. Reprendre la question S1 en remplaçant dans le premier terme du membre de droite de (3.3), $h_{\text{eff}}(m_k)$ par $h_{\text{eff}}(m_k) + \varepsilon\frac{\chi_{k+1}}{\sqrt{\delta t}}$. On reprendra les paramètres et la visualisation de la question S1, et on prendra de plus $\varepsilon = 0.1$. Tracer la courbe de $|m(t)|$ en fonction de t. Commenter.

La question S2 montre que la discrétisation précédente n'est pas efficace pour préserver la contrainte $|m(t)| = 1$, même approximativement. On se propose dans les questions qui suivent d'étudier d'autres discrétisations, et on se concentre, pour commencer, sur l'équation

$$\frac{dm}{dt} = m \wedge \frac{d\xi}{dt}. \tag{3.4}$$

On propose une seconde discrétisation (dite “point milieux", ou semi-implicite) sous la forme

$$\frac{m_{k+1} - m_k}{\delta t} = m_{k+\frac{1}{2}} \wedge \frac{\chi_{k+1}}{\sqrt{\delta t}}, \text{ avec } m_{k+\frac{1}{2}} = \frac{1}{2}(m_k + m_{k+1}). \tag{3.5}$$

T3. **a.** Montrer que si m_k et m_{k+1} vérifient (3.5), alors $|m_{k+1}| = |m_k|$ (on pourra calculer $(m_{k+1} - m_k) \cdot m_{k+\frac{1}{2}}$).

b. Montrer que le schéma (3.5) peut se mettre sous la forme $M^+_{k+1} m_{k+1} = M^-_{k+1} m_k$, avec $M^\pm_{k+1} = I_3 \pm \frac{\sqrt{\delta t}}{2} \Gamma_{k+1}$ où Γ_k est la matrice de l'endomorphisme de $\mathbb{R}^3$ défini par $x \mapsto \chi_k \wedge x$.

c. Montrer que M^+_{k+1} est toujours inversible (on pourra considérer son noyau). En déduire que, étant donné $m_0 \in \mathbb{S}^2$, on peut construire, à l'aide du schéma (3.5) une suite $(m_k)_{0 \le k \le N}$, à valeurs dans $\mathbb{S}^2$, telle que m_{k+1} ne dépend que de m_k et d'un aléa indépendant de $(m_l)_{l \le k}$ (c'est une chaine de Markov).

T4. On fixe $m_0 \in \mathbb{S}^2$.

a. Montrer que si m_{k+1} est solution de (3.5), avec $m_k \in \mathbb{S}^2$, alors m_{k+1} admet un développement limité de la forme

$$\begin{aligned} m_{k+1} &= m_k + \sqrt{\delta t}\, m_k \wedge \chi_{k+1} + \frac{\delta t}{2}(m_k \wedge \chi_{k+1}) \wedge \chi_{k+1} \\ &\quad + \frac{1}{4}(\delta t)^{\frac{3}{2}} \big((m_{k+\frac{1}{2}} \wedge \chi_{k+1}) \wedge \chi_{k+1}\big) \wedge \chi_{k+1}. \end{aligned}$$

b. Montrer que $\mathbb{E}\big((m_k \wedge \chi_{k+1}) \wedge \chi_{k+1}\big) = -2\mathbb{E}(m_k)$.

c. Déduire de a. et b. que $|\mathbb{E}(m_{k+1}) - (1 - \delta t)\mathbb{E}(m_k)| \le C_1 (\delta t)^{\frac{3}{2}}$, pour une constante C_1 ne dépendant ni de k ni de δt (on pourra utiliser le fait que $m_k, m_{k+1} \in \mathbb{S}^2$); montrer ensuite que

$$\sup_{k=0,\dots,N} |\mathbb{E}(m_k) - (1 - \delta t)^k \mathbb{E}(m_0)| \le C_2 \sqrt{\delta t},$$

pour une constante C_2 qui dépend de T, mais pas de k ni de N (on rappelle que $\delta t = \frac{T}{N}$).

d. **(facultatif)** Montrer que si on considère, pour $t \in [0, T]$, la fonction aléatoire m^N définie par $m^N(t) = m_k$ pour $t \in [k\delta t, (k+1)\delta t[$, alors pour tout $t \in [0, T]$, $\mathbb{E}(m^N(t))$ converge, lorsque $N \to \infty$ (ou de manière équivalente lorsque $\delta t \to 0$) vers une fonction $\bar{m}(t)$ que l'on précisera.

S3. Mettre en oeuvre le schéma (3.5) à l'aide d'un programme en Python pour visualiser quelques trajectoires approchées de l'extrémité du vecteur $m(t)$ solution de (3.4) sur la spère $\mathbb{S}^2$ au cours du temps. Tracer sur un même graphique une approximation de l'espérance empirique de $m(t)$, et la fonction $t \mapsto e^{-t}\mathbb{E}(m_0)$ en fonction de $t \in [0, T]$. Commenter.

Indication : On pourra utiliser le module `scipy.linalg` pour la résolution du système linéaire.

T5. Quels sont les avantages et les inconvénients du schéma (3.5) ?

On propose une troisième discrétisation de l'équation (3.4) (appelée schéma projeté) consistant à utiliser un schéma explicite, mais en rajoutant à chaque pas de temps une étape de projection sur la sphère $\mathbb{S}^2$. Ainsi, le schéma s'écrit :

$$v_{k+1} = m_k + \sqrt{\delta t}\, m_k \wedge \chi_{k+1}, \text{ et } m_{k+1} = \frac{v_{k+1}}{|v_{k+1}|}. \tag{3.6}$$

T6. **a.** Montrer que, étant donné $m_0 \in \mathbb{S}^2$, le schéma (3.6) permet de définir une nouvelle suite de v.a. $(m_k)_{k=0,\dots,N}$, à valeurs dans $\mathbb{S}^2$, telle que m_{k+1} ne dépend que de m_k et d'un aléa indépendant de m_k. Quels sont les avantages de ce schéma?

b. **(facultatif)** Effectuer un développement limité à l'ordre $O(\delta t)^{\frac{3}{2}}$ de m_{k+1}, et montrer que si m_{k+1} est donné par le schéma (3.6) et $\tilde{m}_{k+1}$ est donné par le schéma (3.5) (avec le même m_k et la même réalisation de χ_{k+1} dans les deux cas) alors $\mathbb{E}(m_{k+1} - \tilde{m}_{k+1} | m_k) = O(\delta t)^{\frac{3}{2}}$. On pourra utiliser le fait que pour a, b, c, vecteurs de $\mathbb{R}^3$, $(a \wedge b) \cdot c = (b \wedge c) \cdot a$.

S4. Mettre en oeuvre le schéma (3.6) à l'aide d'un programme en Python. On tracera sur un même graphique une réalisation de la trajectoire de l'extrémité du vecteur $m(t)$ pour chacun des deux schémas (3.5) et (3.6), pour la même donnée initiale m_0, et la même réalisation de la suite $(\chi_k)_k$. Commenter.

S5. **(facultatif)** En se basant sur l'idée du schéma projeté, c'est-à-dire en considérant le schéma explicite (3.3) mais en rajoutant à chaque pas de temps une étape de projection de m_{k+1} sur la sphère $\mathbb{S}^2$, reprendre les simulations de la question S2 (avec les mêmes paramètres) en faisant varier l'amplitude du bruit ε entre 0.01 et 0.1. Commenter.

PROJET 4

Modélisation d'une file d'attente

sujet proposé par A. de Bouard

anne.debouard@polytechnique.edu

On souhaite modéliser une situation dans laquelle des clients arrivent à des temps aléatoires à un serveur (qui peut être le guichet d'une banque ou d'un bureau de poste, la caisse d'un supermarché, une station service, ... ou encore un serveur informatique, dont les clients sont les taches que le serveur doit traiter). On supposera que les intervalles de temps entre l'arrivée de deux clients forme une famille i.i.d. de variables aléatoires de loi exponentielle de paramètre λ.

On supposera également que les temps de service (qui sont donc égaux aux intervalles de temps entre le départ de deux clients de la file, tant que celle-ci ne se vide pas) sont i.i.d. de loi exponentielle de paramètre μ.

On conviendra que le client en train d'être servi compte dans l'effectif de la file d'attente.

S1. Ecrire un programme permettant de calculer le nombre de clients X_t dans la file d'attente, en fonction du temps t (on supposera qu'au temps $t = 0$, la file est vide). Expliquer l'algorithme utilisé. On tracera sur un même graphique 5 simulations dans chacun des trois cas : $\mu = \frac{\lambda}{2}$, $\mu = \lambda$ et $\mu = 2\lambda$. On prendra $\lambda = 1$, et on pourra arrêter les simulations à l'arrivée du trentième client. Commenter les graphiques obtenus.

T1. On suppose qu'un client arrive au temps t, et que le serveur est alors occupé par un autre client, depuis un temps $t_s > 0$.

a. Montrer que la loi du temps que devra attendre le client arrivé au temps t avant de voir l'autre client libérer le serveur ne dépend pas de t_s. Quelle est cette loi ?

b. Quelle est la probabilité que le prochain "événement" après le temps t (c'est à dire l'arrivée d'un nouveau client, ou le départ d'un client ayant été servi) soit une arrivée ? un départ ?

On admettra que les intervalles de temps entre deux événements (arrivée d'un nouveau client, ou départ d'un client ayant été servi) forment une suite i.i.d. de v.a. exponentielles de paramètre $\lambda + \mu$, tant qu'il reste au moins un client dans la file.

T2. On considère une suite $(B_k)_{k\geq 1}$ de variables aléatoires de Bernoulli de paramètre $\frac{\lambda}{\lambda+\mu}$ et une suite $(\tau_k)_{k\geq 1}$ de v.a. i.i.d. (indépendante de la suite (B_k)) de loi exponentielle de paramètre $\lambda+\mu$.

a. On pose $N(\omega) = \inf\{k \geq 1,\ B_k(\omega) = 1\}$ et $T = \sum_{k=1}^{N} \tau_k$. Calculer, pour $t > 0$ la probablité $\mathbb{P}(T > t)$, et en déduire que T suit une loi exponentielle de paramètre λ.

b. Déduire de ce qui précède que, quitte à introduire des "événements fictifs" (correspondant à des départs fictifs de clients alors que la file est vide), le nombre de clients X_t dans la file (toujours dans le cas où celle-ci est vide au temps $t = 0$) peut être représenté sous la forme $X_t = Y_{N_t}$, où N_t est le processus de comptage de la suite $(\tau_k)_{k\geq 1}$, c'est à dire que pour chaque $t > 0$,

$$N_t(\omega) = \sup\{n \geq 0,\ \sum_{k=1}^{n} \tau_k(\omega) \leq t\},$$

(N_t représente ainsi le nombre d'événements, éventuellement fictifs, ayant eu lieu avant le temps t), et où $(Y_n)_{n\geq 0}$ est une suite de v.a. à valeurs dans $\mathbb{N}$, définie comme une récurrence aléatoire, c'est à dire de la forme $Y_{n+1} = f(Y_n, B_{n+1})$, où f est une fonction déterministe **que l'on précisera**.

c. En déduire les valeurs de $\mathbb{P}(Y_{n+1} = j | Y_n = i)$ pour $i, j \in \mathbb{N}$ (les valeurs $(p_{i,j})_{i,j\in\mathbb{N}}$ ainsi définies forment la matrice de transition de la chaine de Markov $(Y_n)_{n\in\mathbb{N}}$).

S2. Utiliser la représentation de la question précédente pour effectuer de nouvelles simulations, en reprenant les paramètres de la question S1 (on supposera toujours que la file est vide au temps $t = 0$). Comparer les deux méthodes de simulation.

T3. Dans cette question, on cherche les éventuelles lois stationnaires de la suite $(Y_n)_{n\geq 0}$, c'est à dire que l'on cherche une **mesure de probabilité** $(\pi(k))_{k\geq 0}$ sur $\mathbb{N}$ telle que, pour tout $n \in \mathbb{N}$, si $\mathbb{P}(Y_n = k) = \pi(k)$ pour tout $k \in \mathbb{N}$, alors $\mathbb{P}(Y_{n+1} = k) = \pi(k)$ pour tout $k \in \mathbb{N}$. On admettra qu'une telle loi stationnaire, si elle existe, décrit l'état limite de X_t lorsque t tend vers l'infini.

a. Soit $\pi = (\pi(k))_{k\geq 0}$, une loi stationnaire pour $(Y_n)_{n\geq 0}$. En conditionnant l'événement $\{Y_{n+1} = k\}$ par les valeurs possibles de Y_n, trouver une relation entre $\pi(k-1)$, $\pi(k)$ et $\pi(k+1)$, pour tout $k \geq 1$, ainsi qu'entre $\pi(0)$ et $\pi(1)$.

b. En déduire toutes les lois stationnaires (quand elles existent) dans chacun des cas suivants:

i) $0 < \lambda < \mu$

ii) $\lambda = \mu > 0$

iii) $0 < \mu < \lambda$

Commenter ces résultats en comparant aux simulations des questions S1 et S2.

T4. On suppose dans cette question que $0 < \lambda < \mu$ et on se place en régime stationnaire, c'est à dire que l'on suppose que pour tout n, Y_n est distribuée suivant la loi π telle que $\pi(k) = \left(\frac{\lambda}{\mu}\right)^k \left(1 - \frac{\lambda}{\mu}\right)$, pour $k \in \mathbb{N}$.

a. Calculer la probabilité que le serveur soit occupé et la longueur moyenne de la file (c'est à dire le nombre moyen de clients dans la file). Que se passe t'il lorsque λ tend vers μ ?

b. On note W le temps d'attente d'un client avant d'être servi. Quelle est la probabilité que W soit nul ? Quelle est la loi de W si on suppose qu'il y a exactement k personnes dans la file lors de l'arrivée du client (c'est à dire la loi de W sachant $Y_n = k$) ? En déduire la loi de W, ainsi que sa moyenne, et la loi conditionnelle de W sachant $W > 0$. Calculer le temps total moyen passé par un client dans la file.

c. Montrer que le temps τ entre le départ (réel) d'un client de la file et le départ réel du client suivant suit une loi exponentielle de paramètre λ (on pourra décomposer la loi de τ suivant qu'il reste au moins un client dans la file, ou que la file est vide).

S3. Retrouver numériquement les résultats des questions T3 et T4 en effectuant un grand nombre de simulations. Représenter graphiquement les histogrammes de la mesure π et de W, en prenant pour paramètres $\lambda = 1$ et $\mu = 2$, et les comparer avec les valeurs théoriques obtenues.

On considère maintenant deux situations, A et B, dans lesquelles on dispose de deux serveurs pouvant servir les clients en parallèle. Les clients arrivent toujours après des temps i.i.d exponentiels de paramètre λ et les temps de service de chacun des serveurs sont également i.i.d. exponentiels de paramètre λ.

T5. Dans la situation A, les clients forment une seule file et choisissent le premier serveur qui se libère. Si aucun des deux serveurs n'est occupé, ils choisissent l'un des deux serveur au hasard avec probabilité $1/2$.

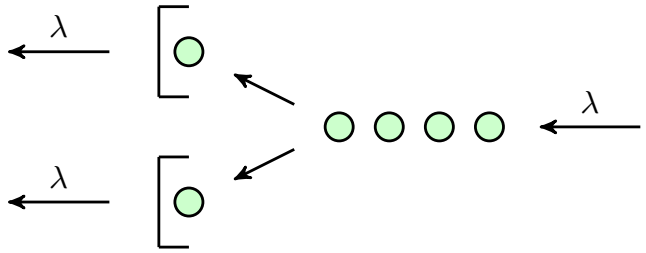


Figure 4.1: situation A

a. Montrer que les clients quittent la file unique pour se diriger vers l'un des serveurs à des temps exponentiels de paramètre 2λ.

On admet que le nombre de clients dans la file (y compris les éventuels clients en train d'être servis) peut se mettre sous la forme $X_t = Z_{N_t}$ où cette fois N_t est le processus de comptage d'une suite i.i.d. de variables aléatoires exponentielles de paramètre 3λ (correspondant aux temps entre deux événements éventuellements fictifs, de type arrivée d'un nouveau client, départ d'un client du premier serveur ou départ d'un client du second serveur), et $(Z_n)_{n\in\mathbb{N}}$ est une chaine de Markov.

b. On se place au temps $t > 0$. Quelle est la probabilité pour que le prochain "événement" soit une arrivée ? le départ d'un client du premier serveur ? le départ d'un client du second serveur ? En déduire les probabilités de transition $\mathbb{P}(Z_{n+1} = j|Z_n = i)$ pour $i, j \in \mathbb{N}$.

c. Montrer que $(Z_n)_{n\in\mathbb{N}}$ admet une unique loi stationnaire que l'on calculera.

d. On note toujours W le temps d'attente d'un client avant d'être servi, en régime stationnaire. Calculer $\mathbb{P}(W = 0)$, puis la loi de W (on pourra commencer par montrer que la loi de W sachant $Y_n = k$ est une loi $\Gamma(k-1, 2\lambda)$ si $k \geq 2$) . En déduire la moyenne de W.

T6. Dans la situation B, il y a une file distincte devant chaque serveur; les clients choisissent une file ou l'autre au hasard avec probabilité $1/2$.

Expliquer pourquoi du point de vue du client, cette situation est équivalente à celle de la question T2 avec des temps d'arrivées de clients suivant des v.a. exponentielles de paramètre $\lambda/2$ et des temps de service exponentiels de paramètre λ. En déduire la probabilité, en régime stationnaire, qu'un client ne doive pas attendre, et la moyenne du temps d'attente. Comparer les situations A et B du point de vue du client.

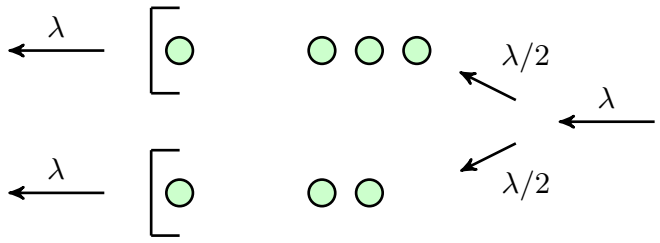


Figure 4.2: situation B

S4. Retrouver les résultats des questions T5 et T6 à l'aide d'un grand nombre de simulations; on comparera pour les situations A et B :

- l'espérance empirique et la variance empirique du temps d'attente du 50ième client.
- la probabilité pour que le 50ième client soit servi tout de suite.
- la longueur moyenne totale de la file.

Commenter.

PROJET 5

Étude et simulation d'un bruit poissonnien (shot noise)

sujet proposé par A. de Bouard

anne.debouard@polytechnique.edu

Le but du projet est d'étudier et de simuler un modèle pouvant décrire l'influx nerveux dans les neurones, sous la forme d'une suite d'impulsions provoquées par un potentiel d'action, apparaissant à des temps aléatoires, et dont l'intensité dépend du potentiel post-synaptique. On ne prend en compte ici que des potentiels excitateurs, et le modèle que l'on considère est couramment appelé bruit poissonnien de grenaille (ou Poisson shot noise en anglais).

Le potentiel est alors donné en fonction du temps t sous la forme

$$V_t = \sum_{n=1}^{\infty} Z_n F(t - T_n), \tag{5.1}$$

où les temps aléatoires T_n représentent les temps d'excitations, les variables aléatoires (positives) Z_n représentent les intensités des impulsions et la fonction (déterministe et positive) F décrit la décroissance du potentiel juste après l'excitation.

On s'intéresse tout d'abord à la modélisation et à la simulation des temps T_n. On considère pour cela une suite $(\tau_k)_{k\in\mathbb{N}^*}$ de variables aléatoires indépendantes de loi $\mathcal{E}(\lambda)$, exponentielle de paramètre λ. On pose alors, pour $n \in \mathbb{N}^*$, $T_n = \sum_{k=1}^{n} \tau_k$, et, par convention, $T_0 = 0$.

T1. Quelle est la loi de la variable aléatoire T_n ? Calculer $\mathbb{E}(e^{-T_n})$ et montrer que, presque surement, e^{-T_n} décroît vers 0 lorsque n tend vers l'infini. En déduire que T_n tend vers l'infini presque surement lorsque n tend vers l'infini.

Dans toute la suite, on supposera que la fonction F intervenant dans l'expression du potentiel V_t (voir l'équation (5.1)) vérifie la propriété suivante :

$$F(s) = 0, \text{ pour tout } s < 0.$$

Cette propriété traduit le fait que l'excitation arrivant au temps T_n ne contribue pas au potentiel V_t avant T_n. Ainsi, on déduit de la question 2) que, si $t \geq 0$ est fixé, alors, presque surement, card$\{n,\ T_n \leq t\}$ est fini, et donc la somme intervenant dans l'expression de V_t est également finie, de sorte que V_t est à valeurs réelles.

Pour $t \geq 0$ fixé, on considère la variable aléatoire N_t décrivant le nombre d'excitations générées avant le temps t, ou de manière équivalente, l'indice N du dernier temps T_N précédent le temps T. Ainsi,

$$N_t = \sup\{n \in \mathbb{N}, 0 \leq T_n \leq t\}.$$

La famille de variables aléatoire (N_t) est appelée processus de Poisson simple issu de 0, d'intensité λ.

T2. Calculer pour $t \geq 0$ fixé, la probabilité $\mathbb{P}(t \in [T_n, T_{n+1}[)$, et en déduire que la variable aléatoire N_t suit une loi de Poisson de paramètre λt. Quel est le nombre moyen d'excitations générées avant le temps t ?

T3. Soit $n \in \mathbb{N}^*$, $t > 0$ et $U_1, U_2, \ldots, U_n$, n v.a. indépendantes de loi commune la loi uniforme sur $[0, t]$. On note $U_{(1)}, U_{(2)}, \ldots, U_{(n)}$ les statistiques d'ordre associées à $(U_1, \ldots, U_n)$, c'est à dire que $(U_{(1)}, U_{(2)}, \ldots, U_{(n)})$ est le vecteur $(U_1, U_2, \ldots, U_n)$ ré-ordonné dans l'ordre croissant. En particulier

$$U_{(1)} = \min\{U_1, \ldots, U_n\} \text{ et } U_{(n)} = \max\{U_1, \ldots, U_n\}.$$

Montrer que le vecteur aléatoire $(U_{(1)}, U_{(2)}, \ldots, U_{(n)})$ a pour densité

$$f_{(U_{(1)}, U_{(2)}, \ldots, U_{(n)})}(u_1, \ldots, u_n) = \frac{n!}{t^n} 1_{\{0 \leq u_1 \leq \cdots \leq u_n \leq t\}}.$$

T4. Montrer que la loi conditionnelle de $(T_1, \ldots, T_n)$ sachant $N_t = n$ est la même que celle de $(U_{(1)}, U_{(2)}, \ldots, U_{(n)})$.

S1. En déduire un algorithme de simulation d'un processus de Poisson sur un intervalle de temps $[0, T]$ fixé. On décrira l'algorithme. Le mettre en oeuvre à l'aide d'un programme en Python, et tracer quelques trajectoires du processus (c'est à dire quelques réalisations de N_t en fonction de $t \in [0, T]$) en prenant $\lambda = 5$ et $T = 1$, puis, sur un graphique différent, $T = 20$. Commenter.

Dans toute la suite, on supposera que la famille de variables aléatoires $(Z_n)_{n \in \mathbb{N}^*}$ donnant les intensités des impulsions est une famille de v.a. indépendantes et identiquement distribuées, indépendante de $(T_n)_{n \in \mathbb{N}^*}$, à valeurs dans $\mathbb{R}^+$, et intégrables. Quant au profil F décrivant la décroissance du potentiel, on supposera que

$$F(s) = \begin{cases} e^{-\alpha s} & \text{si} \quad s \geq 0 \\ 0 & \text{si} \quad s < 0 \end{cases} \tag{5.2}$$

où $\alpha > 0$ est une constante fixée.

On s'intéresse alors à la variable aléatoire $X_n = V_{T_n}$, où V_t est donnée par (5.1) et (5.2).

