MAP361 *Aléatoire* – Projets de Simulation

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 un tiers dans la note des cours MAP 361, 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 (division par deux).

1 Calendrier

— Présentation de la liste des projets : le fascicule est disponible à partir du vendredi 26 avril.

— Choix des projets : les élèves devront exprimer leurs voeux à l'adresse

https://de.polytechnique.fr

du **vendredi 3 mai au lundi 13 mai, 23h59**. Dans les jours qui suivent, un algorithme répartira les sujets de façon automatique, en fonction des voeux émis par les élèves, et les élèves seront informés de leur sujet. Les élèves n'ayant pas informé la scolarité de leurs voeux dans les temps se verront affecter un sujet parmi ceux qui restent disponibles.

— Remise des rapports et programmes Python au plus tard le

lundi 1 juillet 2019 à midi.

Tout projet rendu **en retard** verra sa note réduite (-5). Vous pouvez ou bien

- envoyer un notebook jupyter, qui fait office de code+rapport, par mail à l'enseignant ayant proposé le sujet,
- envoyer un rapport et un code par mail à l'enseignant ayant proposé le sujet,
- déposer un rapport papier à la scolarité et envoyer un code par mail à l'enseignant ayant proposé le sujet.

2 Questions de Python/séances de soutien

Un cours en amphithéâtre et des séances de TP dédiées à Python sont prévus dans l'emploi du temps du tronc commun les 6, 9 et 15 mai.
De plus, une **permanence de soutien** pour ce travail sera organisée par Vincent Lemaire, Céline Bonnet et Juliette Chevallier les

— mercredi 12 juin, 18h à 20h, PC40

— lundi 17 juin, 18h à 20h, PC40

— lundi 24 juin, 18h à 20h, PC40

Enfin, en cas de difficulté proprement liée au sujet, le binôme devra s'adresser par mail à l'enseignant ayant proposé le sujet pour des explications complémentaires.

3 Rapport de projet

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.
Toute initiative personnelle est fortement encouragée. Les questions **T** sont théoriques, les questions **S** relèvent de la simulation.

Partie théorique 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 simulation doit contenir tous les codes (programmes `.py` ou notebook `jupyter`). 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 $n \gg 1$ se fera plutôt en commençant le programme par `n=1000` (ou une autre valeur) et en se référant à `n` tout au long du programme que en recopiant `1000` à chaque fois que ce paramètre apparaît dans le programme.

Table des matières

1 | Algorithme de Metropolis-Hastings

proposé par Giovanni Conforti, giovanni.conforti@polytechnique.edu

Dans ce projet, on se donne une loi de probabilité π, et on considère le problème de tirer des échantillons de cette loi. L'algorithme de Metropolis-Hastings est une des solutions possibles pour ce problème : il construit de façon récursive une suite de variables aléatoires $(X_n)_{n\in\mathbb{N}}$ dont la loi converge vers π quand $n \to +\infty$. En pratique, on va approximer π avec la loi de X_n. Il est donc très important de quantifier l'erreur qu'on fait avec cette approximation. On verra dans la suite qu'on peut contrôler cette erreur, et que la vitesse de convergence des X_n est exponentielle.

1 Hypothèses et description de l'algorithme

On considère un espace de probabilité E fini ou dénombrable et une loi π sur E qu'on veut échantilloner (loi cible). L'algorithme nécessite aussi d'une autre loi q, pour laquelle on est capable simuler des échantillons. On suppose que $\pi(x), q(x) > 0$ pour tout $x \in E$. On pose

$$\frac{1}{K} = \inf_{x\in E} \frac{q(x)}{\pi(x)},$$

et on suppose que $0 < K < +\infty$. Pour initialiser l'algorithme, on va se servir d'une autre loi de probabilité μ (on pourrait prendre $\mu = q$) telle que $\mu(x) > 0$ pour tout $x \in E$.