T5. Montrer que $X_{n+1} = f(X_n, \xi_{n+1})$ où ξ_{n+1} est un vecteur aléatoire (à valeurs dans $\mathbb{R}^2$) indépendant de X_n, et f est une fonction déterministe que l'on précisera. On dit que $(X_n)_n$ est une chaine de Markov (noter que son espace d'état n'est pas dénombrable). On précisera la loi de ξ_n.

S2. On suppose dans cette question que la loi de Z_1 est une loi uniforme sur $[0, 1]$. Préciser la fonction de répartition de la variable aléatoire $e^{-\alpha \tau_1}$. En déduire un algorithme de simulation de $(X_n)_{n \in \mathbb{N}^*}$ (que l'on décrira), à partir de la simulation de v.a. uniformes sur $[0, 1]$. Mettre celui-ci en oeuvre pour simuler quelques réalisations des 100 premières valeurs de la suite X_n, dans chacun des trois cas : a) $\alpha = \lambda/2$, b) $\alpha = \lambda$, et c) $\alpha = 2\lambda$. On prendra $\lambda = 1$. Commenter.

S3. En se basant sur l'algorithme de la question S1, tracer les trajectoires de V_t pour $t \in [0, 10]$, pour 5 réalisations, dans chacun des trois cas de la question S2, en prenant $\lambda = 1$. On décrira la méthode utilisée.

T6. Montrer que pour tout $n \in \mathbb{N}^*$, X_n a la même loi que $\tilde{X}_n = \sum_{k=1}^n e^{-\alpha T_{k-1}} Z_k$. Montrer que $\tilde{X}_n$ converge en moyenne vers une variable aléatoire X, et en déduire que X_n converge vers X en loi.

S4. Reprendre les simulations de la question S2 pour la suite $\tilde{X}_n$. Commenter.

S5. Tracer un histogramme approché de la loi de X, a l'aide de la suite X_n (ou $\tilde{X}_n$) pour n assez grand, pour $\alpha = \lambda$, $\alpha = \lambda/2$ et $\alpha = 2\lambda$. On prendra dans cette question $Z = 1$ avec probabilité 1, et toujours $\lambda = 1$.

T7. a. Soit $n \in \mathbb{N}^*$; montrer que si on note A_n l'événement

$$A_n = \{N_t = n\} = \{T_n \leq t,\ T_{n+1} > t\},$$

alors

$$\sum_{k=n+1}^{+\infty} \mathbb{E}(e^{-\alpha T_k}|A_n) \leq e^{-\alpha t}\left(\frac{\lambda+\alpha}{\alpha}\right)$$

(on pourra utiliser l'indépendance de $(\tau_k)_{k\geq n+2}$ avec A_n).

b. Calculer, pour $t > 0$ fixé, la fonction de répartition de la v.a. T_{N_t}. Cette v.a. admet-elle une densité ? En déduire la valeur de $\mathbb{E}(e^{-\alpha T_{N_t}})$ (on pourra noter que

$$\mathbb{E}(e^{-\alpha T_{N_t}}) = \mathbb{P}(T_{N_t} = 0) + \mathbb{E}(e^{-\alpha T_{N_t}} 1_{T_{N_t}>0}),$$

et que la v.a. $e^{-\alpha T_{N_t}} 1_{T_{N_t}>0}$ admet une densité).

c. Déduire des questions a. et b. que $\tilde{X}_{N_t}$ converge en moyenne vers X lorsque t tend vers l'infini.

S6. Illustrer la convergence précédente en calculant une valeur approchée de $\mathbb{E}(|\tilde{X}_{N_t} - X|)/\mathbb{E}(X)$ pour t assez grand, toujours en se basant sur l'algorithme de la question S1. On choisira t suffisamment grand pour que l'erreur relative soit inférieure à 5%. On prendra $\lambda = \alpha = 1$, et Z constante égale à un.

T8. En utilisant le lemme de Borel-Cantelli, montrer que la suite $(\tilde{X}_n)_n$ converge presque sûrement, et retrouver le résultat de la question T7.c.

T9. (Question facultative) Montrer que la suite $(X_n)_n$ ne peut pas converger presque sûrement.

PROJET 6

Estimation de la densité: application à une population de bactéries

sujet proposé par M. Doumic

marie.doumic@inria.fr

On souhaite estimer la répartition des tailles d'une population de bactéries. Les biologistes savent en effet que cette répartition, dans des conditions expérimentales stables (nutriment et espace suffisamment en excès pour ne pas manquer), reste fixe au cours du temps, phénomène appelé *homéostasie* et confirmé par les modèles mathématiques de dynamique des populations. On suppose de plus que cette répartition de tailles notée f est "suffisamment régulière", dans un sens qui sera précisé ci-dessous, et est en tout cas une mesure à densité par rapport à la mesure de Lebesgue à support dans $(0, +\infty)$.

On dispose, pour estimer cette densité, d'un échantillon de n bactéries dont on mesure les tailles. On note ces tailles $x_1, \cdots, x_n$ et on suppose que ce sont les réalisations de variables aléatoires $X_1, \cdots, X_n$ indépendantes identiquement distribuées de loi f. La mesure empirique ν_n s'écrit donc

$$\nu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}(A),$$

et on veut utiliser la réalisation de la mesure empirique en $(x_1, \ldots, x_n)$ pour estimer f. De même, on définit la fonction de répartition empirique F_n par

$$F_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq x},$$

T1. Prouver/rappeler

$$F_n(x) \xrightarrow[n\to\infty]{p.s.} F(x).$$

Peut-on utiliser directement ν_n pour estimer une moyenne pondéré de f, par exemple le moment d'ordre p, défini par $\mu_p = \int_0^\infty x^p f(x) dx$, en le supposant fini? Peut-on utiliser directement ν_n pour estimer f: quels seraient les avantages et les inconvénients de ce choix?

T2. On note $\hat{f}_n(x) = \hat{f}_n(x; X_1, \ldots, X_n)$ un estimateur de f défini à partir des variables aléatoires (X_i) (en pratique, on définira $\hat{f}_n$ à partir de leurs réalisations $x_1, \ldots, x_n$), et on note $\hat{f}(x) = \mathbb{E}[\hat{f}_n(x)]$. On

définit pour chaque x, le *risque quadratique moyen* (Mean Squared Error ou MSE) par

$$MSE(x) := \mathbb{E}_{f^{\otimes n}}\left[(f(x) - \hat{f}_n(x))^2\right] = \int \dots \int \left(f(x) - \hat{f}_n(x; x_1, \dots, x_n)\right)^2 \prod_{i=1}^{n} f(x_i) dx_i.$$

Démontrer la décomposition biais-variance

$$MSE(x) = b^2(x) + \sigma^2(x)$$

avec

$$b(x) = f(x) - \hat{f}(x).$$

et

$$\sigma^2(x) = \mathbb{E}\left[\left(\hat{f}_n(x) - \hat{f}(x)\right)^2\right] = \mathbb{E}\left[\left(\hat{f}_n(x) - \mathbb{E}_{f^{\otimes n}}[\hat{f}_n(x)]\right)^2\right].$$

T3. Comme on souhaite obtenir un estimateur régulier de f utilisant f_n, l'idée des estimateurs à noyau est de régulariser f_n à l'aide de suites régularisantes. Soit $K : \mathbb{R} \to \mathbb{R}$ une fonction C_0^∞, *i.e.* infiniment dérivable et tendant vers 0 en l'infini ainsi que toutes ses dérivées, telle que $\int K(x)dx = 1$ et $\int x^k K(x)dx = 0$ pour $k = 1, 2, \cdots m$ avec $m \geq 0$ ($m = 0$ correspondant à une hypothèse vide sur les moments supérieurs ou égaux à 1).

Proposer des noyaux K tels que $m = 1$ et $m = 2$. Remarquez que K ne peut pas être de signe constant si $m \geq 2$.

On note (K_h) pour $0 < h \leq 1$ la suite régularisante définie par

$$K_h(x) := \frac{1}{h} K\left(\frac{x}{h}\right).$$

T4. On rappelle que le produit de convolution entre une fonction g et une mesure ν est défini par

$$g * \nu(x) = \int g(x - y)\nu(dy).$$

Donner la formule obtenue dans le cas où $g = K_h$ et $\nu = \nu_n$; on note

$$\hat{f}_n(x) := K_h * \nu_n.$$

Donner les expressions de $b^2(x)$ et de $\sigma^2(x)$ en fonction des X_i, de n, de K, de f et de h.

T5. Montrer que

$$\hat{f} = \mathbb{E}[\hat{f}_n(x)] = K_h * f.$$

Estimation du biais

Dans cette partie, on cherche à estimer

$$b(x) = f(x) - K_h * f(x).$$

T6. En utilisant que

$$f(x) - K_h * f(x) = \int K_h(y)(f(x) - f(x - y))dy, \tag{6.1}$$

puis en faisant le changement de variable $z = \frac{y}{h}$, montrer que si f et f' sont dans L^∞ alors pour tout $x > 0$ on a

$$|f(x) - K_h * f(x)| \leq Ch\|f'\|_{L^\infty},$$

où C est une constante dépendant de K que l'on précisera.

T7. En utilisant (6.1) et un développement de Taylor-Lagrange, montrer que pour $k \leq m$, si f est $k+1$ fois continûment dérivable, on a

$$|f(x) - K_h * f(x)| \leq Ch^{k+1} \|f^{(k+1)}\|_{L^\infty},$$

pour une constante C dépendant de K que l'on précisera.

Estimateur de la variance

T8. En décomposant la variance sous forme du carré d'une somme de n variables aléatoires indépendantes centrées, et en remarquant que pour tout i

$$K_h * f = \hat{f} = \mathbb{E}[K_h(x - X_i)]$$

montrer que

$$\|\sigma^2\|_{L^\infty} \leq \frac{1}{nh} \|K\|_{L^2}^2 \, \|f\|_{L^\infty}.$$

Estimation du risque quadratique

T9. En utilisant T7 et T8, donner une majoration du risque quadratique moyen pour f k fois dérivable, $k \leq m+1$. Qu'en déduire sur le choix de h? Exprimer l'ordre de grandeur d'un h permettant de minimiser le risque quadratique moyen en fonction de n et k. Quel est alors l'ordre de grandeur de l'erreur sur le MSE? Commenter.

Simulations

Pour les méthodes de simulation, vous pouvez vous référer au chapitre 5.5. du poly.

T10. On choisit, pour $p \geq 2$,

$$f(x) = x^p e^{-\frac{x^{p+1}}{p+1}}$$

et le noyau gaussien $K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. Vérifier que f définit bien une densité. Que vaut m pour K?

S1. Ecrire une fonction simulant un $n-$ échantillon de densité f à l'aide de l'algorithme du rejet.

S2. Ecrire une fonction simulant un $n-$ échantillon de densité f à l'aide de la méthode de la fonction inverse. Comparer les vitesses des deux méthodes.

S3. Ecrire une fonction retournant l'estimateur à noyau $\hat{f}_n$ pour un échantillon de taille n donné, un choix de K et un choix de h.

S4. Tracer l'estimateur pour différentes valeurs de h et différentes tailles d'échantillon n.

S5. Comme on connaît ici la vraie densité f, on peut mesurer l'erreur réellement effectuée $(f(x) - \hat{f}_n(x))^2$. A n fixé, tracer en échelle doublement logarithmique (logarithmique en abscisse et en ordonnée) cette erreur en fonction de plusieurs h. Prendre le h qui minimise l'erreur, et recommencer pour plusieurs n. Que remarquez-vous? Cela correspond-il au résultat de la question T9?

S6. Pour chaque n, choisir h d'ordre de grandeur optimal (cf. question T9.) de façon à minimiser l'erreur, et tracer en échelle doublement logarithmique l'erreur en fonction de h. Que remarquez-vous? Cela correspond-il au résultat de la question T9.?

Applications à des données sur les bactéries

S7. Chargez les données que vous trouverez à l'adresse:

`https://team.inria.fr/mamba/fr/francais-projets-map-361-donnees-a-telecharger/`

Ce fichier rassemble des données de colonies de bactéries: la première colonne correspond à des tailles à la naissance (en μm), la deuxième à des tailles à la division, la troisième des âges à la division. Proposer un estimateur pour chacune de ces trois distributions (à partir par exemple du noyau gaussien utilisé ci-dessus), respectivement notées f_b, f_d et f_a. Quel h choisir? Justifier votre réponse et commenter. Comparer les estimateurs choisis de f_b et de $2f_d(2x)$.

PROJET 7

Test d'adéquation entre deux populations: application à des cellules en division

sujet proposé par M. Doumic

marie.doumic@inria.fr

Motivation

On mesure la taille d'un échantillon de bactéries en division. L'analyse d'images permet d'avoir accès d'une part à la taille (bruitée) des cellules mères, d'autre part à la taille (bruitée) des cellules filles: on veut vérifier que la loi de la taille des filles est égale à la loi de la moitié de la taille des mères.

Cela s'inscrit dans une démarche plus générale qu'on peut théoriser ainsi: étant donnée l'observation de deux échantillons $x_1, \cdots, x_{n_1}$ et $y_1, \cdots, y_{n_2}$, que l'on suppose être la réalisation de lois $X_1, \cdots, X_{n_1}$ et $Y_1, \cdots, Y_{n_2}$ indépendantes identiquement distribuées, comment déterminer si la loi des observations de X_i est la même que la loi des observations de Y_i?

On note F_1 la fonction de répartition de la loi des X_i et F_2 celle des Y_i. La question est donc de savoir si l'on a $F_1 = F_2$ ou $F_1 \neq F_2$. En raison de la loi des grands nombres, on voit intuitivement que la réponse sera d'autant plus fiable que les échantillons n_1 et n_2 seront grands.

Définition d'un test d'adéquation

On rappelle tout d'abord le principe d'un test statistique (polycopié partie 10.1). On appelle **test d'adéquation simple** une fonction Ψ des observations à valeurs dans $\{0, 1\}$: on a

$$\Psi = \Psi_{n_1,n_2}(X_1, \cdots, X_{n_1}; Y_1, \cdots, Y_{n_2}) = \begin{cases} 0 \text{ si on accepte l'hypothèse } F_1 = F_2, \\ \\ 1 \text{ si on rejette l'hypothèse } F_1 = F_2. \end{cases}$$

On définit alors les deux types d'erreurs que l'on peut faire:

1. **L'erreur de première espèce** consiste à rejeter l'hypothèse $F_1 = F_2$ alors même qu'en réalité on a bien $F_1 = F_2$. Cela signifie qu'on observe $\{\Psi = 1\}$ sachant que $F_1 = F_2$: sa probabilité est

$$\mathbb{P}^{n_1,n_2}_{F_1=F_2}[\Psi = 1],$$

où $\mathbb{P}^{n_1,n_2}_{F_1=F_2}$ est la loi de $(X_1, \cdots, X_{n_1}; Y_1, \cdots, Y_{n_2})$ sachant que $F_1 = F_2$.

2. **L'erreur de seconde espèce** consiste à accepter l'hypothèse $F_1 = F_2$ alors que $F_1 \neq F_2$. De même, cela se produit avec une probabilité égale à

$$\mathbb{P}^{n_1,n_2}_{F_1 \neq F_2}[\Psi = 0]$$

Le but d'un bon test statistique est donc de "minimiser" les erreurs de première **et** de seconde espèce pour (F_1, F_2) donnés, sous certaines hypothèses concernant F_1 et F_2.

On dit qu'un test est **consistant** si son erreur de seconde espèce tend vers 0 pour tout couple (F_1, F_2) tel que $F_1 \neq F_2$ lorsque n_1 et n_2 tendent vers l'infini.

1 Populations de même taille: test du signe

On suppose dans cette section que $n_1 = n_2 = n$ (ce qui correspond à notre application décrite en introduction). On pose, pour $i = 1, \cdots, n$,

$$Z_i := \mathbb{1}_{X_i \geq Y_i},$$

et on suppose que F_1 et F_2 sont continues.

T1. Montrer que si $F_1 = F_2$, alors $\mathbb{P}[Z_i = 1] = \frac{1}{2}$.

T2. On suppose $F_1 = F_2$. Quelle est la loi de $T_n = \sum_{i=1}^{n} Z_i$?

T3. En déduire que pour tout $\alpha \in (0, 1)$, il existe une constante $t_{n,\alpha}$ ne dépendant que de n et de α telle que le test

$$\Psi_{n,\alpha} := \mathbb{1}_{|T_n - \frac{n}{2}| \geq t_{n,\alpha}}$$

ait une erreur de première espèce plus petite que α : cela définit un **test de niveau α.**

$t_{n,\alpha}$ est-il unique? Pourquoi a-t-on intérêt à le choisir le plus petit possible?

T4. En utilisant par exemple le théorème de la limite centrale, donner un équivalent de $\inf t_{n,\alpha}$ lorsque $n \to \infty$.

S1. Ecrire un programme Python qui, pour deux fonctions de répartition données, avec $F_1 = F_2$ ou pas, simule les $n-$ échantillons (X_i) et (Y_i) (on codera par exemple l'algorithme du rejet, ou bien on choisira une famille de fonctions de répartition connues et on appliquera la méthode de la fonction inverse: voir le poly, ch. 5.5. sur les méthodes d'inversion). Exemple de lois à simuler: $f(x) = x^k e^{-\frac{x^{k+1}}{k+1}}$ avec $k \geq 0$ et $x > 0$; on pourra aussi simuler des lois classiques (loi gaussienne ou loi log-normale etc) et comparer avec les solveurs et méthodes existant dans scipy.stats.

S2. Ecrire un programme Python qui, pour deux échantillons de même taille n (n variable) donnés, calcule $\Psi_{n,\alpha}$ (α variable).

S3. En répétant plusieurs expériences avec des échantillons simulés selon les lois choisies en S1, vérifier numériquement que le test est bien de niveau α. Donner une procédure qui donne un sens rigoureux à la phrase "vérifier numériquement".

S4. Chargez les données que vous trouverez à l'url:

`https://team.inria.fr/mamba/fr/francais-projets-map-361-donnees-a-telecharger/`

La première colonne consiste en des mesures expérimentales de tailles à la naissance, la seconde de tailles à la division d'une colonie de bactéries en division (ignorer la troisième colonne). Appliquer le test sur l'ensemble des données, pour plusieurs niveaux α. Que remarquez-vous? Pour $\alpha = 0.05$, appliquer également le test sur des sous-échantillons de taille $10, 100,$ etc. Que remarquez-vous? Commentez.

S5. Pour $\lambda > 0$, on choisit $F_1(x) = (1 - e^{-x})\mathbb{1}_{x\geq 0}$ et $F_2(x) = (1 - e^{-\lambda x})\mathbb{1}_{x\geq 0}$. Evaluer numériquement la fonction d'erreur de seconde espèce du test en fonction de α. Si on imagine n "grand" et $|1 - \lambda_n| \to 0$, quel est l'ordre de grandeur de λ_n qui permet à votre avis de rejeter l'hypothèse $F_1 = F_2^n$ (où F_2^n est la loi avec λ_n) avec une "bonne" précision?

T5. (Facultatif) Par un développement asymptotique de l'erreur de seconde espèce, justifiez théoriquement votre réponse à la question précédente.

T6. Le test $\Psi_{n,\alpha}$ est-il consistant ? Si la réponse est non, construire deux distributions $F_1 \neq F_2$ telles que l'erreur de seconde espèce ne tende pas vers 0.

S6. (Facultatif) Vérifiez numériquement la réponse proposée à la question précédente.

2 Populations de tailles différentes : test de Kolmogorov-Smirnov

Un défaut de l'approche précédente est (mis à part le problème de la consistance) que les tailles des deux échantillons doivent être identiques. Le test très connu de Kolmogorov-Smirnov permet d'éliminer ces deux inconvénients. On suppose toujours ici que F_1 et F_2 sont continues.

On définit les fonctions de répartition empiriques:

$$\hat{F}_1 := \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbb{1}_{x\geq X_i}, \qquad \hat{F}_2 := \frac{1}{n_2}\sum_{i=1}^{n_2} \mathbb{1}_{x\geq Y_i}.$$

T7. Rappeler/prouver que presque sûrement, pour tout x, on a

$$\hat{F}_1(x) \underset{n_1\to\infty}{\longrightarrow} F_1(x).$$

On admet le théorème de Glivenko-Cantelli qui prouve que cette convergence simple est en fait uniforme: on a

$$\lim_{n_1\to\infty} \sup_x |\hat{F}_1(x) - F_1(x)| = 0.$$

On définit

$$T_{n_1,n_2} := \sup_{x\in\mathbb{R}} \left|\hat{F}_1(x) - \hat{F}_2(x)\right|$$

et à partir de T_{n_1,n_2} on définit la fonction de test

$$\Psi_{n_1,n_2,\alpha} := \mathbb{1}_{T_{n_1,n_2}\geq t_{n_1,n_2,\alpha}}$$

où $t_{n_1,n_2,\alpha}$ correspond, comme pour le test du signe, au plus petit t tel que l'erreur de première espèce soit plus petite que α.

T8. Montrer que si $F_1 = F_2 = F$, la loi de T_{n_1,n_2} ne dépend pas de F. Pour cela, on montrera que $U_i = F(X_i)$ et $V_j = F(Y_j)$ sont des variables de loi uniforme sur $[0,1]$ et qu'on a l'égalité

$$T_{n_1,n_2} = \sup_{x\in\mathbb{R}} \left| \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbb{1}_{U_i\leq F(x)} - \frac{1}{n_2}\sum_{j=1}^{n_2} \mathbb{1}_{V_j\leq F(x)} \right|$$

S7. Ecrire une fonction python qui, pour deux échantillons X_i et Y_j de taille n_1 et n_2, calcule la statistique T_{n_1,n_2}. On pourra trier globalement les X_i et Y_j en un vecteur croissant Z_k de taille n_1+n_2, puis noter

$$S_k := \frac{1}{n_1}\mathbb{1}_{Z_k\in(X_i)} - \frac{1}{n_2}\mathbb{1}_{Z_k\in(Y_j)},$$

et montrer que

$$T_{n_1,n_2} = \max_{1\leq k\leq n_1+n_2} \left| \sum_{l=1}^{k} S_l \right|.$$

S8. Pour calculer $t_{n_1,n_2,\alpha}$ asymptotiquement, on admettra le résultat suivant:

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}} T_{n_1,n_2} \xrightarrow{\text{loi}} W \qquad \text{lorsque } \min(n_1,n_2)\to\infty$$

avec $\mathbb{P}(W\leq x) = \sum_{k=-\infty}^{+\infty} (-1)^k \exp(-2k^2x^2)$.

Ecrire une fonction python qui, pour un seuil $\alpha>0$ donné, calcule numériquement t_α tel que $\mathbb{P}(W\geq t_\alpha)=\alpha$.

S9. En déduire un calcul numérique de $t_{n_1,n_2,\alpha}$ dans l'approximation n_1 et n_2 grands.

S10. Sur un exemple que l'on choisira, vérifier numériquement que le test ainsi obtenu est asymptotiquement de niveau α.

S11. (Facultatif.) Etudier numériquement l'erreur de seconde espèce du test de Kolmogorov-Smirnov pour $F_1(x) = (1-e^{-x})\mathbb{1}_{x\geq 0}$ et $F_2(x) = (1-e^{-\lambda x})\mathbb{1}_{x\geq 0}$ en fonction de $\lambda>0$.

S12. Sur les données vues en S4, appliquer le test de Kolmogorov-Smirnov pour plusieurs niveaux α. Que remarquez-vous? Pour $\alpha = 0.05$, faites de même pour des sous-échantillons de taille n_1 et n_2 variables. Que remarquez-vous? Commentez.

PROJET 8

Propagation d'opinion chez les moutons

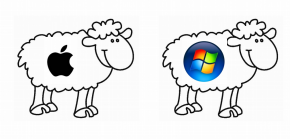
sujet proposé par L. Gerin

gerin@cmap.polytechnique.fr

Outils de MAP : Probabilités conditionnelles, Modélisation, Lois discrètes et continues, Suites récurrentes et matrices, ...

Mots-clés : Renforcement, Sensibilité aux conditions initiales,...

Le but du projet est d'étudier de façon théorique et expérimentale deux modèles simples de propagation d'opinions dans une population de moutons (on considère qu'il y a deux opinions concurrentes : les pro-Mac et pro-Windows). Le but est d'étudier différents aspects du phénomène de renforcement : l'opinion majoritaire a tendance à être renforcée au cours du temps[1].

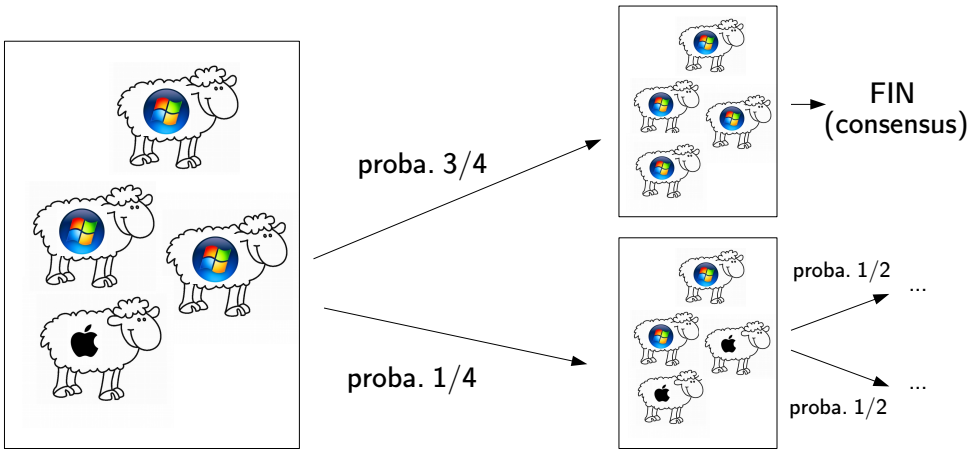


Partie A : Modèle à population fixe

Soit $N > 1$ fixé, on considère une population de N moutons, chacun est soit pro-Mac soit pro-Windows.

Initialement $0 \leq m \leq N$ moutons sont pro-Mac. A chaque instant $t = 0, 1, 2, ...$, un mouton est choisi uniformément dans la population (indépendamment du passé) et il bêle son opinion. Instantanément, un mouton de l'autre camp (s'il en reste un) change d'opinion. Le processus continue jusqu'à ce qu'un consensus pro-Mac ou pro-Windows soit atteint. Voici un exemple pour $N = 4, m = 1$:

[1] Le 2ème modèle a réellement été utilisé en Économie pour décrire une bataille entre deux technologies concurrentes, voir : W.B. Arthur. Self-reinforcing mechanisms in economics. *The economy as an evolving complex system*, vol.5, p.9-31 (1988).



Formellement, le processus $(M(t))_{t\geq 0}$ est donc défini de la façon suivante : $M(0) = m$ et pour $t \geq 0$ si $M(t) = k \notin \{0, N\}$ alors

$$M(t+1) = \begin{cases} k+1 \text{ avec proba. } \frac{k}{N} \\ k-1 \text{ avec proba. } \frac{N-k}{N} \end{cases}.$$

Si jamais $M(t) = 0$ (resp. $M(t) = N$) alors $M(t+1) = 0$ (resp. $M(t+1) = N$).

On se pose les question suivantes (en fonction de N, m):

- Que peut-on dire de la proportion de moutons pro-Mac à un instant donné?
- Quelle est la probabilité d'atteindre le consensus pro-Mac?

Quelques simulations

S1. Écrire un script python qui trace des simulations de trajectoires $(M(t))_{0\leq t\leq T}$ pour T donné. On choisira les paramètres suivants : $N = 2000, T = 4000$. Pour l'initialisation m, il est demandé d'afficher (si possible sur le même graphique) une trajectoire partant de chacune des valeurs suivantes : $m = 200, 600, 1000, 1400, 1800$.

Loi exacte à t fixé

On observe sur les simulations précédentes qu'il y a une forte dépendance en la condition initiale : l'opinion initialement majoritaire a une grande tendance à se répandre dans toute la population. L'objectif des questions qui suivent est de quantifier cette dépendance à l'aide de calculs explicites.

On fixe N pour toute la suite. Pour $k \in \{0, 1, \ldots, N\}$ et $t \geq 0$ on note

$$p(t,k) = \mathbb{P}(M(t) = k) = \mathbb{P}\left(\text{Au temps } t, \text{ exactement } k \text{ moutons sont pro-Mac}\right).$$

On a ainsi

$$p(0,k) = \begin{cases} 1 & \text{si } k = m, \\ 0 & \text{sinon} \end{cases}.$$

T1. Démontrer que pour tout $2 \leq k \leq N-2$ et pour $t \geq 1$ on a

$$p(t,k) = \frac{k-1}{N} p(t-1, k-1) + \frac{N-(k+1)}{N} p(t-1, k+1).$$

T2. Trouver les formules analogues pour $k \in \{0, 1, N-1, N\}$: écrire $p(t,k)$ en fonction de $p(t-1,0), p(t-1,1), \ldots, p(t-1,N)$. (Pour cette question il n'est pas besoin de justifier.)

T3. En déduire qu'il existe une matrice $Q = (q_{k,k'})_{0 \leq k,k' \leq N}$ de taille $(N+1) \times (N+1)$ telle que pour tout $t \geq 1$:

$$\Big(p(t,0), p(t,1), \ldots, p(t,N)\Big) = \Big(p(t-1,0), p(t-1,1), \ldots, p(t-1,N)\Big) \times Q.$$

S2. Ecrire une fonction python qui prend comme paramètres N et génère la matrice Q.
Afficher Q pour $N = 6$.

S3. Utiliser les questions précédentes pour écrire une fonction python qui calcule :

$$(t, k, N, m) \mapsto p(t,k).$$

Pour les paramètres $N = 100, t = 30, m = 70$, tracer la courbe $k \mapsto p(t,K)$.
(Pour vérifier cotre code, je trouve $p(30,86) = 0.1190564\ldots$ *avec* $N = 100, m = 70$*.)*

Probabilité de consensus

On peut démontrer (et on admet) que si l'on attend suffisamment longtemps alors presque-sûrement on atteint le consensus, soit pour Mac soit pour Windows. On définit donc

$$\alpha(m) = \mathbb{P}(\exists t, M(t) = N) = \mathbb{P}(\text{Le consensus est finalement atteint pour Mac})$$
$$\beta(m) = \mathbb{P}(\exists t, M(t) = 0) = \mathbb{P}(\text{Le consensus est finalement atteint pour Windows}).$$

Grâce à la remarque précédente on a $\alpha(m) + \beta(m) = 1$.

T4. Démontrer que

$$\alpha(m) = \lim_{t \to +\infty} p(t,N) \qquad (8.1)$$
$$\beta(m) = \lim_{t \to +\infty} p(t,0).$$

(Indication : Utiliser la propriété de monotonie des mesures de probabilités dans le Polycopié de MAP361.)

T5. Votre fonction python de la question S3 permet donc de calculer une approximation de $\alpha(m)$, en prenant t assez grand. Donner une méthode numérique permettant de calculer $\alpha(m)$ à 10^{-5} près.
(Indication : que peut-on dire de la suite $(p(t,N) + p(t,0))_{t \geq 0}$*?)*

S4. Tracer avec python les probas $m \mapsto \alpha(m)$ pour $N = 30$ (si vous n'avez pas résolu la question T5 vous pouvez considérer que $t = N^2$ est suffisant dans l'équation (8.1) pour que l'approximation soit correcte).
(Pour vérifier votre code : je trouve $\alpha(12) = 0.1324654....$*)*

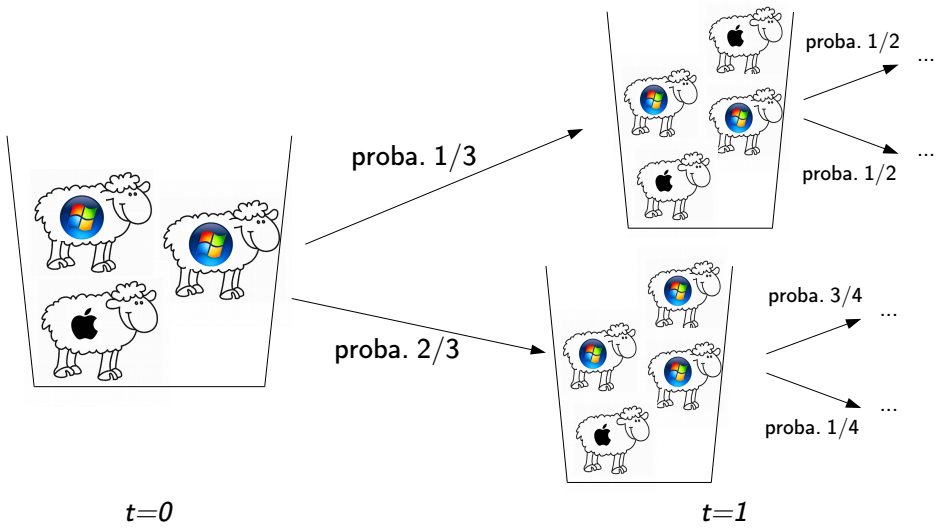
Partie B : Modèle avec population qui augmente

On considère cette fois une population constituée initialement de m moutons pro-Mac et w moutons pro-Windows. À chaque instant $t = 1, 2, ...$, un nouveau mouton arrive dans la population. Pour déterminer son opinion il choisit uniformément au hasard un mouton déjà présent et choisit la même opinion[2].

On note $Z(0) = m, W(0) = w$ et pour $t \geq 1$ on note $Z(t)$ (resp. $W(t)$) le nombre de moutons pro-Mac (resp. pro-Windows) immédiatement après l'instant t, c'est-à-dire lorsque le t-ème nouveau mouton est arrivé. On a pour tout t

$$Z(t) + W(t) = m + w + t.$$

Voici un schéma du début du processus, lorsque $m = 1, w = 2$:



Le processus $(Z(t))_{t\geq 0}$ est donc défini de la façon suivante : $Z(0) = m$ et pour $t \geq 0$ si $Z(t) = k$ et $W(t) = k'$ alors

$$Z(t+1) = \begin{cases} k+1 \text{ avec proba. } \frac{k}{k+k'} \\ k \text{ avec proba. } \frac{k'}{k+k'} \end{cases}.$$

Simulations de trajectoires

On fixe pour l'instant $m = w = 1$. On va s'intéresser au processus $(Z(t))_{t\geq 0}$.

S5. Écrire un script python qui trace (sur le même graphique) 15 trajectoires $(Z(t))_{0\leq t\leq T}$, pour $T = 800$.

Cas $m = w = 1$: uniformité

Sur vos simulations on doit normalement observer que sur chaque trajectoire $(Z(t))_t$ augmente à peu près linéairement avec t. On va donc étudier le comportement de la variable aléatoire $Z(t)/t$.

S6. Toujours dans le cas $m = w = 1$, écrire un script python qui réalise K tirages de la variable aléatoire $\frac{Z(T)}{T}$ et affiche les résultats dans un histogramme.
On prendra $K = 10000$, $T = 500$ et on représentera des histogrammes à 20 bâtons. Avec `matplotlib` un histogramme des valeurs `Donnees` avec 20 bâtons s'affiche avec

[2] Dans ce modèle les moutons ne changent jamais d'opinion.

```
plt.hist(Donnees, bins=20,ec='black')
```

(le `ec='black'` dessine les bords des bâtons).

T6. Sur votre histogramme on doit normalement observer que la variable aléatoire $\frac{Z(T)}{T}$ est répartie à peu près uniformément dans $[0,1]$. Démontrer par récurrence que, en effet, on a pour tout $t \geq 0$ que la variable aléatoire $Z(t)$ est uniforme dans l'ensemble

$$\{1, 2, 3, \ldots, t+1\}.$$

Cas $m, w \neq 1$: non-uniformité

S7. On prend maintenant $m = 2$, $w = 1$, écrire un script python qui réalise K tirages de la variable aléatoire $\frac{Z(T)}{T}$ et affiche les résultats dans un histogramme.
On prendra également $K = 10000$, $T = 500$ et on représentera des histogrammes à 20 bâtons.

T7. Décrivez et commentez l'histogramme. En particulier, est-ce que le résultat est intuitif?

Bonus: Une question théorique

Ces deux dernières questions bonus sont bien plus difficiles et ne seront pas notées très généreusement, elles ne sont destinées qu'aux fans de moutons et de produits infinis.

Les valeurs initiales m, w sont maintenant quelconques. Le but est de démontrer le résultat suivant (apparemment évident) : presque-sûrement la population de moutons pro-Mac ne cesse jamais d'augmenter, autrement dit :

$$\mathbb{P}\left(\exists t \text{ tel que } Z(t) = Z(t+1) = Z(t+2) = Z(t+3) = \ldots\right) = 0.$$

T8. Calculer pour tout $t < t'$ et tout $\mu \in \{1, 2, \ldots, t+1\}$ la probabilité

$$\mathbb{P}\left(Z(t) = Z(t+1) = \cdots = Z(t') \mid Z(t) = \mu\right).$$

T9. En déduire une majoration de

$$\mathbb{P}\left(Z(t) = Z(t+1) = \cdots = Z(t')\right).$$

et démontrer que pour tout t

$$\mathbb{P}\left(Z(t) = Z(t+1) = Z(t+2) = Z(t+3) = \ldots\right) = 0.$$

T10. Conclure.

PROJET 9

Chimie et absorption de dimères : le modèle de Flory

sujet proposé par L. Gerin

gerin@cmap.polytechnique.fr

Paul J. Flory (Prix Nobel de Chimie en 1974) a introduit en 1939 un modèle aléatoire simple pour étudier l'absorption de molécules très simples (des *dimères*) sur un substrat :

P.J.Flory. Intramolecular reaction between neighboring substituents of vinyl polymers.
Journal of the American Chemical Society, vol.61, n.6, p.1518-1521.

L'objectif du projet est d'analyser et simuler ce modèle.

Partie A : Modèle de Flory

Définition et simulation du modèle

Soit $N \geq 2$ fixé, on modélise le *substrat* par l'ensemble $\{1, 2, \ldots, N\}$. Sur ce substrat vont s'attacher des molécules très simples de longueur 2, appelées *dimères*[1]. Pour $t \in \mathbb{N}$ et $1 \leq k \leq N$ on va noter $M_t(k) = 0$ ou 1 selon que la position k est libre ou occupée par un dimère à l'instant t. Initialement tout est vide :

$$M_0(1) = M_0(2) = M_0(3) = \cdots = M_0(N) = 0.$$

L'évolution du processus $t \mapsto (M_t(k))_{t \geq 0, k \leq N}$ est définie ainsi :

- Pour chaque $t \geq 1$ on tire une variable aléatoire U_t uniforme[2] dans $\{1, 2, \ldots, N-1\}$. Les (U_t) sont supposés indépendants.
    - Supposons que $U_t = u$, deux situations sont possibles (voir le schéma plus bas) :

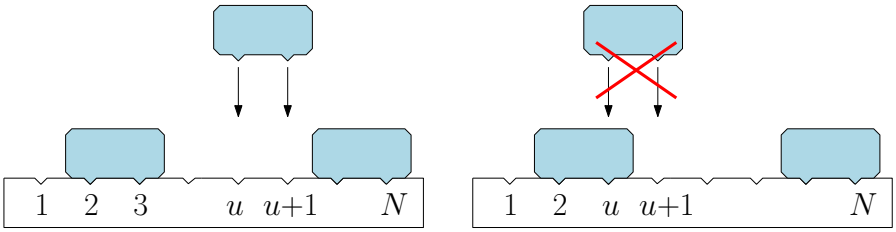
[1] Dans l'article original de Flory il s'agit de méthylvinylcétone.

[2] L'hypothèse d'uniformité est justifiée par le fait que la solution substrat+dimères est suffisamment mélangée pour que les dimères soient répartis de façon homogène.

- Si jamais les positions $u, u+1$ sont libres (c'est-à-dire $M_t(u) = M_t(u+1) = 0$) alors un dimère vient s'accrocher en $(u, u+1)$:

$$M_{t+1}(u) = M_{t+1}(u+1) = 1$$

  et les $N-2$ autres coordonnées restent inchangées : $M_{t+1}(v) = M_t(v)$ pour tout $v \notin \{u, u+1\}$. C'est le cas de gauche dans le schéma.
- Si l'une au moins des positions $u, u+1$ est occupée alors rien ne se passe et $M_{t+1}(v) = M_t(v)$ pour tout v. C'est le cas de droite.



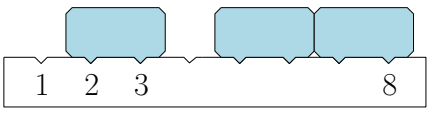
Ainsi le processus continue jusqu'à ce qu'il ne reste sur le substrat que des places vides isolées, on dit que le substrat est *saturé*.

Remarque : *On admet pour l'instant que le processus finit forcément par s'arrêter et atteindre une configuration saturée. Ceci sera démontré à la question T5.*

On pose $M_t = (M_t(1), M_t(2), \dots, M_t(N))$ et on note $M_\infty := (M_\infty(1), M_\infty(2), \dots, M_\infty(N))$ la configuration finale. Les variables aléatoires M_t et M_∞ sont donc à valeurs dans $\{0,1\}^N$. Voici un exemple d'évolution du processus (M_t) dans le cas $N = 8$:

$$\begin{aligned} & & M_0 &= (0,0,0,0,0,0,0,0) \\ U_1 &= 2, & M_1 &= (0,1,1,0,0,0,0,0) \\ U_2 &= 5, & M_2 &= (0,1,1,0,1,1,0,0) \\ U_3 &= 3, & M_3 &= (0,1,1,0,1,1,0,0) \\ U_4 &= 7, & M_4 &= (0,1,1,0,1,1,1,1) \end{aligned}$$

et le processus s'arrête, on a donc $M_\infty = (0,1,1,0,1,1,1,1)$. Schématiquement :



S1. Afficher une trajectoire du processus $(M_t)_{0 \le t \le T}$. On utilisera la représentation suivante: la trajectoire va être codée par une matrice $N \times T$ où le coefficient (k,t) sera la variable $M_t(k)$. Pour afficher une matrice `M`, et donc la trajectoire, une façon simple avec `matplotlib` est d'utiliser `plt.matshow(M)`. On choisira les paramètres suivants : $N = 100, T = 300$.

Densité du substrat

On note $X_t(N) \in \{0, 1, \dots, N\}$ (resp. $X_\infty(N)$) le nombre de positions occupées sur le substrat après l'étape t (resp. à la fin du processus) c'est-à-dire

$$\begin{aligned} X_t(N) &= M_t(1) + M_t(2) + \cdots + M_t(N) \\ X_\infty(N) &= M_\infty(1) + M_\infty(2) + \cdots + M_\infty(N). \end{aligned}$$

Le résultat remarquable obtenu par Flory est que l'espérance de $X_\infty(N)$ croît linéairement et que plus précisément

$$\frac{1}{N}\mathbb{E}[X_\infty(N)] \overset{N\to+\infty}{\to} 1-e^{-2} \approx 0.8646647...$$

Ce résultat théorique semble compatible avec les résultats expérimentaux[3].

T1. Pour le cas $N=4$ déterminer la loi de M_∞, la loi de $X_\infty(4)$ et calculer $\mathbb{E}[X_\infty(4)]$.

S2. Simuler $K=1000$ fois la variable $X_T(N)$ pour $N=100$ et $T=300$. Afficher un histogramme des valeurs obtenues. Calculer également la moyenne de vos K valeurs et comparer avec le résultat de Flory. Avec `matplotlib` un histogramme des valeurs `Donnees` avec 20 bâtons s'affiche avec `plt.hist(Donnees, bins=20, ec='black')` (l'argument `ec='black'` sert à dessiner les bords des bâtons).

On cherche maintenant à valider le résultat de Flory par un calcul numérique exact (et pas par simulation). On pose $e_0=e_1=0$ et pour $N\geq 2$, $e_N=\mathbb{E}[X_\infty(N)]$. On cherche un moyen de calculer numériquement e_N.

T2. Justifier que pour tout $N\geq 2$ on a

$$e_N = 2+\frac{1}{N-1}\sum_{u=1}^{N-2}(e_{u-1}+e_{N-u-1}). \qquad (\star)$$

(Indication : Vous pouvez utiliser l'espérance conditionnelle. Cette question n'est pas forcément évidente à rédiger proprement, des arguments moins rigoureux sont autorisés.)

S3. Écrire un script python qui permet de calculer e_N en utilisant l'équation $(\star)$. Donner la valeur de e_{100}.
(Pour vérifier vos calculs : je trouve $e_{100}=86.195801...$)

Courbe limite du processus de densité X_t

À N fixé le processus $(X_t)_{t\geq 0}$ est croissant au cours du temps et finit par s'arrêter à la valeur X_∞, dont l'espérance est proche de $N(1-e^{-2})$ d'après les questions précédentes. On cherche maintenant à décrire plus précisément le processus $(X_t)_{t\geq 0}$.

On pose

$$\mathbf{x} : t\mapsto 1-\exp\left(-2(1-e^{-t})\right).$$

Il a été démontré[4] le résultat suivant : pour tout réel $T>0$ et tout réel $\varepsilon>0$

$$\mathbb{P}\left(\sup_{0\leq t\leq T}\left|\tfrac{1}{N}X_{\lfloor tN\rfloor}-\mathbf{x}(t)\right|>\varepsilon\right)\overset{N\to+\infty}{\to}0. \qquad (\oplus)$$

S4. Afficher le graphique d'une simulation qui permet d'illustrer la convergence $(\oplus)$. (À vous de choisir N,T.)

[3] C.S.Marvel, C.L.Levesque, *Journal of the American Chemical Society*, vol.60 (1938).

[4] C'est une conséquence de l'éq.(19) dans : P.C.Hemmer. The random parking problem. *Journal of Statistical Physics*, vol.57, n.3, p. 865-869 (1989).

Fin du processus

On cherche à montrer que le processus s'arrête forcément, et donner une estimation du temps nécessaire pour saturer. On note $F_N \in \{0, 1, 2, \dots\}$ la variable aléatoire donnée par le premier instant auquel la configuration est saturée.

T3. Justifier que

$$\{F_N > t\} \subset \bigcup_{u \in \{1,2,\dots,N-1\}} \{ \text{ Pour tout } s \leq t,\ U_s \neq u\}.$$

T4. En déduire que

$$\mathbb{P}(F_N > t) \leq (N-1)\left(1 - \tfrac{1}{N-1}\right)^t. \qquad \text{(\#)}$$

T5. Déduire de l'équation (#) que le processus d'arrête forcément :

$$\mathbb{P}(F_N < +\infty) = 1.$$

Partie B : Modèle uniforme

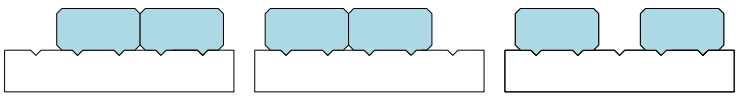
On a vu que le modèle de Flory donne un résultat théorique (densité proche de 86% de densité sur le substrat) compatible avec les expériences chimiques. On cherche maintenant à savoir si un autre modèle assez naturel donne un résultat proche : le modèle uniforme.

On note $\mathcal{S}_N$ l'ensemble des configurations saturées sur un substrat de taille N, et on considère une configuration C uniforme dans $\mathcal{S}_N$. Comme l'était M_∞ précédemment, C est un élément de $\{0,1\}^N$. On écrit $C = (C(1), C(2), \dots, C(N))$.

Remarque : *Le modèle uniforme est donc complètement statique, par opposition au modèle de Flory : il n'y a pas de notion de temps dans le modèle.*

On note $Z_N = C(1) + \cdots + C(N)$ le nombre de positions occupées dans le substrat dans la configuration C. L'objectif est d'étudier numériquement le comportement asymptotique de $\mathbb{E}[Z_N]$.

On pose $s_0 = s_1 = 1$ et pour $N \geq 2$ on pose $s_N = \mathrm{card}(\mathcal{S}_N)$, par exemple $s_5 = 3$:



(Au passage on voit que pour $N = 5$ la variable Z_5 est constante et égale à 4.)