1.1 L'algorithme

On tire au hasard deux suites indépendantes $(Y_n)_{n\geq 0}$ i.i.d. de loi q, et $(U_n)_{n\geq 0}$ i.i.d. de loi uniforme sur $[0, 1]$. Soit X_0 une variable aléatoire de loi μ indépendante ces deux suites. On définit $(X_n)_{n\geq 0}$ par récurrence : pour tout $n \geq 0$,

$$X_{n+1} = \begin{cases} Y_{n+1} & \text{si } U_{n+1} < R(X_n, Y_{n+1}) \quad \text{où} \quad R(x,y) = \min\left(\frac{\pi(y)q(x)}{\pi(x)q(y)}, 1\right), \\ X_n & \text{sinon.} \end{cases}$$

La fonction $R(\cdot,\cdot)$ est dite *règle d'acceptation*. Elle nous dit si on accepte la proposition Y_{n+1} et on pose $X_{n+1} = Y_{n+1}$ ou si on la rejette : dans ce cas $X_{n+1} = X_n$.

2 Convergence exponentielle vers la loi cible

2.1 La dynamique de la suite $(X_n)_{n\in\mathbb{N}}$

On définit la *probabilité de transition* $P(x, y)$

$$\forall x, y \in E, \quad P(x,y) = \mathbb{P}(X_1 = y | X_0 = x).$$

Rappelons ici que $\mathbb{P}(X_1 = y|X_0 = x) = \frac{\mathbb{P}(X_1=y, X_0=x)}{\mathbb{P}(X_0=x)}$

T1. Vérifier que $P(x, y)$ est bien définie pour tout $x, y \in E$. Montrer que

$$P(x,y) = \begin{cases} q(y)R(x,y) & \text{si } x \neq y \\ 1 - \sum_{z\neq x} q(z)R(x,z) & \text{si } x = y \end{cases}$$

et en déduire que $P(x,y) > 0$ pour tout $x, y \in E$.

T2. Montrer que, pour tout $x, y \in E$ on

$$\pi(x)P(x,y) = \pi(y)P(y,x).$$

Cette proprieté est dite *reversibilité*.

T3. Montrer que pour tout $y \in E$

$$\mathbb{P}(X_1 = y) = \sum_{x\in E} \mu(x)P(x,y).$$

De façon génerale, montrer que

$$\mathbb{P}(X_{n+1} = y) = \sum_{x\in E} \mathbb{P}(X_n = x)P(x,y). \tag{1}$$

Pour cela, on peux commencer par la formule

$$\mathbb{P}(X_{n+1} = y) = \sum_{x\in E} \mathbb{P}(X_n = x)\mathbb{P}(X_{n+1} = y|X_n = x),$$

et se servir de l'indépendence entre (U_{n+1}, Y_{n+1}) et (U_n, Y_n).

Dans tout la suite, on note π_n la loi de X_n, $\pi_n(x) = \mathbb{P}(X_n = x)$. Il est important de remarquer que π_n dépend de q, mais aussi de la loi initiale μ.

T4. Se servir de la question précédente pour montrer que si la loi initiale μ est égale a π, alors $\pi_n = \pi$ pour tout $n \in \mathbb{N}$. En pratique, on peut pas prendre $\mu = \pi$, car on a besoin d'être capable de tirer selon π, ce qui est l'objectif final de l'algorithme.

T5. Montrer que

$$\forall x, y \in E, \quad P(x,y) - K^{-1}\pi(y) \geq 0. \tag{2}$$

La distance en variation totale

Pour deux lois de probabilité μ, ν sur E, leur distance en variation totale, est donnée par

$$d_{\mathrm{VT}}(\mu,\nu) = \sup_{A\subseteq E} |\mu(A) - \nu(A)|.$$

On a que d_{VT} satisfait les proprietés fondamentales d'une distance.

(i) $d_{\mathrm{VT}}(\mu,\nu) = 0 \Leftrightarrow \mu = \nu$

(ii) $d_{\mathrm{VT}}(\mu,\nu) = d_{\mathrm{VT}}(\nu,\mu) = 0$

(iii) Inégalité triangulaire

$$\forall \mu, \nu, \eta \quad d_{\mathrm{VT}}(\mu,\nu) \leq d_{\mathrm{VT}}(\mu,\eta) + d_{\mathrm{VT}}(\eta,\nu)$$

On a aussi la formule explicite, valable pour tout μ, ν

$$d_{\mathrm{VT}}(\mu,\nu) = \frac{1}{2}\sum_{x\in E} |\mu(x) - \nu(x)|.$$

2.2 Convergence exponentielle

T6. Se servir de la question précédente et de (1) pour montrer que

$$\forall n \geq 0, \quad d_{\mathrm{VT}}(\pi_{n+1},\pi) = \frac{1}{2}\sum_{y} \left| \sum_{x} (\pi_n(x) - \pi(x))(P(x,y) - K^{-1}\pi(y)) \right|.$$

T7. Se servir de la question précédente et de (2) pour montrer que

$$\forall n \in \mathbb{N}, \quad d_{\mathrm{VT}}(\pi_{n+1},\pi) \leq d_{\mathrm{VT}}(\pi_n,\pi)(1 - K^{-1}),$$

ce qui nous donne la convergence exponentielle de l'algorithme de Metropolis-Hastings.

3 Implémentation et simulation

S1. Implémenter l'algorithme de Metropolis-Hastings pour une loi cible π sur $\mathbb{N}$ donnée par

$$\forall m \in \mathbb{N}, \quad \pi(m) = \frac{1}{Z_\lambda} \frac{1}{\log(m+2)} \frac{\lambda^m}{m!},$$

où Z_λ est la constante de normalisation et $\lambda = 0.5, 1, 2$. Comme loi q, on prendra la loi de Poisson de paramètre λ, qu'on peut simuler en Python avec `numpy.random.poisson`. Est-ce-que on a besoin de connaître la valeur exacte de Z_λ pour implémenter l'algorithme ?
Pour $N = 5, 50, 10^2, 10^3$, tracer des histogrammes de π_N obtenus en lançant l'algorithme un nombre $L >> 1$ de fois et en utilisant l'estimateur $\pi_N^L(m)$ de $\pi_N(m)$ défini par

$$\pi_N^L(m) = \frac{1}{L} \sum_{l=1}^{L} \mathbf{1}_m(X_N^l),$$

où $(X_N^l)_{l=1,..,L}$ sont les valeurs restituées par chaque application de l'algorithme.

S2. Pour les mêmes choix de paramètres, calculer approximativement la valeur de K et vérifier empiriquement la convergence exponentielle en variation totale (indication : il faut évaluer numériquement la constante Z_λ). On pourra utiliser l'estimateur (tronqué) $d_{TV}^{20,L}$ pour la distance en variation totale

$$d_{\mathrm{VT}}^{20,L} = \frac{1}{2} \sum_{m=0}^{20} |\pi(m) - \pi_N^L(m)|$$

S3. (facultatif) Que peut-on dire lorsque la loi cible π est définie par

$$\forall m \in \mathbb{N}, \quad \pi(m) = \frac{1}{Z} |\sin(m)| \frac{1}{m!},$$

où Z est la constante de normalisation ? Faire des tests numériques, avec q la loi de Poisson de paramètre 1, pour voir si l'algorithme converge dans ce cas.

2 | Méthode de Stein et Théorème central limite

proposé par Giovanni Conforti, giovanni.conforti@polytechnique.edu

Un résultat classique de la théorie des Probabilités est le Theorème Central Limite (TCL).

Théorème. *Soit $(X_n)_{n\geq 0}$ une suite de variables aléatoires réelles indépendantes et de même loi, de carré intégrable, de moyenne m et de variance σ^2. Alors les variables*

$$S_n = \frac{(X_1 + \ldots + X_n) - nm}{\sigma\sqrt{n}}$$

convergent en loi vers une variable aléatoire de loi $\mathcal{N}(0,1)$.

Le but du projet est d'introduire la méthode de Stein, qui nous permet de donner des preuves alternatives de ce résultat et de *quantifier* la distance entre la loi normale et la loi de S_n. Il est important de noter que le TCL nous donne juste la convergence, mais ne dit rien sur la vitesse de convergence, ce qu'on peut trouver grâce à la méthode de Stein. Dans la dernière partie du projet on verra aussi comment la méthode de Stein nous permet de dépasser, dans certains cas, l' hypothèse d'indépendance entre les variables $(X_n)_{n\in\mathbb{N}}$.