T6. Démontrer que pour tout $N \geq 3$,

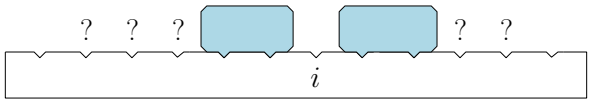
$$s_N = s_{N-2} + s_{N-3}.$$

S5. En déduire une fonction `python` qui calcule s_N.

T7. Pour N fixé et $i \leq N$ la variable $C(i) \in \{0,1\}$ désigne donc l'absence/présence d'un dimère au-dessus de la position i dans une configuration uniforme de taille N. Démontrer que

$$\mathbb{P}(C(i)=0) = \begin{cases} \dfrac{s_{N-3}}{s_N} & \text{si } i=1 \text{ ou } i=N, \\ 0 & \text{si } i=2 \text{ ou } i=N-1, \\ \dfrac{s_{i-3} \times s_{N-i-2}}{s_N} & \text{sinon.} \end{cases}$$

Indication pour le 3ème cas : Voici une représentation schématique de l'événement $\{C(i)=0\}$*:*



S6. En déduire une expression de $\mathbb{E}[Z_N]$ et écrire une fonction `python` qui calcule $\mathbb{E}[Z_N]$. Calculer $\mathbb{E}[Z_{100}]$.

Remarque : Pour le modèle uniforme on observe une densité légèrement inférieure (environ 82%) à celle du modèle de Flory (environ 86%). Il semble que le modèle de Flory colle mieux aux expériences réelles.

PROJET 10

Division cellulaire

sujet proposé par C. Marzouk

cyril.marzouk@polytechnique.edu

On modélise l'évolution de cellules qui grossissent au cours du temps et se divisent aléatoirement en deux. Plus précisément, partant d'une cellule mère de masse m (possiblement aléatoire) à l'instant initial, on considère que sa masse augmente linéairement, à vitesse 1, de façon déterministe, puis, après un temps aléatoire T indépendant de m, de loi exponentielle de paramètre $\lambda > 0$, la cellule, dont la masse est $m + T$, se divise en deux cellules filles, chacune de masse $\frac{1}{2}(m + T)$. Ces deux cellules se comportent ensuite indépendamment et selon la même dynamique que la cellule mère.

On peut représenter l'arbre généalogique associé à cette dynamique, voir la figure 10.1. Dans la suite on s'intéressera à deux quantités, dans un premier temps l'évolution de la masse cellulaire $(X_t)_{t\in\mathbb{R}_+}$ le long d'une branche arbitraire de l'arbre généalogique et dans un second temps l'évolution de la taille de la population $(N_t)_{t\in\mathbb{R}_+}$ au cours du temps.

Remarque: pour les histogrammes, on prendra au moins $N = 1000$ réalisations, mais libre à vous de pousser plus loin selon le temps que cela prend.

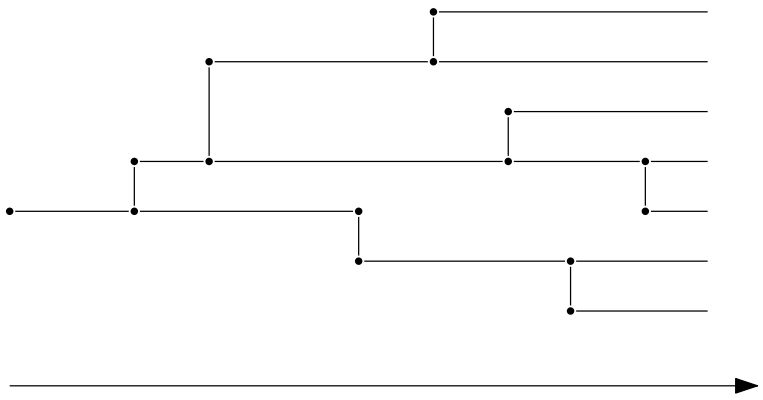


Figure 10.1: Une réalisation de l'arbre généalogique.

1 Évolution de la masse d'une cellule

On note $(X_t)_{t\in\mathbb{R}_+}$ l'évolution de la masse d'une cellule au cours du temps, qui suit la dynamique suivante: partant d'une masse initiale X_0, si $0 = T_0 \leq T_1 \leq T_2 \leq \dots$ sont les instants de division cellulaire, alors pour tout $n \in \mathbb{N}$, pour tout $t \in [T_n, T_{n+1}[$, on a

$$X_t = X_{T_n} + t - T_n \qquad \text{et} \qquad X_{T_{n+1}} = \frac{1}{2}\left(X_{T_n} + T_{n+1} - T_n\right).$$

Un exemple de réalisation possible de $(X_t)_{t\geq 0}$ est donné à la figure 10.2. On suppose que X_0 est une variable aléatoire positive de carré intégrable, et en notant $\Delta T_n = T_n - T_{n-1}$ pour tout $n \geq 1$, alors $(\Delta T_n)_{n\geq 1}$ est une suite de variables aléatoires i.i.d. de loi exponentielle de paramètre $\lambda > 0$ et est indépendante de X_0.

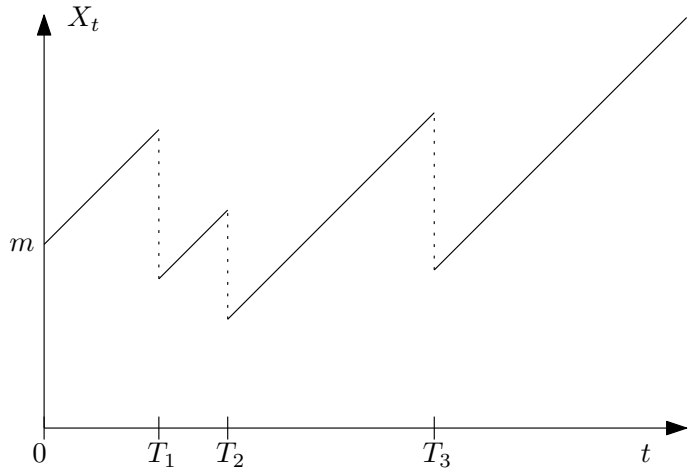


Figure 10.2: Une réalisation du processus $(X_t)_t$ qui correspond à l'évolution de la cellule la plus haute dans l'arbre généalogique de la figure 10.1.

T1. Montrer que $\frac{1}{n}T_n$ converge presque sûrement lorsque $n \to +\infty$ et donner sa limite.

T2. Montrer que $\sqrt{n}(\frac{1}{n}T_n - \frac{1}{\lambda})$ converge en loi lorsque $n \to +\infty$ et donner sa limite.

S1. Représenter un histogramme approché de la loi limite avec $n = 100$ et $\lambda = 1$ et à l'aide de N réalisations.

Pour tout $n \in \mathbb{N}$, on pose $\widetilde{X}_n = X_{T_n}$ la masse cellulaire à l'instant de division.

T3. Montrer que $\widetilde{X}_n$ a la même loi que

$$2^{-n}X_0 + \sum_{k=1}^{n} 2^{-k}\Delta T_k.$$

Montrer ensuite que cette somme converge p.s. vers une variable aléatoire $\widetilde{X}$; en déduire que $\widetilde{X}_n$ converge en loi vers $\widetilde{X}$.

S2. Simuler 5 trajectoires (avec des couleurs différentes) de $(X_t)_{0\leq t\leq 100}$ pour $\lambda = 1$ et $X_0 = 1$.

S3. Donner un histogramme approché de $\widetilde{X}$ pour $\lambda = 1$, $\lambda = 2$ et $\lambda = 10$, avec $X_0 = 1$ dans tous les cas. Quelle vous semble être l'espérance et la variance de $\widetilde{X}$? Commenter.

Montrer que X_t converge en loi lorsque $t \to +\infty$ vers une variable aléatoire limite X est plus compliqué. On se contente ici de montrer que la donnée initiale X_0 n'a que très peu d'importance en comparant les lois de X_t et de Y_t qui suit la même dynamique, mais démarrée de Y_0.

T4. Soit $t \geq 0$; montrer que le nombre de divisions de la cellule survenues avant l'instant t:

$$D_t = \sum_{n=1}^{+\infty} 1_{T_n \leq t},$$

suit la loi de Poisson de paramètre λt. On pourra pour cela commencer par montrer l'identité: pour tout $k \in \mathbb{N}$, pour tout $u \geq 0$,

$$\int_{[0,+\infty[^k} 1_{x_1+\cdots+x_k \leq u}\, dx_1 \cdots dx_k = \frac{u^k}{k!},$$

puis calculer $\mathbb{P}(D_t = k)$.

T5. On suppose que les masses initiales X_0 et Y_0 sont constantes presque sûrement, égales à x et y respectivement, et que les processus $(X_t)_{t\geq 0}$ et $(Y_t)_{t\geq 0}$ suivent la même dynamique, au sens où les divisions cellulaires se produisent aux mêmes instants $(T_n)_{n\geq 1}$ (on dit que les deux processus sont *couplés*). Montrer que pour tout $t \geq 0$, on a

$$\mathbb{E}[|X_t - Y_t|] = |x - y|\, \mathrm{e}^{-\lambda t/2}\,.$$

(On pourra pour cela conditionner par rapport aux événements $\{D_t = k\}$ avec $k \in \mathbb{N}$.)

S4. Sur des graphiques différents, simuler 5 trajectoires de paires de processus couplés $(X_t, Y_t)_{0\leq t\leq 100}$ pour $\lambda = 1$ et différentes valeurs de X_0 et Y_0.

S5. Donner un histogramme approché de la loi de X pour $\lambda = 1$, $\lambda = 2$ et $\lambda = 10$, avec $X_0 = 1$ dans tous les cas. Comparer avec celle de $\widetilde{X}$.

2 Évolution de la taille de la population

On s'intéresse à présent à l'évolution du nombre total $(N_t)_{t\geq 0}$ de cellules présentes à chaque instant t, voir un exemple à la figure 10.3. Cela correspond au nombre total d'extrémités de l'arbre si on le coupe au temps t. On rappelle que les longueurs des arêtes de l'arbre sont des variables aléatoires i.i.d. exponentielles de paramètre λ.

S6. Simuler 5 trajectoires (avec des couleurs différentes) de $(N_t)_{0\leq t\leq 10}$ pour $\lambda = 1$ et $X_0 = 1$. Que pensez-vous être le comportement de N_t lorsque $t \to +\infty$?

T6. Soient n variables aléatoires indépendantes $Y_1, \ldots, Y_n$ telles que Y_i suit une loi exponentielle de paramètre $\lambda_i > 0$ pour tout $i \in \{1, \ldots, n\}$. Donner la loi du minimum $\min_{1\leq i\leq n} Y_i$.

T7. Soient n variables aléatoires indépendantes $Z_1, \ldots, Z_n$ telles que Z_i suit une loi exponentielle de paramètre $i\lambda$ pour tout $i \in \{1, \ldots, n\}$ avec $\lambda > 0$ fixé. Montrer par récurrence que pour tout $z > 0$,

$$\mathbb{P}(Z_1 + \cdots + Z_n \leq z) = \left(1 - \mathrm{e}^{-\lambda z}\right)^n.$$

(On pourra pour cela exprimer la probabilité sous la forme d'une intégrale sur $[0, z]^n$, intégrer par rapport à une des variables et utiliser la formule du binôme de Newton.)

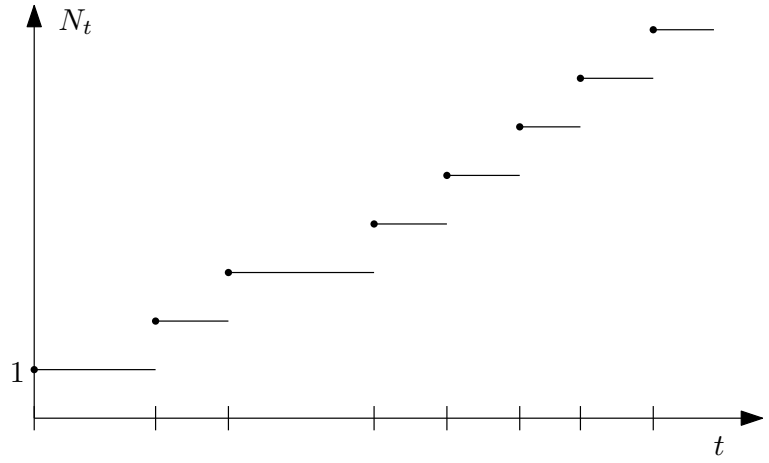


Figure 10.3: Une réalisation du processus $(N_t)_t$ qui correspond à l'arbre généalogique de la figure 10.1.

T8. En déduire que pour tout $t \geq 0$, la variable N_t a la même loi que

$$1 + \sum_{n=1}^{+\infty} 1_{Z_1+\cdots+Z_n \leq t},$$

et que cette loi est la loi géométrique sur $\mathbb{N}^*$ de paramètre $\mathrm{e}^{-\lambda t}$.
(On se souviendra de la propriété d'absence de mémoire des lois exponentielles, équation 5.7 du polycopié.)

T9. Conclure que $\mathrm{e}^{-\lambda t}\, N_t$ converge en loi lorsque $t \to +\infty$ vers une limite que l'on explicitera.

S7. Tracer un histogramme approché de $\mathrm{e}^{-\lambda t}\, N_t$ avec t grand (si possible). Tracer la loi limite sur le même graphique et comparer.

PROJET 11

Permutations aléatoires

sujet proposé par C. Marzouk

cyril.marzouk@polytechnique.edu

Ce projet porte sur l'étude et la simulation de permutations aléatoires d'un ensemble à n éléments avec $n \geq 2$ un entier fixé. On notera σ et τ des permutations, éléments du groupe symétrique $\mathfrak{S}_n$.

On rappelle les fonctions prédéfinies `shuffle()` et `numpy.random.permutation()` qui permettent de tirer des permutations aléatoires uniformes de taille donnée.

Remarque: pour les simulations, on propose au moins $N = 1000$ réalisations, mais libre à vous de pousser plus loin selon le temps que cela prend.

S1. Simuler un échantillon de N permutations de taille $n = 10$ et tracer l'histogramme de $\sigma(1)$, $\sigma(5)$ et $\sigma(10)$.

1 Points fixes

On rappelle que $k \in \{1, \dots, n\}$ est un *point fixe* d'une permutation σ si $\sigma(k) = k$; on note F_n le nombre de points fixes d'une permutation aléatoire uniforme de $\{1, \dots, n\}$ et on définit sa fonction génératrice des moments sur $\mathbb{R}$ par

$$f_n \colon s \mapsto \mathbb{E}[s^{F_n}] = \sum_{j=0}^{n} \mathbb{P}(F_n = j) s^j.$$

S2. À l'aide d'un échantillon de N permutations aléatoires uniformes, tracer un histogramme de F_n pour $n = 100$.

T1. Montrer que $\mathbb{E}[F_n] = 1$; on pourra pour cela écrire $F_n = \sum_{i=1}^{n} 1_{\sigma(i)=i}$.

Plus généralement, pour tout $j \geq 1$, la quantité $F_n(F_n-1)\cdots(F_n-j)$ représente le nombre de $(j+1)$-uplets de points fixes, tous distincts, on a donc

$$F_n(F_n-1)\cdots(F_n-j) = \sum_{\substack{i_1,\ldots,i_{j+1} \\ \text{distincts}}} 1_{\forall k\in\{1,\ldots,j+1\},\sigma(i_k)=i_k}.$$

Bien sûr, cette quantité est nulle si $j \geq n$, et pour $j \in \{0,\ldots,n-1\}$, on montre de même que

$$\mathbb{E}\left[F_n(F_n-1)\cdots(F_n-j)\right] = 1.$$

T2. Exprimer l'espérance précédente avec la fonction f_n. En déduire que pour tout $s \in \mathbb{R}$,

$$f_n(s) = \sum_{j=0}^{n} \frac{(s-1)^j}{j!}.$$

Conclure enfin que F_n converge en loi vers une v.a. de loi de Poisson de paramètre 1 lorsque $n \to +\infty$.

On souhaite à présent générer une permutation aléatoire uniforme parmi celles qui n'ont pas de point fixe. Une méthode de simulation est celle dite *du rejet*: on se fixe un sous-ensemble $A_n \subset \mathfrak{S}_n$ (ici les permutations sans point fixe) et on tire des v.a. i.i.d. uniformes dans $\mathfrak{S}_n$ jusqu'à ce qu'on en obtienne une qui appartienne à A_n; la théorie nous indique que cette dernière v.a. est alors uniformément distribuée dans A_n. De plus, le nombre de tirages effectués suit une loi géométrique de paramètre $|A_n|/|\mathfrak{S}_n|$ où $|\cdot|$ représente le cardinal.

T3. Donner le nombre moyen de tirages dans la méthode du rejet pour obtenir une permutation aléatoire uniforme parmi celles qui n'ont pas de point fixe.

S3. Implémenter cette méthode du rejet; la mettre en œuvre N fois et vérifier le résultat précédent (avec $n = 10$); enfin tracer l'histogramme de $\sigma(1)$, $\sigma(5)$ et $\sigma(10)$.

S4. À nouveau à l'aide d'un échantillon de taille N, tracer l'histogramme de $\sigma(1)$, $\sigma(5)$ et $\sigma(10)$ lorsque σ est une permutation aléatoire uniforme parmi celles qui ont exactement 1 point fixe puis parmi celles qui ont au plus 2 points fixes.

2 Longueur des cycles

On considère un algorithme de génération d'une permutation qui va nous permettre d'étudier le nombre de cycles et leur longueur. On rappelle qu'un *cycle* de longueur ℓ est une suite $(i_1,\ldots,i_\ell)$ telle que $i_{k+1} = \sigma(i_k)$ pour tout $k \in \{1,\ldots,\ell-1\}$ et $\sigma(i_\ell) = i_1$. Chaque permutation peut alors se décomposer d'une unique façon en cycles.

Algorithme 1. On initialise avec

$$m_1 = 1, \qquad r_1 = 1, \qquad I_1 = \{1,\ldots,n\}.$$

Pour k allant de 1 à n, on applique la procédure suivante:

1. On tire $\sigma(m_k)$ uniformément au hasard dans I_k, indépendamment du passé.
2. On pose $I_{k+1} = I_k \setminus \{\sigma(m_k)\}$.
3. Si $\sigma(m_k) \neq r_k$, alors on pose $m_{k+1} = \sigma(m_k)$ et $r_{k+1} = r_k$; sinon, on pose $m_{k+1} = r_{k+1} = \min I_{k+1}$.

En quelques mots, partant de $m_1 = r_1 = 1$, la suite $(m_k)_{k\geq 1}$ suit le cycle $1, \sigma(1), \sigma(\sigma(1)), \ldots$ jusqu'à ce qu'on revienne à 1; on a alors construit le cycle contenant 1 et on construit ensuite de la même façon le cycle contenant le plus petit élément n'appartenant pas à ce premier cycle, etc.

T4. Montrer que l'algorithme produit une permutation aléatoire uniforme: pour tout $\tau \in \mathfrak{S}_n$, on a $\mathbb{P}(\sigma = \tau) = 1/n!$.
(On pourra pour cela conditionner successivement par rapport aux valeurs de $\sigma(m_k)$.)

S5. Implémenter l'algorithme; simuler un échantillon de N permutations et tracer l'histogramme de $\sigma(1)$, $\sigma(5)$ et $\sigma(10)$.

On note U_n la longueur du cycle contenant 1 et C_n le nombre de cycles de σ.

T5. Montrer que U_n suit la loi uniforme sur $\{1, \dots, n\}$.

T6. Grâce à l'algorithme 1, montrer que C_n a la même loi que la somme de n v.a. de Bernoulli indépendantes, de paramètre respectif $1, \frac{1}{2}, \dots, \frac{1}{n}$.

T7. Montrer que $\mathbb{E}[C_n] \sim \ln n$ et $\mathrm{Var}(C_n) \sim \ln n$ lorsque $n \to +\infty$.

T8. À l'aide des fonctions caractéristiques, montrer que la suite C_n correctement renormalisée converge en loi lorsque $n \to +\infty$ et donner sa limite.

S6. Simuler un échantillon de N permutations et tracer l'histogramme de U_n et de C_n. Faites apparaître la loi limite de C_n sur la même figure que l'histogramme.

3 Cycles de longueur paire

On souhaite à présent générer une permutation σ aléatoire uniforme de l'ensemble $\{1, \dots, 2n\}$ dont tous les cycles ont une longueur paire. On modifie pour cela légèrement l'algorithme précédent.

Algorithme 2. On initialise avec

$$m_1 = 1, \qquad r_1 = 1, \qquad I_1 = \{1, \dots, 2n\}.$$

Pour k allant de 1 à n, on applique la procédure suivante:

1. Indépendamment du passé, on tire $\sigma(m_k)$ uniformément au hasard dans $I_k \setminus \{r_k\}$, puis on tire $\sigma(\sigma(m_k))$ uniformément au hasard dans $I_k \setminus \{\sigma(m_k)\}$.
2. On pose $I_{k+1} = I_k \setminus \{\sigma(m_k), \sigma(\sigma(m_k))\}$.
3. Si $\sigma(\sigma(m_k)) \neq r_k$, alors on pose $m_{k+1} = \sigma(\sigma(m_k))$ et $r_{k+1} = r_k$; sinon, on pose $m_{k+1} = r_{k+1} = \min I_{k+1}$.

T9. Comparer cet algorithme au précédent et montrer que si τ est une permutation de $\mathfrak{S}_{2n}$ à cycles de longueur paire, alors

$$\mathbb{P}(\sigma = \tau) = \prod_{k=1}^{n} \frac{1}{(2k-1)^2}.$$

T10. En déduire que:

(a) La permutation σ suit la loi uniforme parmi celles à cycles de longueur paire.

(b) Si l'on tire une permutation aléatoire uniforme de $\mathfrak{S}_{2n}$, la probabilité que tous ses cycles aient longueur paire est équivalente à $1/\sqrt{\pi n}$ lorsque $n \to +\infty$.

Ce dernier équivalent justifie a posteriori le fait de ne pas utiliser la méthode du rejet: l'ensemble que l'on vise est beaucoup trop petit et il faudrait de nombreux essais (en moyenne $\sqrt{\pi n}$) pour obtenir une permutation de la loi souhaitée.

On note U'_n la longueur du cycle contenant 1 et C'_n le nombre de cycle de σ.

S7. Simuler un échantillon de $N = 1000$ permutations avec l'algorithme 2 et tracer l'histogramme de U'_n et de C'_n.

- Comparer la loi de $U'_n/(2n)$ à celle d'une v.a. de densité $x \mapsto \frac{1}{2\sqrt{1-x}}$ sur $[0,1]$.
- Les quantités $\mathbb{E}[C'_n]/\ln n$ et $\mathrm{Var}(C'_n)/\ln n$ vous semblent-elles converger? La variable C'_n correctement normalisée semble-t-elle converger en loi?

(On pourra si besoin prendre n plus grand que 10 pour cette question.)

PROJET 12

Ruine du casino

sujet proposé par C. Marzouk

cyril.marzouk@polytechnique.edu

L'étude de ruine au casino concerne souvent celle d'un joueur, pariant jusqu'à se retrouver sans le sou. Ici on se concentre au contraire sur le gérant de l'établissement qui se demande si son affaire est bonne.

Le modèle est le suivant: on se donne une suite de variables aléatoires $(\xi_i)_{i\geq 1}$ i.i.d. de loi exponentielle de paramètre 1 et on pose pour tout $n \geq 1$ et tout $t \geq 0$,

$$T_n = \xi_1 + \cdots + \xi_n \qquad \text{et} \qquad N_t = \sum_{n=1}^{+\infty} 1_{T_n \leq t}.$$

Par ailleurs, indépendamment, on se donne une suite de variables aléatoires i.i.d. $(X_i)_{i\geq 1}$ positives, d'espérance $\mu = \mathbb{E}[X_1]$ que l'on suppose finie et non nulle.