1 Le cas des variables i.i.d.

On suppose que $(X_n)_{n\geq 0}$ est une suite de variables aléatoires réelles qui satisfont les hypothèses du TCL avec $m = 0$. On note γ la loi $\mathcal{N}(0,1)$,

$$\mathbb{E}_\gamma[f] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(z) \exp(-\frac{z^2}{2}) \mathrm{d}z.$$

1.1 La distance de Wasserstein d'ordre 1

Pour deux lois μ, ν sur $\mathbb{R}$ admettant un moment d'ordre 1, la distance de Wasserstein d'ordre 1 est

$$d_W(\mu,\nu) = \sup_{f\in\mathcal{F}} |\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]|, \quad \mathcal{F} = \{f : \mathbb{R} \to \mathbb{R} : \forall\, x, y \in \mathbb{R},\ |f(x) - f(y)| \leq |x - y|\}.$$

T1. Vérifier que d_W est une vraie distance, c'est a dire que

i) $d_W(\mu,\nu) = 0 \Leftrightarrow \mu = \nu$,

ii) $d_W(\mu,\nu) = 0 = d_W(\nu,\mu) = 0$,

iii) $d_W(\mu,\nu) \leq d_W(\mu,\eta) + d_W(\eta,\nu)$.

La distance de Wasserstein mesure l'écart maximal entre μ et ν sur la classe des fonctions 1-Lipschitz. On peut voir (on ne le fait pas ici) qu'on peut rémplacer $\mathcal{F}$ dans la définition de d_W par $\mathcal{F}' = \mathcal{F} \cap C^\infty(\mathbb{R};\mathbb{R})$.

1.2 Équation de Stein et simulations des solutions

La formule d'intégration par parties suivante est fondamentale dans la méthode.

T2. Vérifier que pour toute $f \in \mathcal{F}'$ on a

$$\mathbb{E}_\gamma[f'(X) - Xf(X)] = 0. \tag{1}$$

T3. (facultatif) Montrer l'implication inverse ; si (1) est valide pour toute $f \in \mathcal{F}'$ pour une loi γ, alors γ est la loi $\mathcal{N}(0,1)$.

Pour $f \in \mathcal{F}'$ donnée, on s'intéresse a l'équation de Stein ; on cherche $g_f : \mathbb{R} \to \mathbb{R}$ telle que

$$\forall z \in \mathbb{R}, \quad g_f'(z) - zg_f(z) = f(z) - \mathbb{E}_\gamma[f]. \tag{2}$$

Il se trouve qu'une solution g_f existe, et qu'on a une formule pour la calculer. De plus, on peut même borner g_f et ses deux premières derivées. Dans toute la suite on admettra le Lemme suivant, sans en faire la preuve.

Lemme. *Pour toute $f' \in \mathcal{F}$, la solution de l'équation (2) est donnée par*

$$g_f(z) = \exp(\frac{z^2}{2}) \int_z^{+\infty} \exp(-t^2/2)(\mathbb{E}_\gamma[f] - f(t))\mathrm{d}t$$

Aussi on a, uniformément en $f \in \mathcal{F}'$:

$$\sup_{z\in\mathbb{R}} |g_f(z)| \leq 2, \quad \sup_{z\in\mathbb{R}} |g_f'(z)| \leq \sqrt{\frac{2}{\pi}}, \quad \textit{and} \quad \sup_{z\in\mathbb{R}} |g_f''(z)| \leq 2.$$

S1. Écrire un code qui permet de calculer des solutions approchées de l'équation de Stein avec une methode de type Monte Carlo. Vérifier empiriquement la convergence de la solution approchée pour différentes fonctions f :

a) $f(z) = z$
b) $f(z) = \sin(z)$
c) $f(z) = \sqrt{1+z^2}$
d) $f(z) = \frac{1}{z^2+1}$
e) $f(z) = e^{-|z|}$

On va se servir du Lemme pour montrer ce Théorème d'approximation.

Théorème 1. *Soit $(X_n)_{n\in\mathbb{N}}$ une suite des variables indépendantes telles que $\mathbb{E}[X_i] = 0, \mathbb{E}[X_i^2] = 1$ pour tout i et*

$$C_3 = \sup_{i\in\mathbb{N}} \mathbb{E}[|X_i^3|] < +\infty, \quad C_4 = \sup_{i\in\mathbb{N}} \mathbb{E}[|X_i^4|] < +\infty.$$

Alors, on a

$$d_W(\gamma_n, \gamma) \leq \frac{1}{n^{1/2}} C_3 + \sqrt{\frac{2C_4}{\pi n}}, \tag{3}$$

où γ_n est la loi de $S_n = \frac{X_1+\ldots+X_n}{\sqrt{n}}$

En particulier, ce Théorème implique la convergence de γ_n vers γ a la vitesse $n^{-1/2}$ pour $n \to +\infty$.

T4. Se servir du point précédent pour montrer que si $f \in \mathcal{F}'$, alors,

$$\mathbb{E}[f(S_n)] - \mathbb{E}_\gamma[f] = \mathbb{E}[g_f'(S_n) - S_n g_f(S_n)]$$

et en déduire que

$$d_W(\gamma_n, \gamma) \leq \sup_{f\in\mathcal{F}'} |\mathbb{E}[g_f'(S_n) - S_n g_f(S_n)]|.$$

T5. Pour tout $i \leq n$ on note S_n^{-i} la variable $\frac{\sum_{j\neq i, 0\leq j\leq n} X_j}{\sqrt{n}}$. Montrer que pour tout $f \in \mathcal{F}'$

$$\mathbb{E}[g_f'(S_n) - S_n g_f(S_n)] = A + B$$

avec

$$A = \mathbb{E}[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i (g_f(S_n) - g_f(S_n^{-i}) - (S_n - S_n^{-i}) g_f'(S_n))]$$

et

$$B = \frac{1}{\sqrt{n}} \mathbb{E}[\sum_{i=1}^{n} (X_i (S_n - S_n^{-i}) g_f'(S_n))]$$

T6. Montrer que pour tout $f \in \mathcal{F}'$, $n \in \mathbb{N}$ on a

$$A \leq \frac{1}{n^{1/2}} C_3$$

(*Suggestion :* Utiliser la formule de Taylor et le Lemme, puis l'indépendance entre les X_i)

T7. Montrer que pour tout $f \in \mathcal{F}'$,$n \in \mathbb{N}$

$$B \leq \frac{\sqrt{2}}{\sqrt{\pi n}} C_4$$

(*Suggestion :* Utiliser le Lemme et l'inegalité de Cauchy Schwartz) Déduire la preuve du Théorème 1.2.

On verifiera les résultats théoriques avec des simulations.

S2. Pour les mêmes fonctions f qu'avant et en partant de different choix de lois pour la suite $(X_n)_{n\in\mathbb{N}}$ (pour exemple, uniformes, Bernoulli, Poisson..) [1] et pour $n = 1, ..., 100$, répéter les passages suivantes

(i) Calculer $\mathbb{E}_\gamma[f]$ avec une methóde Monte Carlo

(ii) Calculer $\mathbb{E}_{\gamma_n}[f]$ avec une methóde Monte Carlo

(iii) Comparer les deux quantités et verifiér que la borne (3) est (presque) satisfaite.

Pour chaque f et chaque loi choisies, illustrer graphiquement les résultats. Commenter. Pour quelles fonctions les bornes semblent plus précises ?

2 Sommes des variables localement dépendantes (Facultatif)

On se met ici dans le cadre où les variables $(X_n)_{n\in\mathbb{N}}$ ne sont plus indépendantes, mais "pas trop dépendantes". Plus précisement on dira que le voisinage de la variable X_i est l'ensemble $N_i \subseteq \mathbb{N}$ défini par [2]

$$N_i = \{j \in \mathbb{N} : X_j \text{ est indépendante de } X_i\}^C.$$

Si la taille du voisinage est uniformément bornée on peut généraliser le Théorème 1.

Théorème. *Soit $(X_n, n \in \mathbb{N})$ une suite de variables aléatoires telles que $\mathbb{E}[X_i] = m$ pour tout i et*

$$C_3 = \sup_{i\in\mathbb{N}} \mathbb{E}[|X_i^3|] < +\infty, \quad C_4 = \sup_{i\in\mathbb{N}} \mathbb{E}[|X_i^4|] < +\infty, \quad D = \sup_{i\in\mathbb{N}} |N_i| < +\infty$$

Si on note $\sigma_n^2 = \mathbb{E}[(\sum_{i=1}^n X_i - m)^2]$, alors on a

$$d_W(\gamma_n, \gamma) \leq \frac{nD^2}{\sigma_n^3} C_3 + \frac{\sqrt{28 C_4 n}\, D^{3/2}}{\sqrt{\pi}\sigma_n^2} C_4, \tag{4}$$

où γ_n est la loi de $\frac{X_1+...,X_n-mn}{\sigma_n}$.