Les instants T_n représentent ceux auxquels un joueur gagne et alors X_n représente son gain; le nombre N_t est le nombre total de gagnants au temps t. Par ailleurs, on suppose que les rentrées d'argent du casino (la somme des dépenses des différents clients) sont déterministes et à taux constant α. Ainsi, si le gérant démarre avec une somme $y > 0$, alors au temps $t > 0$, sa fortune vaut

$$Y_t = y + \alpha t - \sum_{n=1}^{N_t} X_n,$$

où une somme vide vaut 0. La probabilité de ruine du gérant est donc

$$r(y) = \mathbb{P}(\exists t > 0 : Y_t \leq 0 \mid Y_0 = y).$$

1 La ruine est-elle évitable?

S1. Simuler 5 trajectoires (avec des couleurs différentes) de $(Y_t)_{0\leq t\leq 100}$ pour $y = 1,\ 5,\ 10,\ 20,\ 50$ avec $\alpha = 1$ et X_1 qui suit la loi exponentielle de paramètre 1. Faire aussi figurer en pointillés la ligne horizontale $y = 0$.

T1. Montrer que pour tout $y > 0$, on a $r(y) > 0$. Ainsi, la ruine ne peut jamais être complètement exclue!

2 Majoration de la probabilité de ruine

S2. On fixe $y = 10$, $\alpha = 1$ et on prend X_1 une loi exponentielle de paramètre $1/\mu$. Dans les trois cas $\mu = 3/4$, $\mu = 1$ et $\mu = 4/3$, simuler un échantillon de taille $N = 1000$ de $(Y_t)_{t\leq 100}$. Quelle est la proportion de casinos ruinés à l'instant $t = 100$ (ou avant)? Qu'attendez-vous du comportement de $(Y_t)_t$ dans les trois cas?

Pour étudier la fonction r, on introduit la marche aléatoire $(S_n)_{n\in\mathbb{N}}$ définie par

$$S_n = \sum_{k=1}^{n}(X_k - \alpha\xi_k).$$

Notons que les accroissements $(X_k - \alpha\xi_k)_{n\geq 1}$ sont i.i.d.

T2. Montrer que

$$r(y) = \mathbb{P}\Big(\max_{n\geq 1} S_n \geq y\Big) = \lim_{n\to+\infty} \mathbb{P}\Big(\max_{1\leq j\leq n} S_j \geq y\Big).$$

Dans la suite, on notera $r_n(y) = \mathbb{P}(\max_{1\leq j\leq n} S_j \geq y)$.

T3. Montrer que S_n/n converge presque sûrement vers une constante. En déduire que si $\alpha < \mu$, alors $r(y) = 1$ pour tout $y > 0$.

On suppose désormais que $\alpha > \mu$ et on cherche à majorer $r(y)$. Pour cela on fait l'hypothèse suivante:

$$\exists A > 0 \text{ tel que} \qquad \mathbb{E}[\mathrm{e}^{AX_1}] < +\infty \qquad \text{et} \qquad \mathbb{E}[\mathrm{e}^{A(X_1-\alpha\xi_1)}] = 1.$$

T4. En étudiant la fonction $g\colon t \mapsto \mathbb{E}[\mathrm{e}^{t(X_1-\alpha T_1)}]$, montrer que si un tel A existe, il est unique.

T5. Montrer que $r_1(y) \leq \mathbb{E}[\mathrm{e}^{A(S_1-y)}\, 1_{S_1\geq y}]$ pour tout $y > 0$. Par récurrence, montrer que $r_n(y) \leq \mathrm{e}^{-Ay}$ pour tout $y > 0$ et pour tout $n \geq 1$.
Pour simplifier le cadre, on pourra supposer que les v.a. $X_k - \alpha\xi_k$ ont une densité (ce qui est le cas si les X_k ont une densité) et regarder séparément selon que $S_1 \geq y$ ou non.

S3. Dans le même cadre que la question S2, tracer un histogramme tronqué du temps de ruine dans les trois cas.

3 Simulations de la fonction de ruine

On se propose de regarder numériquement une borne de la forme $r(y) \leq \mathrm{e}^{-Ay}$ semble optimale. Pour cela, à l'aide d'échantillons de taille $N = 100$ ou plus, on tracera la courbe empirique associée $\{r(y); 0 < y \leq 50\}$ (en prenant des valeurs de y discrètes, par exemple $\{k/2; 1 \leq k \leq 100\}$) dans les trois cas suivants.

S4. On fixe $\alpha = 1$ et on suppose que X_1 suit la loi exponentielle de paramètre $1/\mu$, avec pour $\mu \in \{0.5, 0.8, 1, 1.2, 1.5\}$.

S5. On considère à présent une loi pour laquelle A n'existe pas; précisément, on choisit X_1 tel que

$$\mathbb{P}(X_1 = n) = \frac{4}{n(n+1)(n+2)}, \qquad n \in \mathbb{N}^*,$$

dont la moyenne est $\mu = 2$ et on prend $\alpha = 5$.

S6. On considère à présent une loi pour X_1 a une moyenne infinie:

$$\mathbb{P}(X_1 = n) = \frac{1}{n(n+1)}, \qquad n \in \mathbb{N}^*,$$

et on prend $\alpha = 25$.

4 Le cas de gains exponentiels

On suppose que les X_i suivent la loi exponentielle de paramètre $\frac{1}{\mu} > \frac{1}{\alpha}$. Le but est de calculer explicitement la fonction de ruine:

$$r(y) = \frac{\mu}{\alpha}\,\mathrm{e}^{-Ay} \qquad \text{pour tout } y > 0, \tag{$\star$}$$

avec un A explicite.

T6. Donner la valeur de $A > 0$ tel que $\mathbb{E}[\mathrm{e}^{A(X_1 - \alpha\xi_1)}] = 1$.

En raisonnant comme à la question T5 et en passant à la limite, on obtient que la fonction $r = \lim_n r_n$ vérifie: pour tout $y > 0$,

$$r(y) = \mathbb{E}[1_{S_1 \geq y} + r(y - S_1)1_{S_1 < y}].$$

On rappelle que $S_1 = X_1 - \alpha\xi_1$ avec ξ_1 indépendante de X_1 et de loi exponentielle de paramètre 1.

T7. À l'aide de la formule précédente, montrer que la fonction $R\colon u \mapsto \int_0^u \mathrm{e}^{\frac{x}{\mu}}\, r(x)dx$, vérifie l'équation différentielle

$$R''(y) = \left(\frac{1}{\mu} + \frac{1}{\alpha}\right) R'(y) - \frac{1}{\alpha\mu} R(y) - \frac{1}{\alpha}.$$

T8. Résoudre l'équation précédente et en déduire la formule ($\star$).

S7. Reprendre l'approximation de la courbe $\{r(y); 0 < y \leq 50\}$ pour $\alpha = 1$ et $\mu = 0.8$ de la question S4 et, sur le même graphique, tracer la courbe théorique donnée par ($\star$).

5 Épilogue

Pourquoi n'observe-t-on jamais de ruine de casino?

PROJET 13

Croissance par aggrégation

sujet proposé par C. Marzouk

cyril.marzouk@polytechnique.edu

On étudie un modèle de croissance dans lequel des blocs se déposent au cours du temps sur une surface; de telles croissances *par aggrégation* se retrouvent dans beaucoup de modélisations de phénomènes physiques, chimiques ou biologiques. On se placera ici dans un espace discret et à une dimension, plus précisément sur le *cycle* de longueur $L \geq 5$ donné par $\mathbb{Z}/L\mathbb{Z}$; pour tout $i \in \mathbb{Z}/L\mathbb{Z}$, on appellera $i-1 \bmod L$ et $i+1 \bmod L$ les *voisins* du site i. Le but de ce projet est d'étudier la hauteur maximale H_n à l'instant n et son comportement asymptotique.

Remarque: on prendra $L = 50$ pour les simulations et on propose au moins $N = 500$ réalisations, mais libre à vous de pousser plus loin selon le temps que cela prend.

1 Un modèle jouet

Le modèle le plus simple pour commencer consiste à considérer que les blocs arrivent un par un, avec une position uniformément distribuée sur $\mathbb{Z}/L\mathbb{Z}$ et indépendamment. Chaque bloc occupe le site sur lequel il se dépose, et ainsi chaque site supporte une pile de blocs, voir un exemple à la figure 13.1. On note $(Y_{n,0}, \ldots, Y_{n,L-1})$ le vecteur des nombres de blocs sur chaque site et $H_n = \max\{Y_{n,0}, \ldots, Y_{n,L-1}\}$ la hauteur maximale lorsque n blocs sont tombés.

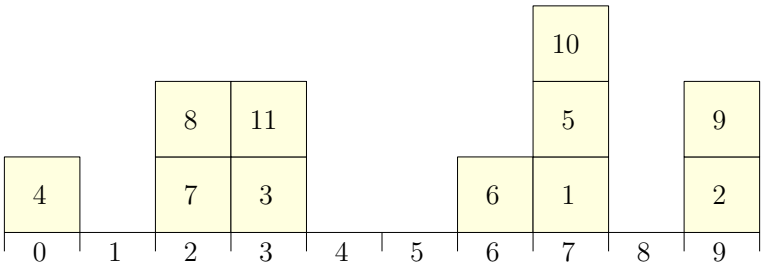


Figure 13.1: Une évolution possible du modèle avec $L = 10$ jusqu'à l'instant $n = 11$. Les blocs sont représentés par des carrés et sont numérotés dans l'ordre d'arrivée. La hauteur maximale est $H_{11} = 3$.

S1. Représenter graphiquement une réalisation, avec $L = 50$ et $n = 5000$. (Par exemple via la fonction `matshow()` de matplotlib qui permet de représenter une matrice numpy graphiquement.)

S2. Représenter graphiquement la moyenne d'un échantillon de taille N avec $n = 100$. Représenter également l'histogramme associé de H_n.

S3. À l'aide d'échantillons de taille N, représenter l'évolution moyenne des deux suites $(H_n)_{1\leq n\leq 100}$ et $(H_n/n)_{1\leq n\leq 100}$.

T1. Donner la loi de $Y_{n,i}$ pour tout $i \in \{0, \dots, L-1\}$ puis montrer que H_n/n converge presque sûrement vers une constante à déterminer.

2 Un modèle à trous

On suppose à présent que les blocs arrivent toujours un à un indépendamment et uniformément au hasard sur $\mathbb{Z}/L\mathbb{Z}$, mais désormais on imagine qu'un bloc occupe tout l'espace (strictement) entre les deux voisins du site sur lequel il se dépose, de sorte qu'un bloc qui arrive au site $i \in \mathbb{Z}/L\mathbb{Z}$ peut se déposer par dessus un bloc qui se trouve positionné en $i-1$ ou en $i+1$ et qui déborde sur i, voir un exemple à la figure 13.2.

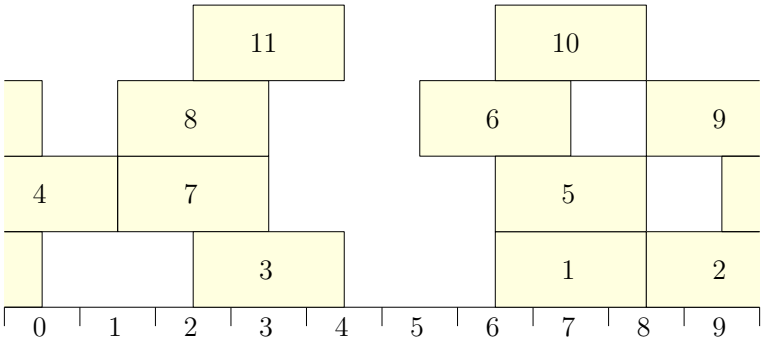


Figure 13.2: Une évolution possible du modèle avec $L = 10$ jusqu'à l'instant $n = 11$. La valeur du vecteur $(Y_{n,0}, \dots, Y_{n,L-1})$ est $(2, 0, 3, 4, 0, 0, 3, 4, 0, 3)$.

On note encore $Y_{n,i}$ la hauteur du plus haut bloc qui s'est déposé au site i à l'instant n (un bloc au-dessus déposé en $i \pm 1$ ne compte pas!). Les $Y_{n,i}$ vérifient alors: $Y_{0,i} = 0$ pour tout $i \in \mathbb{Z}/L\mathbb{Z}$ et pour tout $n \in \mathbb{N}$, si le bloc $n+1$ se dépose en i, alors

$$Y_{n+1,i} = 1 + \max\{Y_{n,i-1}, Y_{n,i}, Y_{n,i+1}\}.$$

On note encore $H_n = \max\{Y_{n,0}, \dots, Y_{n,L-1}\}$; on cherche à étudier son comportement asymptotique. On notera également $(X_k)_{k\geq 1}$ les positions des blocs, qui sont donc des variables aléatoires i.i.d. uniformes sur $\mathbb{Z}/L\mathbb{Z}$.

S4. Reprendre les questions S1, S2 et S3 avec ce modèle.

On a le résultat suivant: il existe une constante $h \in \mathbb{R}$ telle que

$$\lim_{n\to+\infty} \frac{H_n}{n} = h \qquad \text{presque sûrement.}$$

Dans les questions suivantes, on montrera que $\limsup_n H_n/n = h$ p.s. puis on s'attachera à borner h.

S5. À l'aide des simulations précédentes, estimer la valeur de h: donner un intervalle auquel appartiennent 95% des valeurs simulées de H_n/n.

3 Existence d'une limite

Pour tout $m \in \mathbb{N}^*$, la variable H_m est une fonction déterministe des $X_1, \dots, X_m$; pour $n, m \in \mathbb{N}^*$, on note $H(n, n+m]$ la même fonction, mais appliquée en les $X_{n+1}, \dots, X_{n+m}$. Autrement dit, $H(n, n+m]$ représente la hauteur maximale à l'instant $n+m$ si, entre les instants n et $n+1$, on enlève tous les blocs déjà arrivés pour repartir de 0. Notons que $H_m = H(0, m]$.

T2. Montrer que pour tous $n, m \in \mathbb{N}^*$, presque sûrement,

$$H(0, n+m] \leq H(0, n] + H(n, n+m].$$

En particulier, la suite donnée par $h_n = \mathbb{E}[H_n]$ est *sous-additive* et le lemme de Fekete implique que

$$h := \inf_{n \to +\infty} \frac{h_n}{n} = \lim_{n \to +\infty} \frac{h_n}{n}.$$

T3. Montrer que si n, k, q, r sont des entiers naturels tels que $n = kq + r$, alors

$$H_n \leq \sum_{i=1}^{q} H_k^{(i)} + r,$$

où les variables $(H_k^{(i)})_{1 \leq i \leq q}$ sont i.i.d. de même loi que H_k. En utilisant la loi des grands nombres sur ces v.a. avec k fixé et $q \to +\infty$, en déduire que $\limsup_n H_n/n \leq h_k/k$ p.s. et ainsi que

$$\limsup_{n \to +\infty} \frac{H_n}{n} \leq h \qquad \text{presque sûrement.}$$

Conclure finalement à l'aide du lemme de Fatou que

$$\limsup_{n \to +\infty} \frac{H_n}{n} = h \qquad \text{presque sûrement.}$$

On admettra dans la suite que la limite inférieure est égale à h aussi, de sorte que

$$\lim_{n \to +\infty} \frac{H_n}{n} = h \qquad \text{presque sûrement.}$$

4 Minoration

Modifions la dynamique de la façon suivante: à chaque fois qu'un bloc ne fait pas augmenter H_n, on l'efface immédiatement. Ainsi, si le dernier bloc conservé est en position $i \in \mathbb{Z}/L\mathbb{Z}$, par la suite, on ne garde un bloc que s'il tombe en position $i-1$, i ou $i+1$, et on l'efface sinon. Pour $n, m \geq 1$, on note G_n la hauteur de la pile après que n blocs sont tombés, et $\tau_m = \inf\{n \geq 1 : G_n = m\}$ le temps nécessaire pour que la pile atteigne la hauteur m. En particulier, $\tau_1 = 1$ presque sûrement.

T4. Montrer que les v.a. $(\tau_{i+1} - \tau_i)_{i \geq 1}$ sont i.i.d. de loi géométrique de paramètre $3/L$ sur $\mathbb{N}^*$. En déduire que τ_m/m converge presque sûrement $m \to +\infty$ et exprimer sa limite.

T5. En déduire que $h \geq 3/L$.

S6. Un argument (intéressant, mais un peu long) permet de montrer que $h \leq 1/a_0$ où a_0 est la plus petite solution de

$$\frac{L - a_0}{L} + \ln\left(\frac{3a_0}{L}\right) = 0.$$

Donner finalement une valeur approchée de $1/a_0$.

5 Conclusion et généralisations

S7. L'intervalle pour h obtenu par simulation dans la question S5 est-il inclus dans l'intervalle théorique?

S8. Reprendre les questions S1, S2, S3 et S5 avec le modèle de la section 2 mais sur l'intervalle plutôt que le cycle, i.e. lorsque 0 et $L - 1$ ne sont pas voisins, comme dans la figure 13.3.

S9. Reprendre les questions S1, S2, S3 et S5 avec un modèle (sur le cycle) dans lequel les blocs sont de longueurs variables et débordent sur les k voisins ($k = 0$ est donc le modèle jouet et $k = 1$ le modèle de la section 2) avec $k \in \{0, 1, 2, 3\}$ choisi uniformément au hasard.

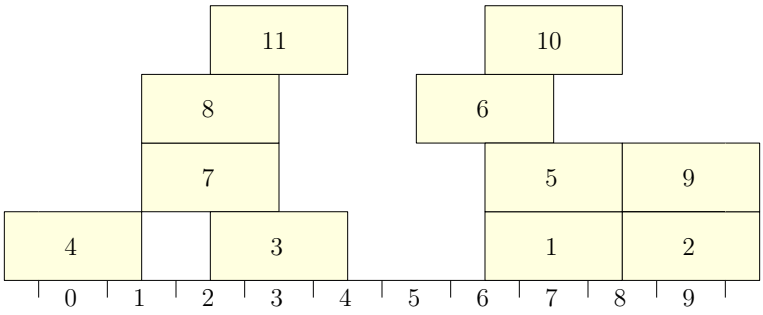


Figure 13.3: Une évolution possible du modèle sur l'intervalle jusqu'à l'instant $n = 11$.

PROJET 14

Paradoxe de Parrondo

sujet proposé par C. Marzouk

cyril.marzouk@polytechnique.edu

Les jeux de Parrondo sont des jeux de lancers de pièce: un “pile” fait gagner 1 et un “face” fait perdre 1. On fixe un $\varepsilon \in]0, \frac{1}{10}[$ et on pose

$$p = \frac{1}{2} - \varepsilon, \qquad p_0 = \frac{1}{10} - \varepsilon, \qquad p_1 = \frac{3}{4} - \varepsilon.$$

En pratique, on prendra ε petit (par exemple $1/1000$). Il y a alors deux jeux:

- le jeu A consiste à lancer une pièce qui donne pile avec probabilité p;
- dans le jeu B, on lance une pièce qui donne pile avec probabilité p_0 si le gain (positif ou négatif) cumulé jusqu'à présent est divisible par 3, et on lance une pièce qui donne pile avec probabilité p_1 sinon.

On note $S_0 = 0$ le capital initial et S_n le gain après n lancers.

Le comportement si l'on joue de manière répétée au jeu A est assez clair, celui si l'on joue de manière répétée au jeu B le deviendra au fil des questions, et nous verrons que jouer alternativement à plusieurs parties de A, puis plusieurs de B et ainsi de suite, ou bien jouer à chaque étape soit à A soit à B aléatoirement peut révéler des comportements surprenants.

L'étude introduit un peu à la théorie des chaînes de Markov qui fait l'objet du cours MAP 432 en deuxième année.

Nota bene: Les calculs numériques fastidieux pourront être délégués à un ordinateur.

1 Étude des deux jeux

T1. On joue au jeu A de façon répétée; comment se comporte S_n lorsque $n \to \infty$ (loi des grands nombres, théorème central limite)? Avez-vous envie de jouer à ce jeu?

S1. Illustrer ce comportement via des simulations en superposant plusieurs trajectoires $(S_k)_{0 \leq k \leq n}$ ainsi qu'un histogramme de la loi de S_n pour un n grand.

S2. On joue à présent au jeu B de façon répétée; à l'aide de simulations, conjecturer le comportement de S_n lorsque $n \to \infty$; avez-vous envie de jouer à ce jeu?

On se propose d'étudier théoriquement le comportement en temps long du jeu B; il semble naturel de considérer la suite $X = (X_n)_{n\geq 0}$ à valeurs dans $\{0,1,2\}$ donnée par $X_n \equiv S_n \bmod 3$. La suite X est un exemple de *chaîne de Markov* homogène, sa loi est entièrement caractérisée par sa *matrice de transition* P_B, où $P_B(i,j) = P(X_{n+1} = j \mid X_n = i)$ pour toute paire $(i,j) \in \{0,1,2\}^2$.

Ainsi, on a ici

$$P_B = \begin{pmatrix} 0 & p_0 & 1-p_0 \\ 1-p_1 & 0 & p_1 \\ p_1 & 1-p_1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{10}-\varepsilon & \frac{9}{10}+\varepsilon \\ \frac{1}{4}+\varepsilon & 0 & \frac{3}{4}-\varepsilon \\ \frac{3}{4}-\varepsilon & \frac{1}{4}+\varepsilon & 0 \end{pmatrix}.$$

T2. Montrer que pour toute paire $(i,j) \in \{0,1,2\}^2$, il existe un entier k tel que $P(X_{n+k} = j \mid X_n = i) \neq 0$; on dit que la chaîne X est *irréductible*.

Un résultat général montre qu'une chaîne de Markov irréductible à valeurs dans un ensemble fini E, de matrice de transition P, possède une unique *loi invariante*, c'est-à-dire une loi π sur E telle que pour tout $j \in E$,

$$\sum_{i\in E} \pi(i)P(i,j) = \pi(j).$$

Le terme de gauche correspond simplement au produit matriciel $(\pi P)(j)$ si l'on voit π comme un vecteur ligne; il s'agit de la loi de X_{n+1} lorsque X_n est choisi aléatoirement selon π (c'est la formule des probabilités totales); la relation dit qu'alors la loi de X_{n+1} est encore π, d'où le nom de loi invariante (dans le temps).

T3. Montrer que la loi invariante π_B associée à la suite X est donnée par

$$\begin{aligned} \pi_B(0) &= C\,(1-p_1(1-p_1)) \\ \pi_B(1) &= C\,(1-p_1(1-p_0)) \\ \pi_B(2) &= C\,(1-p_0(1-p_1)) \end{aligned}$$

avec $C > 0$ telle que $\pi_B(0) + \pi_B(1) + \pi_B(2) = 1$.

T4. Exprimer en fonction de π_B et des paramètres la probabilité de gagner au prochain lancer, ainsi que le gain moyen γ_B, si X_n suit la loi π_B.

Bien que S_n ne soit pas une somme de variables aléatoires i.d.d. une version généralisée de la loi des grands nombres que l'on admettra permet de conclure que S_n/n converge presque sûrement vers une constante c qui ne dépend pas de X_0. Par convergence dominée (comme $|S_n/n| \leq 1$), la convergence a également lieu dans L^1 et ainsi cette constante c est le gain moyen γ_B de la question précédente.

S3. Représenter γ_B en fonction de $\varepsilon > 0$. Commenter par rapport aux simulations précédentes.

2 Mélange de jeux

Fixons deux entiers $r, s \geq 1$; on définit un nouveau jeu, que l'on note A^rB^s, qui consiste en r parties successives selon la règle A, suivies de s parties successives selon la règle B. On note toujours S_n le gain après n parties (donc $n(r+s)$ lancers !).

S4. À l'aide de simulations, conjecturer le comportement de S_n lorsque $n \to \infty$ pour $r = s = 1$, pour $r = s = 2$, ainsi que deux autres couples (r, s) de votre choix; commenter.

On se propose d'étudier théoriquement ces jeux en utilisant à nouveau la suite X dont la matrice de transition est désormais $P_{A^r B^s} = P_A^r P_B^s$ où P_B est la matrice utilisée dans la première partie.

T5. On fixe $\varepsilon = 0$ et on choisit $r = s = 1$; montrer que la matrice de transition de X est

$$P_{AB} = \frac{1}{40} \begin{pmatrix} 20 & 5 & 15 \\ 15 & 7 & 18 \\ 5 & 2 & 33 \end{pmatrix},$$

dont l'unique loi invariante est donnée par

$$\pi_{AB}(0) = \frac{3}{13}, \qquad \pi_{AB}(1) = \frac{1}{13}, \qquad \pi_{AB}(2) = \frac{9}{13}.$$

Quel est le gain moyen sous la loi invariante γ_{AB}?