1. Il faut bien s'assurer che $\mathbb{E}[X_i] = 0$ et $\mathbb{E}[|X_i|^2] = 1$. Donc, si on fait le choix d'une loi de Poisson, il faudra utiliser de transformations affines pour se reconduire a ce cas.

2. Pour un ensemble $A \subseteq \mathbb{N}$, on note A^C son complémentaire

On ne va pas faire la preuve de ce Theoréme, mais on se limitera a le vérifier empiriquement.

S3. (facultatif) Soit $(Y_n)_{n\in\mathbb{N}}, (Z_n)_{n\in\mathbb{N}}$ deux suites i.i.d. telles que $Y_0 \sim \text{Be}(p), Z_0 \sim \text{Be}(q)$. On va construire la suite $(X_n)_{n\geq 0}$

$$\begin{cases} X_0 = Y_0, \\ X_n = Z_n Y_n + (1 - Z_n)Y_{n+1}, \qquad n \geq 1. \end{cases}$$

Montrer que

(a) $X_n \sim \text{Be}(p)$ pour tout $n \geq 0$.

(b) X_j est indépendante de X_i si et seulement si $|i-j| > 1$. En déduire que $|N_i| = 3$ pour tout $i \in \mathbb{N}$

(c) $\text{Cov}(X_n, X_{n+1}) = (1-q)qp^2$. Calculer σ_n dans ce cas.

Dans la suite, on verifiéra empiriquement la borne (4) pour la suite $(X_n)_{m\in\mathbb{N}}$ qu'on vient de construire. Pour cela, il faut calculer les constantes C_3, C_4.

S4. (facultatif) Pour les mêmes fonctions f qu'avant et pour $n = 1, ..., 100$, répéter les passages suivantes

(i) Calculer $\mathbb{E}_\gamma[f]$ avec une methode Monte Carlo

(ii) Calculer $\mathbb{E}_{\gamma_n}[f]$ avec une methode Monte Carlo. Pour faire ça , il faut d'abord écrire un code qui, pour une valeur de n fixée, nous permet d' échantilloner la loi du vecteur aléatoire $(X_1, \ldots, X_n)$.

(iii) Comparer les deux quantités et vérifier que la borne (4) est (presque) satisfaite.

Illustrer graphiquement les résultats obtenus.

3 | Désintégrations radioactives

proposé par Aymeric Dieuleveut, aymeric.dieuleveut@gmail.com

Les noyaux d'isotopes radioactifs se désintègrent naturellement au cours du temps. La période de demie-vie (ou période radioactive) est le temps nécessaire pour que la moitié des noyaux d'une source se soient désintégrés. On souhaite estimer la période de demie–vie d'un atome radioactif en observant un grand nombre d'atomes identiques. On présente et compare plusieurs méthodes.

1 Modélisation

On suppose que les durées de vie des atomes sont des variables aléatoires $T_1, \dots, T_n$ indépendantes et de même loi.
On suppose également que la désintégration est un processus sans mémoire : sachant qu'un atome ne s'est pas désintégré au bout d'un temps t, la loi de sa durée de vie restante $T - t$ est la même que la loi de sa durée de vie T. Formellement, pour tout $s, t \geq 0$:

$$\mathbb{P}(T > t + s | T > t) = \mathbb{P}(T > s). \tag{1}$$

On suppose que T est à valeurs dans $\mathbb{R}_+$ et sans atome.

T1. On définit $G(t) = \mathbb{P}(T > t)$. Montrer que G satisfait l'équation $G(t + s) = G(t)G(s)$. En déduire une expression de G et montrer que la loi de T est une loi exponentielle de paramètre θ (de densité $f_\theta(x) = \theta e^{-\theta x}$ sur $\mathbb{R}_+$).

T2. À quelle quantité reliée à T la durée de demie-vie correspond-elle ? Calculer la fonction de répartition de T. Montrer que la durée de vie moyenne $\mathbb{E}[T]$, la durée de demie -vie, et le taux de désintégration (défini par $\lim_{h\to 0}(\mathbb{P}(T < h)/h)$), sont tous trois égaux.

2 Méthode du maximum de vraisemblance

On observe des durées de vie $T_1, \dots, T_n$.