Si l'on prend $r = s = 2$, toujours avec $\varepsilon = 0$, on trouve comme matrice de transition

$$P_{A^2B^2} = \frac{1}{320} \begin{pmatrix} 162 & 59 & 99 \\ 151 & 58 & 111 \\ 111 & 47 & 162 \end{pmatrix},$$

dont l'unique loi invariante est donnée par

$$\pi_{A^2B^2}(0) = \frac{2783}{6357}, \qquad \pi_{A^2B^2}(1) = \frac{1075}{6357}, \qquad \pi_{A^2B^2}(2) = \frac{2499}{6357}.$$

Des calculs plus pénibles (demandez-vous comment calculer la probabilité qu'après ces quatre lancers, le gain soit $+2$ par exemple) montrent que le gain moyen sous la loi invariante est de

$$\gamma_{A^2B^2} = \frac{16}{163}.$$

T6. Que pouvez-vous conclure, dans les deux cas $r = s = 1$ et $r = s = 2$, quant au comportement asymptotique de S_n lorsque $\varepsilon > 0$ est petit? Commenter par rapport aux simulations.

3 Mélange aléatoire

On propose une dernière variation: le jeu C qui consiste à jouer aléatoirement soit selon la règle A, soit selon la règle B; précisément, on lance une première pièce non biaisée, si elle donne "pile" alors on joue un lancer au jeu A, sinon on joue un lancer au jeu B; on répète à chaque étape de sorte que le choix de la règle est une suite i.d.d. de Bernoulli$(1/2)$; on note toujours S_n le gain après n parties.

S5. À l'aide de simulations, conjecturer le comportement de S_n lorsque $n \to \infty$ et commenter.

T7. Une dernière fois, pour $\varepsilon = 0$, montrer que la matrice de transition de X est

$$P_C = \frac{1}{40} \begin{pmatrix} 0 & 12 & 28 \\ 15 & 0 & 25 \\ 25 & 15 & 0 \end{pmatrix},$$

dont l'unique loi invariante est donnée par

$$\pi_C(0) = \frac{245}{709}, \qquad \pi_C(1) = \frac{180}{709}, \qquad \pi_C(2) = \frac{284}{709},$$

de sorte que le gain moyen sous la loi invariante est

$$\gamma_C = \frac{18}{709}.$$

T8. Que pouvez-vous conclure quant au comportement asymptotique de S_n lorsque $\varepsilon > 0$ est petit?

S6. Simuler et commenter à nouveau le comportement asymptotique de S_n lorsque le choix du jeu A ou du jeu B à chaque étape est plus généralement une variable de Bernoulli(α) pour différentes valeurs de $\alpha \in]0, 1[$.

PROJET 15

Percolation et probabilité critique

sujet proposé par L. Massoulié

laurent.massoulie@inria.fr

On considère la grille $\mathbb{Z}^2$ comme un graphe de sommets $x = (x_1, x_2) \in \mathbb{Z}^2$ et dont les arcs sont donnés par les paires (x, y) telles que $|x-y| := |x_1-y_1|+|x_2-y_2|$ est égal à 1. A chaque arc $e = (x, y)$, on associe une variable aléatoire de Bernoulli de paramètre $p \in [0, 1]$, ξ_e. Les variables aléatoires ξ_e sont supposées i.i.d.. Elles nous permettent de définir un sous-graphe de $\mathbb{Z}^2$, obtenu en conservant tous les sommets $x \in \mathbb{Z}^2$, mais en conservant uniquement les arcs e tels que $\xi_e = 1$.

On note C la composante connexe contenant l'origine 0 de $\mathbb{Z}^2$ dans ce sous-graphe aléatoire.

1 Partie théorique

T1. On note $|C|$ le nombre de sommets dans la composante connexe C, et $f(p) = \mathbb{P}(|C| = +\infty)$. Montrer que la fonction f est croissante en $p \in [0, 1]$. Indice: on se donnera des variables aléatoires U_e i.i.d. uniformes sur $[0, 1]$. On considérera, pour chaque valeur de $p \in [0, 1]$, les variables aléatoires de Bernoulli $\xi_e(p)$ définies selon $\xi_e(p) = \mathbf{1}_{U_e \leq p}$.

T2. On appelle chemin auto-évitant de longueur n issu de l'origine dans la grille $\mathbb{Z}^2$ toute suite de points $i_0, i_1, \ldots, i_n$ de $\mathbb{Z}^2$ telle que $i_0 = 0$, $|i_k - i_{k-1}| = 1$ pour tout $k \in [n]$, et enfin telle que les i_k, $\{k = 0, \ldots, n\}$ sont tous distincts. On note σ_n le nombre de tels chemins.

Etablir pour tout $n \geq 1$ l'inégalité

$$f(p) \leq \sigma_n p^n.$$

On définit la probabilité critique de percolation

$$p_c := \inf\{p \in [0, 1] : f(p) > 0\}.$$

Proposer un majorant de σ_n et en déduire que pour $p > 0$ suffisamment petit (par exemple, $p < 1/3$), $f(p) = 0$, et donc $p_c \geq 1/3$.

T3. On considère la grille duale $\tilde{G}$ de $\mathbb{Z}^2$, constituée des sommets $\tilde{x} = x + (1/2, 1/2)$, $x \in \mathbb{Z}^2$, et des arcs $\tilde{e}$ de paires $(\tilde{x}, \tilde{y})$ de sommets telles que $|\tilde{x} - \tilde{y}| = |\tilde{x}_1 - \tilde{y}_1| + |\tilde{x}_2 - \tilde{y}_2| = 1$. Notant $e_1 = (1, 0)$ et $e_2 = (0, 1)$, à un arc $\tilde{e} = (\tilde{x}, \tilde{x} + e_1)$ de $\tilde{G}$, on associe la variable aléatoire $\xi_{\tilde{e}} := \xi_{(\tilde{x}+(1/2,-1/2),\tilde{x}+(1/2,1/2))}$. de même à un arc $\tilde{e} = (\tilde{x}, \tilde{x} + e_2)$ on associe la variable aléatoire $\xi_{\tilde{e}} := \xi_{(\tilde{x}+(-1/2,1/2),\tilde{x}+(1/2,1/2))}$.

Ces variables aléatoires $\xi_{\tilde{e}}$ définissent un sous-graphe aléatoire de $\tilde{G}$. On appelle circuit de longueur n dans $\tilde{G}$ une suite $\tilde{i}_0, \ldots, \tilde{i}_n$ telle que pour tout $k \in [n]$, $(\tilde{i}_{k-1}, \tilde{i}_k)$ est un arc de $\tilde{G}$, $i_0 = i_n$, et les i_k, $k \in [n]$ sont tous distincts.

Montrer que chaque circuit de longueur n de $\tilde{G}$ qui entoure l'origine 0 de $\mathbb{Z}^2$ contient un point de la forme $(k + 1/2, 1/2)$ avec $0 \leq k < n$. En déduire que le nombre de tels circuits est majoré par $n\sigma_{n-1}$.

T4. Justifier brièvement que, lorsque la composante connexe C est finie, nécessairement il existe un chemin auto-évitant de $\tilde{G}$ entourant l'origine 0 tel que pour tous les arcs $\tilde{e}$ du circuit, $\xi_{\tilde{e}} = 0$. En déduire la majoration

$$\mathbb{P}(|C| < +\infty) \leq \sum_{n \geq 1} (1 - p)^n n\sigma_{n-1}.$$

En déduire que $\lim_{p \to 1} f(p) = 1$, et donc que $p_c < 1$.

T5. Extension à la grille $\mathbb{Z}^d$, $d \geq 3$: à chaque arc $e = (x, y)$ de la grille $\mathbb{Z}^d$, qui est par définition une paire de sommets $x, y \in \mathbb{Z}^d$ tels que $|x - y|_1 = 1$, où $|x - y|_1 := \sum_{j=1}^{d} |x_j - y_j|$, on associe une variable aléatoire ξ_e de Bernoulli de paramètre p, les ξ_e étant supposés i.i.d.. Notant $C(d)$ la composante connexe du sous-graphe de $\mathbb{Z}^d$ obtenu en ne conservant que les arcs e tels que $\xi_e = 1$, on définit encore $f_d(p) = \mathbb{P}(|C(d)| = +\infty)$, et $p_c(d) = \inf\{p \in [0, 1] : f_d(p) > 0\}$. Montrer que pour $d \geq 3$ arbitraire, on a encore $0 < p_c(d) < 1$.

2 Partie simulation

S1. On considère la sous-grille $\mathbb{Z}_n^2 = \{(x_1, x_2) \in \mathbb{Z}^2 : |x_1| \leq n, |x_2| \leq n\}$. On fixe $n = 50$ dans ce qui suit. Dessiner quatre réalisations des sous-graphes aléatoires de $\mathbb{Z}_n^2$, correspondant respectivement aux valeurs $p = 0.2, 0.4, 0.6, 0.8$.

S2. On note $f_n(p)$ la probabilité que, dans le sous-graphe aléatoire construit sur $\mathbb{Z}_n^2$, l'origine 0 soit connectée à la "frontière" $F_n := \{x \in \mathbb{Z}_n^2 : |x_1| = n \text{ ou } |x_2| = n\}$. Toujours pour n =50, en faisant $N = 100$ expériences pour chaque valeur de p dans $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$, donner l'estimateur empirique correspondant de $f_n(p)$.

S3. Sur la base des expériences de la question S2, proposer un intervalle de confiance de la forme $[0, p_{max}]$ pour la valeur de p_c, la probabilité critique de percolation.

S4. On se donne, comme dans la question T1, pour chaque arc e de la grille $\mathbb{Z}_n^2$ une variable aléatoire U_e uniforme sur $[0, 1]$, les U_e étant i.i.d.. Sur la base de ce jeu de variables U_e, et pour chaque $p \in [0, 1]$, on construit les variables aléatoires $\xi_e(p) = \mathbf{1}_{U_e \leq p}$, qui nous définissent un sous-graphe de $\mathbb{Z}_n^2$. On définit $p_c(\{U_e\})$ comme l'inf des $p \in [0, 1]$ tels que dans le sous-graphe considéré, il existe un chemin de l'origine 0 à la frontière F_n.

Pour $n = 50$, puis pour $n = 100$, obtenir un histogramme de la distribution de $p_c(\{U_e\})$, en l'évaluant pour une centaine de choix des jeux de variables uniformes $\{U_e\}$. Commenter le contraste entre ces deux histogrammes.

PROJET 16

Equilibrage de charge randomisé

sujet proposé par L. Massoulié

laurent.massoulie@inria.fr

On considère un système de n serveurs. A des instants aléatoires $T_1 < T_2 < \cdots$, des clients arrivent dans le système, et choisissent un serveur pour les traiter. Lorsqu'un client a choisi son serveur, il est placé dans une file d'attente associée. Les serveurs traitent leurs clients par ordre d'arrivée, et les clients quittent le système dès la fin de leur service. On suppose que les temps de service des différents clients sont i.i.d., exponentiellement distribués de paramètre $\mu > 0$. On suppose de plus les intervalles $T_i - T_{i-1}$ entre deux arrivées i.i.d., exponentiellement distribués de paramètre $n * \lambda$, où $\lambda > 0$. On note $X_i(t)$ le nombre de clients ayant choisi le serveur i, encore présents dans le système à l'instant t. On notera $\rho = \lambda/\mu$, et on supposera $\rho < 1$.

1 Partie théorique

T1. On suppose d'abord que les clients choisissent leur serveur de manière aléatoire, i.i.d., uniformément parmi les n serveurs. On admettra que sous ces hypothèses, les processus $X_i(t)$ évoluent indépendamment pour chaque serveur, et que la probabilité $\pi_k(t)$ qu'un serveur arbitraire ait k clients à l'instant t satisfait l'équation différentielle:

$$\frac{d}{dt}\pi_k(t) = \lambda[-\pi_k(t) + \mathbf{1}_{k>0}\pi_{k-1}(t)] + \mu[\pi_{k+1}(t) - \mathbf{1}_{k>0}\pi_k(t)].$$

En déduire que, si les variables $X_i(0)$ sont i.i.d., de loi géométrique $\pi_k = (1-\rho)\rho^k$, $k \geq 0$, alors pour tout $t > 0$ les variables $X_i(t)$ sont encore i.i.d. de même loi.

T2. On suppose que les $X_i(0)$ sont i.i.d. de loi géométrique $\{(1-\rho)\rho^k\}_{k\in\mathbb{N}}$, et que n est grand. Etablir que, lorsque $n \to \infty$, $\frac{\sup_{i\in[n]} X_i(0)}{\ln(n)}$ converge en probabilité vers $-1/\ln(\rho)$, et que pour tout $k \geq 0$, la fraction $f_{\geq k}$ de serveurs i ayant au moins k clients présents à $t = 0$ converge p.s. vers ρ^k.

T3. On suppose maintenant que chaque client, lors de son arrivée à un instant t, consulte deux serveurs choisis indépendamment et uniformément au hasard dans $[n]$, soit i et i', et rejoint celui des deux qui a le plus petit nombre de clients à traiter, $X_i(t)$ ou $X_{i'}(t)$.

Si lors de son arrivée, $f_{\geq k}(t)$ représente la fraction des serveurs ayant au moins k clients présents, pour tout $k \geq 0$, établir que le choix de ce client fera augmenter $f_{\geq \ell}$ de $1/n$ avec probabilité $f_{\geq \ell-1}(t)^2 - f_{\geq \ell}(t)^2$.

On admettra que, pour n grand, les fractions $f_{\geq k}(t)$ satisfont les équations différentielles suivantes:

$$\frac{d}{dt} f_{\geq k}(t) = \lambda[f_{\geq k-1}(t)^2 - f_{\geq k}(t)^2] - \mu[f_{\geq k}(t) - f_{\geq k+1}(t)],\ k \geq 1.$$

Etablir qu'un point fixe de ces équations est donné par

$$f_{\geq k} = \rho^{2^k - 1},\ k \geq 0.$$

En déduire que le nombre moyen de serveurs, pour ces fractions limites $f_{\geq k}$, ayant au moins $k = \frac{1}{\ln 2} \ln(1 + 2\ln(n)/\ln(1/\rho))$ tend vers 0 lorsque $n \to \infty$. Avec forte probabilité, on a donc dans ce cas $\sup_{i \in [n]} X_i = O(\ln(\ln(n)))$, un ordre de grandeur négligeable devant l'ordre $\ln(n)$ de la question précédente.

T4. On suppose enfin que les clients, lors de leur arrivée, rejoignent un serveur ayant le nombre minimal de clients en attente. En particulier ils choisissent systématiquement un serveur oisif (i.e. tel que $X_i = 0$) s'il y en a un. Dénotant à nouveau par $f_{\geq k}$ la fraction des n serveurs ayant au moins k clients lors de l'arrivée d'un nouveau client, quelle est la probabilité que ce nouveau client fasse augmenter de $1/n$ la fraction $f_{\geq \ell}$?

On admettra que lorsque n est grand, ces fractions $f_{\geq k}$ satisfont approximativement les équations différentielles:

$$\frac{d}{dt} f_{\geq k} = -\mu[f_{\geq k} - f_{\geq k+1}] + \lambda \mathbf{1}_{f_{\geq k-1}=1} \mathbf{1}_{f_{\geq k}<1},\ k \geq 1.$$

En donner un point fixe. Dans le cas présent, on s'attend à ce que, pour n grand, avec forte probabilité le maximum des X_i en régime stationnaire soit égal à 1.

2 Partie simulation

S1. Ecrire un simulateur à événements discrets traquant: le temps total simulé, le temps résiduel de service des clients en cours de service à chacun des n serveurs, le nombre de clients présents à chaque serveur, et le temps résiduel avant la prochaine arrivée de clients.

Faire le graphique donnant l'évolution de $\sup_{i \in [n]} X_i(t)$ pour les trois politiques de choix du serveur considérées plus haut, pour le choix de paramètres $n = 100$, $\lambda = 1$, $\mu = 2$, pour une plage de temps $[0, T]$ pour $T = 100$. On effectuera ce graphique pour deux choix de conditions initiales: d'une part, la condition initiale d'un système vide ($X_i(0) = 0$, $i \in [n]$) et d'autre part la condition initiale où chaque serveur a initialement 20 clients à traiter.

S2. Pour les mêmes paramètres $n = 100$, $\lambda = 1$, $\mu = 2$ et $T = 100$, représenter les fractions $f_{\geq k}$ de serveurs ayant au moins k clients à l'instant T, pour les trois politiques de choix de serveur, ainsi que les fractions théoriques correspondantes déterminées précédemment. On effectuera ces simulations avec deux choix de conditions initiales: d'une part, la condition initiale d'un système vide ($X_i(0) = 0$, $i \in [n]$) et d'autre part la condition initiale où chaque serveur a initialement 20 clients à traiter.

S3. En conservant $\lambda = 1$, $\mu = 2$, $T = 100$, prendre la moyenne empirique de $\max_{i \in [n]} X_i(T)$ sur 10 réalisations indépendantes, pour les trois politiques d'équilibrage de charge considérées (choix du serveur: aléatoire, meilleur parmi deux choix aléatoires, meilleur parmi tous) et pour les valeurs de $n = 100, 500, 1000, 2000, 4000$. Commenter par rapport aux prédictions théoriques attendues (en $\ln(n)$, $\ln(\ln(n))$ et 1 respectivement) pour ces quantités.

PROJET 17

Estimation de la taille d'un graphe par marches aléatoires

sujet proposé par L. Massoulié

laurent.massoulie@inria.fr

Etant donné un graphe $G = (V, E)$ fini, non orienté et connexe, on cherche à estimer sa taille. On suppose qu'on a accès au graphe de la manière suivante: on peut accéder à un sommet particulier $i_0 \in V$ du graphe, et ayant accédé à un sommet i du graphe, on peut alors accéder à n'importe quel voisin j de i dans G. Cette situation se produit par exemple lorsque le graphe représente les individus participant à un réseau pair-à-pair et les connexions existant entre ces individus. On peut aussi considérer l'estimation de la taille du web, en supposant qu'on sait uniquement passer d'une page à une autre conectée à la précédente par un lien hypertexte.

On va considérer l'algorithme suivant pour l'estimation de cette taille. On fixe une longueur τ, et un nombre cible ℓ. A chaque étape $t \geq 1$, on construit une marche aléatoire de longueur τ, soit $X_0^t, \ldots, X_\tau^t$ sur le graphe, issue de $X_0^t = i_0$. Par définition, sachant les choix $X_0^t, \ldots, X_{s-1}^t$, X_s^t est sélectionné uniformément au hasard parmi les voisins de X_{s-1}^t dans G. On note $Y_t = X_\tau^t$ le dernier sommet obtenu à la t-ème marche aléatoire.

On dit qu'on a une **collision** à l'étape t si $Y_t \in \{Y_1, \ldots, Y_{t-1}\}$. On note C_ℓ l'instant de la ℓ-ème collision, soit:

$$C_\ell = \inf\{t \geq 1 : \sum_{s=1}^{t} \mathbf{1}_{Y_s \in \{Y_1, \ldots, Y_{s-1}\}} = \ell\}.$$

L'estimateur $\hat{N}$ du nombre de sommets $N := |V|$ du graphe est alors donné par

$$\hat{N} = \frac{C_\ell^2}{2\ell}.$$

La partie théorique établit des garanties théoriques sur cette procédure d'estimation. La partie simulation consiste à la mise en oeuvre de cet algorithme.

1 Partie théorique

On va supposer par la suite que le graphe G considéré est d-régulier, i.e. chaque sommet $i \in V$ a d voisins dans G. On notera par ailleurs A la matrice d'adjacence du graphe G, qui est par définition la matrice carrée symétrique, dont les lignes et les colonnes sont indicées par les sommets $i \in V$ du graphe, et telle que $A_{ij} = 1$ si (i, j) est un arc du graphe, et $A_{i,j} = 0$ sinon.

T1. Exprimer la loi de probabilité de chaque échantillon Y_t, soit $\pi := \{\mathbb{P}(Y_t = i)\}_{i \in V}$ au moyen de la matrice P^τ, où $P := d^{-1}A$.

T2. En supposant que le spectre de la matrice P est constitué de la valeur propre 1 et d'autres valeurs propres $\lambda_2, \ldots, \lambda_N$ dont les modules sont majorés par $1 - \epsilon$, pour $\epsilon > 0$, et en utilisant le théorème spectral pour P, en déduire que la loi π des Y_t converge pour la norme $\|\cdot\|_2$, lorsque τ tend vers l'infini, vers la loi uniforme sur V.

T3. On fait maintenant l'approximation suivante au vu de la question précédente: on suppose que τ est choisi suffisamment grand, de sorte qu'on peut supposer les Y_t i.i.d., et **uniformément distribués** sur V. Notant $C_1, \ldots, C_\ell$ les instants des ℓ collisions s'étant produites dans l'implémentation de l'algorithme, établir la formule suivante pour $n, m > 0$:

$$\mathbb{P}(C_\ell - C_{\ell-1} > n | C_{\ell-1} = m) = \frac{(N-m+\ell-1)(N-m+\ell-2)\cdots(N-m+\ell-n)}{N^n}.$$

T4. Etablir, pour tout ℓ fixé et tous $a, b > 0$ fixés, la convergence

$$\lim_{N\to\infty} \mathbb{P}(C_\ell - C_{\ell-1} > b\sqrt{N} | C_{\ell-1} = a\sqrt{N}) = e^{-ab-b^2/2}.$$

En déduire la convergence, pour tous $x, y > 0$ fixés:

$$\lim_{N\to\infty} \mathbb{P}\left([C_\ell^2 - C_{\ell-1}^2]/(2N) > y | C_{\ell-1}^2/(2N) = x\right) = e^{-y}.$$

Ce dernier résultat entraîne, par récurrence sur ℓ, la convergence en distribution $(2N)^{-1}C_\ell^2 \overset{N\to\infty}{\to} E_1 + \ldots + E_\ell$, où $E_1, \ldots E_\ell$ sont i.i.d. et exponentiellement distribuées. Avec quelques étapes supplémentaires (non demandées!), on peut en déduire que l'estimateur proposé $\hat{N} := C_\ell^2/(2\ell)$ vérifie $\mathbb{E}\hat{N}/N \sim 1$, et Var$(\hat{N}/N) = \Theta(1/\ell)$. Ces propriétés suggèrent que $\hat{N}$ est un estimateur raisonnable de N lorsque $N \gg 1$ et ℓ est suffisamment grand.

2 Partie simulation

S1. Implémenter l'algorithme proposé initialement, en remplaçant la marche aléatoire standard par la variante suivante: pour passer de X_s^t à X_{s+1}^t, avec probabilité 1/10 on prend $X_{s+1}^t = X_s^t$, et avec probabilité $9/10$ on prend pour X_{s+1}^t un voisin de X_s^t uniformément au hasard dans G. Cette variante est dite marche aléatoire paresseuse, elle a pour intérêt qu'elle vérifie la propriété spectrale de la question T2, ce qui n'est pas le cas pour la marche aléatoire standard sur un graphe bi-partite, pour lequel la valeur propre -1 appartient au spectre de la matrice P.

S2. Tester l'algorithme avec comme graphe $G = (V, E)$ l'hypercube à δ dimensions, i.e. $V = \{u = (u_1, \ldots, u_\delta) \in \{0,1\}^\delta\}$, et

$$E = \{(u, v) \in V^2 : \sum_{i=1}^{\delta} |u_i - v_i| = 1\}.$$

On prendra en particulier $\delta = 10$, et on dessinera les histogrammes des valeurs $\hat{N}$ de l'estimation de $N = 1024$ correspondant aux choix de paramètres $\tau = 5, 10, 50, 100$, et $\ell = 1, 10, 100$, avec pour chaque choix de valeurs (τ, ℓ) 100 réalisations de l'estimateur $\hat{N}$.

On commentera l'impact des deux paramètres τ, ℓ sur la qualité des estimations obtenues, mesurées par l'erreur quadratique relative $\sqrt{(1/100)\sum_{i=1}^{100}(\hat{N}_i/N - 1)^2}$ sur les 100 réalisations.

S3. Tester maintenant l'algorithme sur le cycle C_N, par définition constitué des sommets $1, \ldots, N$ et des arcs $(i, i+1)$, $i = 1, \ldots, N-1$, et de l'arc $(N, 1)$. On dessinera les histogrammes des valeurs $\hat{N}$ de l'estimation de $N = 50$ correspondant aux choix de paramètres $\tau = 10, 100, 1000, 10000$, et $\ell = 1, 10, 100$, avec pour chaque choix de valeurs (τ, ℓ) 100 réalisations de l'estimateur $\hat{N}$. Comme dans la question précédente on commentera l'impact de τ et ℓ sur la qualité des estimations obtenues.

On commentera la différence entre la performance de l'estimateur dans le cas précédent de l'hypercube et le cas précédent du cycle.

PROJET 18

Algorithme de Wilson pour la génération d'arbres couvrants uniformes

sujet proposé par L. Massoulié

laurent.massoulie@inria.fr

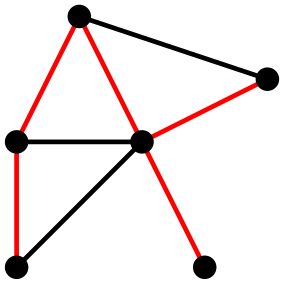


Figure 18.1: Exemple de graphe avec un arbre couvrant en rouge.

Un graphe $G = (V, E)$ non orienté est défini par l'ensemble de sommets V et l'ensemble E d'arcs, qui sont des paires de sommets[1]. Un graphe est dit connexe si chaque paire de sommets i, j est reliée par un chemin (i, i_1), (i_1, i_2),... (i_k, j) d'arcs du graphe. Un graphe est un **arbre** s'il est connexe et ne contient aucun cycle non trivial, ou de manière équivalente, s'il est connexe et tel que $|E| = |V| - 1$.

Pour un graphe $G = (V, E)$, un **arbre couvrant** de G est un graphe arbre $T = (V(T), E(T))$, tel que $V(T) = V$, et $E(T) \subset E$. En d'autres termes, T est un arbre constitué d'arcs de G et connectant tous les sommets de G. Le nombre d'arbres couvrants d'un graphe G pouvant être exponentiel en le nombre de sommets de G, il n'est a priori pas évident d'obtenir un arbre couvrant choisi uniformément distribué au hasard.

L'algorithme de Wilson permet, pour un graphe non orienté connexe G, avec ensemble fini de sommets V, de simuler un arbre couvrant uniformément au hasard parmi tous les arbres couvrants de G. Il procède de la manière suivante.

Initialisation: $T_0 = \{r\}$, arbre réduit à un unique sommet $r \in V$ choisi arbitrairement.

[1] les graphes considérés sont non orientés donc E vérifie la propriété de symétrie: si $(i, j) \in E$, alors $(j, i) \in E$.

Itération, étape i: ayant construit un arbre T_{i-1}, prendre un sommet $u \in V \setminus V(T_{i-1})$, où $V(T_{i-1})$ désigne l'ensemble des sommets de l'arbre T_{i-1}.

(a) Engendrer une marche aléatoire $\{U_0, \ldots, U_N\}$ sur G partant de u, avec probabilités de transition du sommet k vers le sommet l donnée par $P_{k,\ell} = d_k^{-1} 1_{(k,\ell)\in E}$, où d_k est le nombre de voisins de k dans G[2], issue de $U_0 = u$, et arrêtée au premier instant N auquel elle atteint l'un des sommets de T_{i-1}.

(b) Appliquer la procédure suivante d'effacement de boucles à cette marche: pour

$$n = \inf\{t \in \{1, \ldots, N\} : \exists s < t \text{ tel que } U_s = U_t\},$$

on efface la boucle $U_s, \ldots, U_{n-1}$ pour conserver uniquement $U_0, \ldots, U_{s-1}, U_n, \ldots, U_N$. Itérer cette procédure d'effacement jusqu'à absence de telles boucles. Le chemin résultant est noté $V_0 = u, V_1, \ldots, V_M = U_N$. On définit alors l'arbre T_i comme l'union de T_{i-1} et du chemin $V_0, \ldots, V_M$.

L'algorithme termine lorsqu'un arbre couvrant est obtenu.

La partie théorique établit que l'algorithme de Wilson produit bien un arbre couvrant uniforme. La partie simulation consiste à mettre en oeuvre cet algorithme.

1 Partie théorique

Pour chaque sommet k de G distinct de r, on se donne une séquence i.i.d. $\{A_k^t\}_{t\geq 1}$ de voisins de k dans G choisis uniformément au hasard. Ces séquences sont mutuellement indépendantes entre elles.

On interprète ces variables de la manière suivante: dans notre implémentation de l'algorithme de Wilson, la première fois qu'un sommet k est atteint par une des marches construites, son étape suivante est A_k^1; la deuxième fois, l'étape suivante est A_k^2; la t-ème fois, c'est A_k^t. L'ensemble des arcs orientés (k, A_k^1) définit un graphe orienté.

T1. On envisage de telles séquences comme des piles (d'épaisseur infinie) de choix aléatoires de voisins pour chaque sommet $k \neq r$. Considérant le graphe dirigé défini par les choix A_k^1 au sommet des piles, si celui-ci contient un cycle orienté $C = k_0, k_1 = A_{k_0}^1, \ldots, k_0 = A_{k_\ell}^1$, on s'autorise à le supprimer, c'est à dire à supprimer du haut des piles associées les choix correspondants $A_{k_0}^1, \ldots, A_{k_\ell}^1$. Un tel cycle est dit **directement suppressible**.

On appelle cycle toute séquence $C = (k_0, k_1 = A_{k_0}^{t_0}, \ldots, k_0 = A_{k_\ell}^{t_\ell})$. On dira qu'un cycle $C = (k_0, k_1 = A_{k_0}^{t_0}, \ldots, k_0 = A_{k_\ell}^{t_\ell})$ est **suppressible** si il existe une séquence de cycles $C_1, \ldots, C_k = C$ tels qu'on peut supprimer séquentiellement les cycles C_1, puis C_2, etc. jusqu'à $C_k = C$.

Prouver que pour tout cycle C suppressible, et tout cycle C' directement suppressible, alors on peut supprimer C via une séquence de cycles qui commence par le cycle C'.

T2. En déduire que seuls deux scénarios sont possibles: soit après chaque suppression d'un cycle arbitraire, il reste toujours un cycle qu'on peut encore supprimer (le processus de suppression des cycles ne termine jamais), soit le processus de suppression des cycles termine, auquel cas les cycles enlevés ne dépendent pas de l'ordre dans lequel on les a enlevés, et les valeurs au sommet des piles dans la configuration terminale ne dépendent pas non plus de cet ordre.

T3. Montrer qu'avec probabilité 1, l'algorithme de Wilson termine en un nombre d'étapes fini. En interprétant l'algorithme de Wilson comme un enlèvement de cycles, en déduire que avec probabilité 1 le processus de suppression des cycles termine.

[2]Autrement dit, il s'agit d'une marche qui se déplace en choisissant à chaque pas un voisin uniformément parmi les voisins de sa position actuelle.

T4. Soit $C_1, \ldots, C_M$ l'ensemble de cycles enlevables étant données les piles $\{A_k^t\}_{t\geq 1}$, $k \in V \setminus \{r\}$, et soit $\vec{T}$ le graphe orienté, libre de cycles, obtenu dans la configuration terminale après suppression des cycles. Justifier que la collection de cycles $\{C_1, \ldots, C_M\}$ est indépendante de $\vec{T}$. En déduire que l'arbre T obtenu en retirant l'orientation des arcs de $\vec{T}$ est uniformément distribué parmi les arbres couvrants de G.

2 Partie simulation

S1. Implémenter l'algorithme de Wilson: en prenant comme entrée un graphe non orienté G, obtenir en sortie un échantillon uniforme d'un arbre couvrant de G.

S2. Tester l'algorithme avec $G = (V, E)$ une grille à deux dimensions, c'est-à-dire pour $V = \{(x, y), x \in \{1, \ldots, n\}, y \in \{1, \ldots, n\}\}$ et $E = \{\{(x, y), (x', y')\} \in V^2 : |x - x'| + |y - y'| = 1\}$. Engendrer des images de 2 arbres couvrants résultants pour la grille avec $n = 10$, puis avec $n = 50$.

S3. Tester l'algorithme avec $G = (V, E)$ une grille triangulaire, i.e. $V = \{(x, y), x \in \{1, \ldots, n\}, y \in \{1, \ldots, x\}\}$ et

$$E = \{\{(x, y), (x', y')\} \in V^2 : \text{ soit } x = x' \text{ et } |y - y'| = 1, \text{ soit } x' = x + 1 \text{ et } y' \in \{y, y + 1\}\}.$$

Engendrer des images de 2 arbres couvrants résultants pour la grille triangulaire avec $n = 10$, puis avec $n = 50$.

PROJET 19

Composante géante des graphes aléatoires

sujet proposé par L. Massoulié

laurent.massoulie@inria.fr

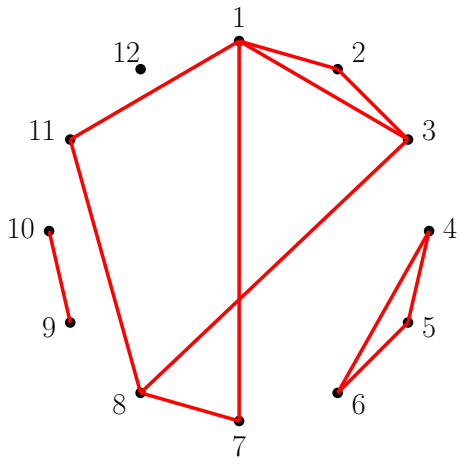


Figure 19.1: Exemple de graphe d'Erdős-Rényi avec 12 sommets. Il y a 4 composantes connexes de tailles 6, 3, 2 et 1.

On s'intéresse à la taille des composantes connexes de graphes aléatoires dits d'Erdős-Rényi. Par definition un tel graphe de paramètres $n \in \mathbb{N}$ et $p \in [0,1]$, qu'on note $\mathcal{G}(n,p)$ est constitué de n sommets, qu'on identifie à $[n] := \{1,\dots,n\}$, et pour chaque paire non orientée de sommets $(u,v) \in [n]$, $u \neq v$, l'arc (u,v) est présent dans le graphe avec probabilité p, indépendamment de la présence des autres arcs.

On note $C(u)$ la composante connexe du graphe contenant le sommet u. On note aussi X_1 la taille de la plus grande composante connexe, mesurée en nombres de sommets, et X_2 la taille de la seconde plus grande composante connexe.

On suppose dans la suite que $p = \lambda/n$ pour une constante $\lambda > 0$ fixe.

T1. Pour un sommet $u \in [n]$ fixé, on note $\mathcal{X}_d(u)$ l'ensemble des sommets à distance d de u dans $G = \mathcal{G}(n, \lambda/n)$, $X_d(u) = |\mathcal{X}_d(u)|$ son cardinal, et $X_{\leq d}(u) = 1 + \sum_{i\in[d]} X_i(u)$. On note aussi $\mathcal{F}_d(u) = \sigma(\mathcal{X}_1(u), \ldots, \mathcal{X}_d(u))$.

Montrer que, conditionnellement à $\mathcal{F}_d(u)$, $X_{d+1}(u)$ suit la loi Bin$(n - X_{\leq d}, 1 - (1 - \lambda/n)^{X_d(u)})$.

T2. En utilisant l'inégalité $1 - (1 - p)^m \leq mp$, valable pour $p \in [0, 1]$ et $m \in \mathbb{N}$, déduire que pour tout $\theta > 0$, on a

$$\mathbb{E}(e^{\theta X_{d+1}(u)} | \mathcal{F}_d(u)) \leq e^{X_d(u)\lambda(e^\theta - 1)}.$$

En déduire, lorsque $\lambda < 1$, l'existence de $r \in (0, 1)$ et de $\theta^* > 0$ tels que pour tout $\theta \in (0, \theta^*]$,

$$\mathbb{E}(e^{\theta X_{d+1}(u)} | \mathcal{F}_d(u)) \leq e^{r\theta X_d(u)}. \tag{19.1}$$

T3. Déduire de (19.1) que, pour $\theta > 0$ suffisamment petit, $\mathbb{E}(e^{\theta|C(u)|}) < +\infty$, où $|C(u)| = 1 + \sum_{d\geq 1} X_d(u)$ est la taille de la composante connexe $C(u)$.

T4. Déduire de la question précédente que pour $\theta > 0$ suffisamment petit,

$$\lim_{n\to+\infty} \mathbb{P}(\exists u \in [n] : |C(u)| \geq 2\frac{\log(n)}{\theta}) = 0.$$

Ainsi, pour $\lambda < 1$, et une constante $c > 0$ convenablement choisie, alors

$$\lim_{n\to\infty} \mathbb{P}(X_1 \leq c \ln(n)) = 1.$$

En d'autres termes la plus grande composante géante est de taille logarithmique. A fortiori, on a pour $\lambda < 1$, et pour tout $\epsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}(X_1/n > \epsilon) = 0, \tag{19.2}$$

c'est à dire que la taille renormalisée X_1/n tend vers 0 en probabilité.

Par contraste, lorsque $\lambda > 1$, on a pour tout $\epsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}(|X_1/n - [1 - p_{ext}(\lambda)]| \leq \epsilon) = 1 \tag{19.3}$$

c'est à dire convergence en probabilité de X_1/n vers la constante $1-p_{ext}(\lambda)$, où $p_{ext}(\lambda)$ est par définition la probabilité d'extinction d'un processus de branchement de Galton-Watson, démarré avec un unique ancêtre, et où chaque individu a un nombre d'enfants distribué selon la loi de Poisson de paramètre λ.

On a de plus, pour une fonction $g(\lambda) > 0$,

$$\lim_{n\to\infty} \mathbb{P}(X_2 \leq g(\lambda) \ln(n)) = 1$$

c'est à dire que la seconde plus grande composante connexe est de taille logarithmique en n.

On peut donner une justification heuristique de ces résultats comme suit: pour un sommet arbitraire $u \in [n]$, le voisinage du sommet u dans le graphe $\mathcal{G}(n, \lambda/n)$ jusqu'à une distance fixée d tend en loi, lorsque $n \to \infty$, vers un arbre de Galton-Watson avec u pour ancêtre et nombre d'enfants de loi Poisson(λ). Pour une fraction $p_{ext}(\lambda)$ des sommets $u \in [n]$, leur voisinage a une taille petite, ce qui correspond à l'extinction du processus de branchement. Par contre tous les autres sommets ont un voisinage qui s'étend, jusqu'à rejoindre un unique "composant géant", la plus grande composante connexe du graphe, de taille $X_1 \sim (1 - p_{ext}(\lambda))n$.

Le résultat classique de Galton-Watson caractérise $p_{ext}(\lambda)$ comme la plus petite racine $z \in [0, 1]$ de l'équation

$$z = \phi(z),$$

où $\phi(z) = e^{-\lambda(1-z)}$ est la fonction caractéristique de la variable de Poisson de paramètre λ. La propriété (19.2) est en fait un cas particulier de (19.3), puisque $p_{ext}(\lambda) = 1$ pour $\lambda \leq 1$.

S1. Pour $n = 200$, et pour des valeurs de $\lambda = 0.9,\ 1.1,\ 1.5,\ 2,\ 2.5$, effectuer une vingtaine de simulations de chaque $\mathcal{G}(n, \lambda/n)$.

Comparer la moyenne empirique des tailles observées pour X_1/n à la valeur prédite en théorie, soit $1 - p_{ext}(\lambda)$. On utilisera un schéma itératif $p(t+1) = \phi(p(t))$, $p(0) = 0$, pour déterminer $p_{ext}(\lambda)$ lorsque $\lambda > 1$. La prédiction théorique est-elle précise pour les paramètres considérés?

On considère maintenant une version multi-types de graphes aléatoires décrite comme suit. Chacun des n sommets $u \in [n]$ possède un type $\sigma(u) \in \{-1, 1\}$ déterminé de manière i.i.d., avec $\mathbb{P}(\sigma(u) = +1) = r$ pour un $r \in [0, 1]$. Conditionnellement aux $\sigma(u)$, un arc (u, v) est présent avec probabilité $M_{\sigma(u),\sigma(v)}/n$ pour une matrice symétrique $M \in \mathbb{R}_+^2$.

Le voisinage d'un sommet dans ce graphe aléatoire est encore asymptotiquement distribué comme un arbre de branchement de Galton-Watson, mais cette fois-ci avec plusieurs types $\sigma = +1$ ou -1. Plus précisément, chaque individu de type σ donne naissance à X_+ enfants de type $+$ et X_- enfants de type $-$, où X_-, X_+ sont indépendants, de lois de Poisson de paramètres respectifs: $\lambda_{\sigma-} = (1-r)M_{\sigma-}$, et $\lambda_{\sigma+} = rM_{\sigma+}$.

Pour un tel processus de branchement, on sait que la probabilité $p_{\sigma,ext}$ d'extinction, sachant que l'ancêtre est de type $\sigma \in \{+, -\}$, est égale à 1 pour tout σ si la matrice $\Lambda := (\lambda_{\sigma,\tau})_{\sigma,\tau\in\{+,-\}}$ a un rayon spectral $\rho(\Lambda)$ inférieur ou égal à 1. Le vecteur $(p_{\sigma,ext})_{\sigma\in\{+,-\}}$ des probabilités d'extinction est la plus petite solution (pour l'ordre partiel de comparaison de chaque coordonnée) de l'équation de point fixe:

$$\forall \sigma \in \{+, -\}, \quad p_{\sigma,ext} \in [0, 1],\ p_{\sigma,ext} = e^{-\lambda_{\sigma+}(1-p_{+,ext})} e^{-\lambda_{\sigma-}(1-p_{-,ext})}.$$

De ces probabilités d'extinction on peut déterminer la taille et la constitution de la plus grande composante géante du graphe: si on note X_σ le nombre de sommets de type $\sigma \in \{+, -\}$ dans la plus grande composante géante, alors on a la convergence en probabilité suivante:

$$\lim_{n\to\infty} \frac{1}{n} X_+ = r(1 - p_{+,ext}),\ \lim_{n\to\infty} \frac{1}{n} X_- = (1-r)(1 - p_{-,ext}). \tag{19.4}$$

S2. Pour $r = 0.3$, $M_{++} = M_{--} = 4$, $M_{+-} = M_{-+} = 1$, et pour $n = 200$, simuler une vingtaine de répliques du graphe aléatoire multi-types décrit ci-dessus. Comparer les moyennes empiriques des tailles observées pour X_+/n, X_-/n aux valeurs prédites en théorie données en (19.4). La prédiction théorique est-elle précise pour les paramètres considérés?

PROJET 20

Connectivité des graphes aléatoires

sujet proposé par L. Massoulié

laurent.massoulie@inria.fr

On s'intéresse à des graphes aléatoires dits d'Erdős-Rényi. Par définition un tel graphe de paramètres $n \in \mathbb{N}$ et $p \in [0,1]$, qu'on note $\mathcal{G}(n,p)$ est constitué de n sommets, qu'on identifie à $[n] := \{1,\dots,n\}$, et pour chaque paire non orientée de sommets $(u,v) \in [n]$, $u \neq v$, l'arc (u,v) est présent dans le graphe avec probabilité p, indépendamment de la présence des autres arcs.

Un graphe est connecté si et seulement si il est constitué d'une unique composante connexe, i.e. de chaque sommet u il existe un chemin d'arcs dans le graphe le reliant à tout autre sommet v.

On s'intéresse à la probabilité que le graphe $\mathcal{G}(n,p)$ soit connecté. Un résultat dû à Erdős et Rényi établit que pour toute constante $c \in \mathbb{R}$, on a:

$$\lim_{n\to\infty} \mathbb{P}\left(\mathcal{G}(n,(\ln(n)+c)/n) \text{ connecté}\right) = e^{-e^{-c}}. \tag{20.1}$$

Ceci montre que les graphes d'Erdős-Rényi $\mathcal{G}(n,p)$ sont connectés avec probabilité approchant 1 si le degré moyen des sommets $D := (n-1)p$ vérifie $D - \ln(n) \gg 1$ (i.e. c positif et grand), tandis qu'ils sont déconnectés avec probabilité approchant 1 si $\ln(n) - D \gg 1$ (i.e. c négatif et grand en valeur absolue). On considère une constante $c \in \mathbb{R}$ fixée, on pose $p = (\ln(n)+c)/n$ et on considère une réalisation G de $\mathcal{G}(n,p)$.

On dit qu'un sommet u est isolé s'il n'a aucun voisin (en d'autres termes, si son degré est nul). Pour tout $u \in [n]$ on pose $Z_u = \mathbf{1}_u$ sommet isolé de G. Enfin on note $X = \sum_{u\in[n]} Z_u$ le nombre des sommets isolés dans le graphe.

Pour un entier k fixe, on note $X^{\underline{k}} = X(X-1)\cdots(X-k+1)$.

T1. Montrer que pour tout $k \in \mathbb{N}$, on a

$$\begin{aligned}\mathbb{E}(X^{\underline{k}}) &= \textstyle\sum_{u_1,\dots u_k} \mathbb{E}(Z_{u_1}\cdots Z_{u_k}) \\ &= n^{\underline{k}}\mathbb{E}(Z_1\cdots Z_k) \\ &= n^{\underline{k}}(1-p)^{k(n-k)+\binom{k}{2}},\end{aligned}$$

où la somme porte sur toutes les suites d'entiers distincts $u_1,\dots,u_k$ de $[n]$, et en déduire:

$$\lim_{n\to\infty} \mathbb{E}(X^{\underline{k}}) = e^{-kc}.$$

On admettra que cette convergence pour tout $k \in \mathbb{N}$ des moments descendants $\mathbb{E}(X^{\underline{k}})$ de X vers ceux d'une variable aléatoire de Poisson de paramètre $\lambda = e^{-c}$ lorsque $n \to \infty$ entraîne la convergence en loi de X vers la loi de Poisson de paramètre e^{-c}.

T2. En déduire que la probabilité $\mathbb{P}(X = 0)$ que $\mathcal{G}(n,p)$ n'ait aucun sommet isolé vérifie

$$\lim_{n\to\infty} \mathbb{P}(X = 0) = e^{-e^{-c}}. \tag{20.2}$$

T3. Montrer que pour un entier $k > 1$, la probabilité que $\mathcal{G}(n,p)$ ait une composante connexe de taille k est majorée par:

$$\binom{n}{k}\mathbb{P}([k] \text{ composante connexe de } \mathcal{G}(n,p)) = \binom{n}{k}\mathbb{P}(\mathcal{G}(k,p) \text{ connecté})(1-p)^{k(n-k)}. \tag{20.3}$$

On peut alors majorer grossièrement $\mathbb{P}(\mathcal{G}(k,p)$ connecté$)$ par $T_k p^{k-1}$, où T_k est le nombre d'arbres sur $[k]$, et p^{k-1} est la probabilité que chacun des $k-1$ arcs d'un arbre particulier sur $[k]$ est présent dans $\mathcal{G}(k,p)$.

Un théorème de Cayley établit que $T_k = k^{k-2}$; la borne donnée en (20.3) est donc à son tour majorée par:

$$\binom{n}{k} k^{k-2} p^{k-1} (1-p)^{k(n-k)}.$$

De cette majoration, on peut déduire que

$$\lim_{n\to\infty} \mathbb{P}(\mathcal{G}(n,p) \text{ a une composante connexe de taille } k \in \{2,\ldots,n-2\}) = 0.$$

T4. Etablir que cette dernière propriété, combinée avec le résultat (20.2) de la première question, implique le résultat d'Erdős-Rényi (20.1).

S1. Ecrire un programme qui détermine si un graphe non orienté G est connexe.

Réaliser une vingtaine de simulations de $\mathcal{G}(n,p)$ pour $n = 1000$ et pour chaque valeur de $p = k/n$, $k = 6.5,\ 7.0,\ 7.5,\ 8.0,\ 8.5,\ 9.0$.

S2. Comparer, pour chaque valeur de p, la fraction des graphes simulés qui sont connexes, et comparer cette fraction à $e^{-e^{-c}}$, pour $c = np - \ln(n)$. Le résultat asymptotique (20.1) d'Erdős-Rényi fournit-il une bonne approximation de $\mathbb{P}(\mathcal{G}(n,p)$ connecté$)$ pour les valeurs de n et p considérées?

On considère maintenant un graphe orienté $\mathcal{G}'(n,p)$ sur les n sommets $[n]$, où chaque arc orienté (u,v), $u \neq v$ est présent avec probabilité p, et ce indépendamment de la présence des autres arcs. Chaque sommet u du graphe a alors un degré entrant $d_u^{in} = \sum_{v\neq u} \mathbf{1}_{\text{arc } (v,u) \text{ présent}}$, et un degré sortant $d_u^{out} = \sum_{v\neq u} \mathbf{1}_{\text{arc } (u,v) \text{ présent}}$. On note alors X_{in} (respectivement, X_{out}) le nombre de sommets $u \in [n]$ de degré rentrant d_u^{in} (respectivement, sortant d_u^{out}) égal à zéro.

T5. Justifier brièvement que, pour $p = (\ln(n) + c)/n$ avec $c \in \mathbb{R}$ fixé, X_{in} et X_{out} ont une loi binomiale de paramètres $(n-1, (1-p)^{n-1})$, et que celle-ci tend vers la loi de Poisson de paramètre e^{-c} lorsque $n \to \infty$.

On dit qu'un graphe orienté est fortement connexe si pour toute paire de sommets distincts u, v, il existe

un chemin orienté allant de u à v. Une condition nécessaire pour qu'un graphe soit fortement connexe est que chaque sommet u ait ses degrés entrant d_u^{in} et sortant d_u^{out} non nuls.

Un argumentaire semblable à celui fait pour les graphes non-orientés établit que, pour $p = (\ln(n) + c)/n$ et $c \in \mathbb{R}$ fixé, on a

$$\lim_{n\to\infty} \mathbb{P}(\mathcal{G}'(n,p) \text{ connexe}) = \lim_{n\to\infty} \mathbb{P}(X_{out} = X_{in} = 0) = e^{-2e^{-c}}. \tag{20.4}$$

S3. Ecrire un programme déterminant si un graphe orienté est fortement connexe. Tester, pour $n = 1000$ et $p = k/n$, $k = 6.5,\ 7.0,\ 7.5,\ 8.0,\ 8.5$, la validité de la formule asymptotique (20.4) en simulant une vingtaine de graphes $\mathcal{G}'(n,p)$ pour les paramètres correspondants.

PROJET 21

Modélisation d'une épidémie

sujet proposé par A. Véber

amandine.veber@polytechnique.edu

On cherche à modéliser la propagation d'une épidémie et à comprendre à quelle condition celle-ci va perdurer ou, au contraire, s'éteindre en touchant un nombre restreint d'individus. Dans la suite, on considèrera plusieurs types de populations et on comparera l'évolution de l'épidémie au sein de chacune. *Attention, les modèles abordés ici ne visent pas à modéliser la pandémie de covid-19, ce projet n'a qu'un intérêt pédagogique lié au cours de MAP361.*

1 Propagation dans une population infinie

Supposons que la population est initialement composée d'une infinité d'individus sains et d'un seul individu infecté. Le temps est discret (par exemple, on considère l'évolution jour après jour) et on suppose que chaque individu infecté à l'étape n disparaît (*i.e.*, se rétablit) à l'étape $n+1$ en ayant contaminé un nombre d'individus encore sains distribué suivant la loi d'une variable aléatoire Z à valeurs dans $\mathbb{N}$. Les contaminations se font indépendamment les unes des autres, de sorte que les nombres d'infections émanant d'individus distincts au temps n sont indépendants et identiquement distribués. En notant I_n le nombre d'infectés au temps n, la suite $(I_n)_{n\geq 0}$ forme donc un processus de Galton-Watson de loi de reproduction (la loi de) Z. Pour fixer les notations, on écrira pour tout $n \geq 0$

$$I_{n+1} = \sum_{k=1}^{I_n} Z_{n,k},$$

où les $Z_{n,k}$ forment une famille de variables aléatoires indépendantes et de même loi que Z. Par convention, la somme ci-dessus est vide si $I_n = 0$ et on a alors $I_{n+1} = 0$.

Notons $m = \mathbb{E}[Z]$ et supposons que cette quantité est finie.

S1. Simuler l'évolution de $(I_n)_{n\geq 0}$ lorsque Z suit une loi de Poisson de paramètre a, pour $a = 0.9$, $a = 1$ et $a = 1.1$. Pour chaque valeur de a, on tracera sur un même graphe 10 réalisations de $(I_n)_{0\leq n\leq 100}$ sur l'intervalle de temps $0 \leq n \leq 100$.

T1. Montrer que si $k \in \mathbb{N}$ et $Z_1, \ldots, Z_k$ sont des variables aléatoires indépendantes de loi de Poisson de paramètres respectifs $\lambda_1, \ldots, \lambda_k$, alors $Z_1 + \cdots + Z_k$ suit une loi de Poisson de paramètre $\lambda_1 + \cdots + \lambda_k$. En déduire une manière d'accélérer le code de la question précédente.

S2. Quel problème se pose lorsque l'on essaie de répondre à la question S1 avec $a = 2$?

S3. Pour chaque valeur de $a \in \{0.9, 1, 1.1\}$, en utilisant 5000 réalisations de $(I_n)_{n\geq 0}$, donner une estimation de la probabilité que I_{50} soit égal à 0 (autrement dit, que l'épidémie s'éteigne au plus tard au temps 50), ainsi qu'un intervalle de confiance au niveau 95% de cette probabilité. (*Indication:* on pourra appliquer le théorème de la limite centrale à

$$\frac{1}{N}\sum_{k=1}^{N} 1_{\{I_{50}^{(k)}=0\}}$$

où $I^{(k)}$ est la $k$ième réalisation de la suite I.)

T2. Dans le cas général où Z est une variable aléatoire intégrable, montrer que pour tout $n \geq 0$ et tout $j \geq 0$, $\mathbb{E}[I_{n+1} \mid I_n = j] = mj$ et par conséquent que $\mathbb{E}[I_{n+1} \mid I_n] = mI_n$. En déduire que $\mathbb{E}[I_n] = m^n$.

T3. A quel comportement s'attend-on pour I_n lorsque $n \to \infty$, en fonction de m ? (On ne demande pas de preuve ici.)

On supposera dans la suite que $\mathbb{P}[Z = 1] < 1$.

T4. Que se passerait-il, sinon ?

T5. Posons pour tout $s \in [0, 1]$, $\phi_1(s) = \mathbb{E}[s^Z] = \mathbb{E}[s^{I_1}]$ et $\phi_n(s) = \mathbb{E}[s^{I_n}]$. En calculant $\mathbb{E}[s^{I_{n+1}} \mid I_n = j]$ pour tout $j \in \mathbb{N}$, montrer que

$$\mathbb{E}\big[s^{I_{n+1}} \mid I_n\big] = \mathbb{E}\big[s^Z\big]^{I_n}.$$

En déduire que pour tout $n \geq 1$, $\phi_n(s) = \phi_1 \circ \phi_1 \circ \cdots \circ \phi_1(s)$, où ϕ_1 est composé n fois.

T6. Montrer que pour tout $n \geq 0$, $\mathbb{P}[I_n = 0] = \phi_n(0)$. En notant T_e le temps d'extinction de l'épidémie, c'est-à-dire le premier temps où $I_n = 0$, montrer que

$$\mathbb{P}[T_e < \infty] = \lim_{n\to\infty} \phi_n(0).$$

S4. Déduire des questions précédentes un algorithme de calcul de $\mathbb{P}[T_e \leq M]$, pour un entier M donné et une loi de Z donnée (dont on connaît l'expression de la fonction génératrice).

S5. Tracer la fonction $M \mapsto \mathbb{P}[T_e \leq M]$ pour les 3 lois utilisées dans la question S1, c'est-à-dire lorsque Z suit une loi de Poisson de paramètre $a = 0.9$, $a = 1$ et $a = 1.1$. On choisira la bonne fourchette de valeurs de M à considérer pour observer correctement le comportement de la probabilité d'extinction et on placera sur le graphe les 3 valeurs approchées de $\mathbb{P}[T_e \leq 50]$ obtenues à la question S3.

T7. Dans les 3 scénarios précédents, comparer le comportement pour M grand de $\mathbb{P}[T_e \leq M]$ à l'intuition demandée à la question T3.

Tous ces calculs nous indiquent si l'épidémie va s'éteindre ou persister lorsque le "stock" d'individus sains est infini. Dans le paragraphe suivant, nous allons voir comment cette probabilité est modifiée par le fait que la population initiale est en réalité finie.

2 Propagation dans une population finie

On suppose cette fois que la population initiale est constituée de N individus sains et un individu infecté. A nouveau, on suppose que chaque individu infecté k au temps n se rétablit au temps $n+1$ après avoir contaminé un nombre $Z_{n,k}$ d'individus encore sains, où les $Z_{n,k}$ sont des variables aléatoires indépendantes et identiquement distribuées, de même loi que Z. Les individus qui se rétablissent ne peuvent pas être réinfectés. La différence avec la partie précédente est que ce nombre d'infections est borné par le nombre S_n d'individus sains encore susceptibles d'être infectés au temps n. Formellement, on a donc $I_0 = 1$, $S_0 = N$ et pour tout $n \geq 0$,

$$I_{n+1} = \min\left\{S_n, \sum_{k=1}^{I_n} Z_{n,k}\right\}$$
$$S_{n+1} = \max\left\{0, S_n - \sum_{k=1}^{I_n} Z_{n,k}\right\}.$$

Rappelons la notation T_e pour le temps d'extinction de l'épidémie, c'est-à-dire le premier temps où I touche 0. On notera également T_i le premier temps auquel la population est totalement infectée, c'est-à-dire où $S_n = 0$ (avec la convention $T_i = +\infty$ si I_n touche 0 avant S_n).

T8. Montrer que l'épidémie se termine en au plus $N+1$ jour (autrement dit, que $I_{N+1} = 0$ presque-sûrement).

S6. Coder l'évolution du système $(S_n, I_n)_{n\geq 0}$ dans le cas où Z suit une loi de Poisson de paramètre $a > 0$ et pour une valeur de N donnée. La fonction demandée aura donc comme arguments les paramètres a et N, ainsi que le nombre de pas de temps P à simuler. Tracer un exemple de réalisation de la suite $(S_n, I_n)_{n\geq 0}$ pour $N = 1000$, $P = 1000$ et a valant 0.9, 1 et 1.1.

S7. Pour les mêmes valeurs de N et a qu'à la question précédente, donner une estimation de $\mathbb{E}[T_e]$ ainsi que de $\mathbb{E}[T_i \,|\, T_i < \infty]$.

S8. Donner une estimation empirique de $\mathbb{P}[T_e < T_i]$ pour $N = 50, 100, 1000$ et 10000, uniquement dans le cas où $a = 0.9$. Comparer à la probabilité d'extinction du processus de Galton-Watson obtenue dans ce cas.

T9. En se basant sur les résultats de la question précédente, dans le cas d'une épidémie sous-critique ($m < 1$), l'épidémie est-elle "aidée" par le fait que la population soit finie ?

S9. Reprendre la question S8 dans le cas où $a = 1.1$.

T10. Une épidémie surcritique ($m > 1$) est-elle aidée par le fait que la population soit finie ?

[Bonus] Expliquer pourquoi diviser une grande population en des petites sous-populations isolées peut constituer une politique de contrôle de l'épidémie intéressante. On illustrera la réponse en simulant, pour $a = 1.1$, 4 populations indépendantes partant chacune d'un seul individu infecté et $N = 2500$ individus sains et en traçant une réalisation de l'évolution de la somme des nombres d'individus infectés à chaque temps, $I_n^{(1)} + \ldots + I_n^{(4)}$, sur l'intervalle de temps $0 \leq n \leq 1000$ (soit $P = 1000$ dans les notations précédentes). On tracera sur le même graphe une réalisation de l'évolution du nombre d'individus infectés dans une seule population partant d'un seul individu infecté et $N = 10000$ individus sains.

PROJET 22

Un modèle d'arbres généalogiques

sujet proposé par A. Véber

amandine.veber@polytechnique.edu

Considérons une population de grande taille dans laquelle on a échantillonné n individus. On souhaite modéliser leur arbre généalogique, en supposant pour simplifier que chaque individu n'a qu'un seul parent à la génération précédente (on pensera par exemple à des bactéries ou à certains champignons). Pour ce faire, on fixe un paramètre $\psi \in]0,1]$ et on suppose que l'arbre se construit en remontant les générations de la manière suivante :

- on part de $X_0 = n$ lignées ancestrales;
- si le nombre X_k d'ancêtres de l'échantillon qui vivent à la génération $-k$ (dans le passé) est x, alors chacun d'entre eux lance une pièce qui tombe sur 'pile' avec probabilité ψ ou sur 'face' avec probabilité $1-\psi$, indépendamment les unes des autres. On décide alors que tous les ancêtres ayant tiré 'pile' ont le même parent à la génération $-(k+1)$, tandis que les autres ont tous des parents différents. Autrement dit, les lignées ancestrales de tous les individus de l'échantillon dont l'ancêtre à la génération $-k$ a tiré 'pile' *fusionnent* en une seule lignée à la génération $-(k+1)$, que l'on continue à suivre avec les lignées restantes;
- l'arbre est complet lorsqu'on a atteint un ancêtre commun à tout l'échantillon, i.e. lorsque X prend la valeur 1.

Un exemple est donné dans la Figure 22.1.

1 Étude de la suite $(X_k)_{k\geq 0}$

T1. Montrer que si X a pour valeur $x \in \{1,\dots,n\}$ à une génération $-k$ donnée, alors pour tout $i \in \{1,\dots,x-1\}$,

$$\mathbb{P}[X_{k+1} = x-i \,|\, X_k = x] = \binom{x}{i+1}\psi^{i+1}(1-\psi)^{x-i-1} := p_{x,x-i}. \tag{22.1}$$

Que vaut $\mathbb{P}[X_{k+1} = x \,|\, X_k = x]$?

Pour les questions S1 et S2, on prendra $n = 100$ et $\psi \in \{0.1, 0.5, 0.9\}$.

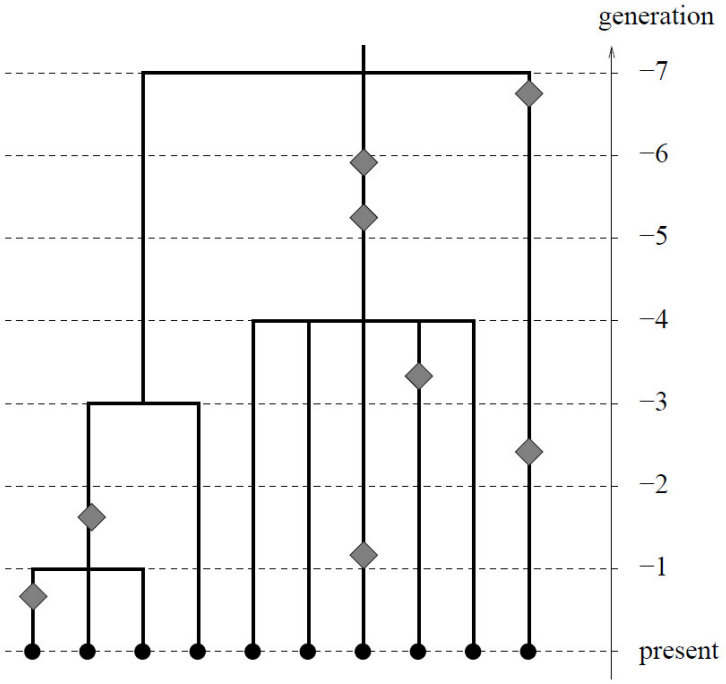


Figure 22.1: Une réalisation d'arbre avec $n = X_0 = 10$, $X_1 = X_2 = 8$, $X_3 = 7$, $X_4 = X_5 = X_6 = 3$ et enfin $X_7 = 1$. Les losanges gris représentent les mutations qui touchent les ancêtres de l'échantillon.

S1. Mettre en place un algorithme de simulation des trajectoires de X. Pour chacune des valeurs de ψ, tracer sur un même graphique 20 réalisations de cette trajectoire. Le nombre de générations à représenter sur ce graphique pourra dépendre de la valeur de ψ et on réfléchira à un choix judicieux.

T2. Notons T la première génération à laquelle X vaut 1. Montrer que T est fini avec probabilité 1. On pourra commencer par montrer que partant d'un nombre $m \leq n$ de lignées ancestrales, la première génération à laquelle X diminue d'au moins une unité, notée $\tilde{T}_m$, suit une loi géométrique d'un paramètre borné inférieurement en $m \in \{2, \dots, n\}$ par une constante strictement positive.

S2. En effectuant 1000 réalisations de trajectoires pour chaque jeu de paramètres, tracer l'histogramme des valeurs prises par T pour chacun des jeux de paramètres considérés.

S3. Donner une valeur approchée pour $\mathbb{E}[T]$ pour chacune des 3 valeurs de ψ.

On rappelle que si Z suit une loi binomiale de paramètres $N \in \mathbb{N}$ et $p \in [0, 1]$, alors

$$Z \overset{(loi)}{=} \sum_{i=1}^{N} B_i,$$

où les B_i sont des variables de Bernoulli indépendantes et de paramètre p.

T3. Quelle est la loi de X_1 en fonction de n? Vers quoi et en quel sens converge $n^{-1}X_1^{(n)}$ lorsque $n \to \infty$? (L'exposant de X_1 ne sert ici qu'à rappeler sa dépendance en n.)

T4. Vers quoi et en quel sens converge $(X_1^{(n)} - (1 - \psi)n)/\sqrt{n}$ lorsque $n \to \infty$?

T5. En décomposant le problème suivant les valeurs prises par $n^{-1}X_1^{(n)}$, montrer que $n^{-1}X_2^{(n)}$ converge en probabilité vers $(1-\psi)^2$, lorsque $n \to \infty$. On peut montrer (mais c'est un peu plus dur) que la convergence a également lieu presque-sûrement.

S4. Qu'est-ce que cela suggère sur le début de la suite $\left(n^{-1}X_k^{(n)}\right)_{k\geq 0}$ lorsque la taille n de l'échantillon est très grande? Vérifier cette intuition pour $\psi = 0.1$ et $\psi = 0.5$ en juxtaposant sur un même graphique la limite suggérée et des réalisations de la trajectoire de X pour $n = 10, 100, 1000, 10000$ (un graphique par valeur de ψ).

2 Longueur de l'arbre et mutations

On suppose à présent qu'on peut observer la séquence ADN des individus de l'échantillon à un gène (i.e., un emplacement du génome) donné. En réalité, c'est d'ailleurs la seule observable donnant de l'information sur les généalogies dans la plupart des populations réelles : en cueillant une poignée de fleurs (par exemple), il est a priori impossible de savoir quels sont leurs liens généalogiques rien qu'en les regardant... On suppose donc que des mutations apparaissent régulièrement sur les lignées ancestrales et qu'elles conduisent toujours à un *allèle* nouveau, de sorte que des individus de l'échantillon ne peuvent partager une mutation donnée que si elle est portée par leur plus récent ancêtre en commun.

Une manière classique de modéliser la manière dont les mutations apparaissent sur l'arbre consiste à supposer que celles-ci forment un *processus ponctuel de Poisson* d'intensité $\mu > 0$, c'est-à-dire que conditionnellement à la réalisation de l'arbre obtenue,

- si une branche de l'arbre a longueur ℓ, alors le nombre de mutations qu'elle porte suit une loi de Poisson de paramètre $\mu\ell$;
- les nombres de mutations portées par des branches distinctes sont des variables aléatoires indépendantes.

Les mutations forment donc un aléa supplémentaire qui vient se superposer à celui de l'arbre généalogique lui-même. Dans la suite, on fera l'hypothèse qu'on peut observer toutes les mutations qui tombent sur l'arbre. Ce sera le cas par exemple si le gène considéré est très long et que chaque mutation touche une paire de bases différente.

Pour rappeler sa dépendance en n (la taille de l'échantillon, qui est également la valeur de X_0), notons maintenant T_n la génération à laquelle X atteint la valeur 1, puis

$$L_n := \sum_{k=0}^{T_n-1} X_k$$

la longueur de l'arbre entier (i.e., la somme des longueurs de toutes ses branches) et enfin S_n le nombre total de mutations observées sur l'arbre, lorsque la taille de l'échantillon est n.

S5. Pour $n = 100$ et pour $\psi \in \{0.1, 0.5, 0.9\}$, donner l'histogramme des valeurs prises par L_n sur 1000 réalisations.

S6. Pour ces mêmes valeurs de paramètres, donner une approximation de $\mathbb{E}[L_n]$.

T6. Montrer que si $Z_1, \ldots, Z_k$ sont des variables aléatoires indépendantes suivant des lois de Poisson de paramètres respectifs $p_1, \ldots, p_k$, alors $Z_1 + \ldots + Z_k$ suit également une loi de Poisson de paramètre $p_1 + \ldots + p_k$.

On admettra que conditionnellement à L_n, S_n suit une loi de Poisson de paramètre μL_n (la question précédente le justifie partiellement).

T7. En écrivant que $\mathbb{E}[S_n] = \mathbb{E}[\mathbb{E}[S_n|L_n]]$, montrer que $\mathbb{E}[S_n] = \mu\mathbb{E}[L_n]$.

T8. On rappelle la notation $p_{x,y}$ introduite dans (22.1). Montrer que

$$\mathbb{E}[L_n] = n + \sum_{k=0}^{n-1} p_{n,n-k}\mathbb{E}[L_{n-k}],$$

où par convention on a posé $\mathbb{E}[L_1] = 0$ (pour rappel, l'évolution s'arrête lorsque X touche 1 et donc l'arbre partant de 1 individu est réduit à un point).

S7. En déduire un algorithme de calcul de $\mathbb{E}[L_n]$ et donner les valeurs de $\mathbb{E}[L_n]$ pour les mêmes jeux de paramètres qu'à la question S6. Comparer les résultats de la question S6 au calcul rigoureux de $\mathbb{E}[L_n]$ effectué dans cette question.

S8. Pour $\psi = 0.1$, jusqu'à quel ordre de grandeur des valeurs de n peut-on utiliser l'algorithme de la question S7 pour obtenir un résultat en moins de 1 min?

On s'intéresse maintenant aux mutations portées par un seul individu de l'échantillon. Celles-ci sont donc portées par les *branches externes* de l'arbre, c'est-à-dire les n branches débutant à chaque individu de l'échantillon et terminant à la première fusion à laquelle la lignée ancestrale de cet individu prend part. On note L_n^1 la longueur totale de ces branches externes et S_n^1 le nombre de mutations portées par un seul individu, lorsque l'échantillon est de taille n.

T9. En reprenant la méthode de la question T7 , montrer que $\mathbb{E}[S_n^1] = \mu\mathbb{E}[L_n^1]$.

Pour toutes les questions qui suivent, on prendra $n = 100$, $\psi \in \{0.1, 0.5\}$ et $\mu \in \{0.1, 1, 10\}$.

S9. Pour chaque valeur de ψ et μ ci-dessus, générer 1000 simulations conjointes de L_n^1 et de S_n^1 (c'est-à-dire, de couples (ℓ, s) tels que conditionnellement à ℓ, s est une réalisation d'une loi de Poisson de paramètre $\mu\ell$). Donner l'histogramme des valeurs prises par L_n^1 sur ces 1000 réalisations, ainsi que l'histogramme des valeurs prises par S_n^1.

S10. Donner une approximation de $\mathbb{E}[L_n^1]$ et de $\mathbb{E}[S_n^1]$. Calculer le ratio $\mathbb{E}[S_n^1]/\mathbb{E}[L_n^1]$ et le comparer à μ.

S11. A partir des couples de valeurs obtenues à la question S9, donner une approximation de $\mathbb{E}[S_n^1/L_n^1]$ pour chaque valeur du couple (ψ, μ) et la comparer à la valeur de μ.