T3. Écrire le modèle statistique engendré par l'observation de $T_1, \dots, T_n$. Calculer l'estimateur du maximum de vraisemblance $\hat{\theta}_{1,MV}$ pour θ. Cet estimateur fournit-il un estimateur sans biais de θ ? Montrer que $\hat{\theta}_{1,MV}$ est asymptotiquement normal et calculer sa variance limite. Proposer un intervalle de confiance pour le paramètre θ.

S1. Proposer une méthode pour simuler un échantillon de variables aléatoires exponentielles de paramètre θ. Pour $n = 100$, calculer l'estimateur du maximum de vraisemblance. En effectuant cette expérience de façon répétée, illustrer la normalité asymptotique de l'estimateur.

T4. Montrer que $2\theta(T_1 + \dots + T_n)$ suit une loi du χ^2 à $2n$ degrés de liberté. En déduire un intervalle de confiance non asymptotique de niveau $1 - \alpha$ pour θ.

S2. Représenter la fonction de répartition empirique d'un échantillon de même loi que $2\theta(T_1 + \dots + T_n)$. Quelle théorème illustre-t-on ?

Afin de mettre en pratique une telle méthode, on a besoin d'attendre que tous les atomes soient désintégrés. On dénote $M_n = \max(T_1, \dots, T_n)$ ce temps d'attente.

T5. Calculer la fonction de répartition de M_n. Montrer que $\theta M_n - \ln(n)$ converge en loi, quand $n \to \infty$, vers une loi de *Gumbel* (une loi de densité $x \mapsto e^{e^{-x}}$).
Montrer que $\theta M_n / \ln(n)$ converge vers 1 en probabilité, puis presque surement.

S3. Pour $n \in \{10, 100, 1000\}$, représenter un histogramme empirique d'un échantillon de variables aléatoires de loi M_n (renormalisée ou non).
Quelles sont les limites de cette première méthode ? En pratique, cette méthode vous semble-t-elle efficace ?

3 Utilisation des statistiques d'ordre.

Dans cette partie, on se propose de choisir d'utiliser seulement une partie des observations (les d premières, avec $d \leq n$).
On dénote $T_{(1)}, \dots, T_{(n)}$ les statistiques d'ordre de $T_1, \dots, T_n$, c'est à dire les variables aléatoires telles que

$$\{T_{(1)}, \dots, T_{(n)}\} = \{T_1, \dots, T_n\}, \quad \text{et } T_{(1)} \leq T_{(2)} \leq \dots \leq T_{(n)}.$$

T6. Déterminer la densité du $n-$uplet $T_{(1)}, \dots, T_{(n)}$. Montrer que le $n-$uplet

$$\left(nT_{(1)}, (n-1)(T_{(2)} - T_{(1)}), \dots, 2(T_{(n-1)} - T_{(n-2)}), T_{(n)} - T_{(n-1)} \right) \tag{2}$$

est un échantillon de loi exponentielle de paramètre θ. Quelle est l'interprétation de ce résultat ? (on expliquera pourquoi le résultat aurait pu être anticipé sans aucun calcul).
A partir de l'observation des d premières décompositions, quel est l'estimateur du maximum de vraisemblance $\hat{\theta}_{2,d}$ pour θ ?

S4. Générer un échantillon de taille $n = 1000$. Représenter l'estimateur du maximum de vraisemblance en fonction de d (on utilisera aussi comme axe des abscisses les temps $T_{(d)}$). En répétant l'expérience 100 fois, illustrer l'évolution de l'estimateur et sa variance en fonction de d (on pourra représenter les résultats sous forme de boites a moustaches, pour une grille de valeurs de d).

S5. Calculer une estimation de l'erreur quadratique moyenne ($\mathbb{E}[|\hat{\theta} - \theta|^2]$) de l'estimateur et représenter son évolution en fonction du temps. Quelle valeur de d vous semble la plus pertinente ?

S6. Comparer la variance de l'estimateur obtenu dans la première partie et de l'estimateur obtenu avec $d = n/2$. Quel temps moyen faut-il attendre pour pouvoir calculer chacun de ces deux estimateurs ?

4 Observation de durée fixe

On se propose à présent de choisir a priori une durée t et d'observer les désintégrations jusqu'à ce temps. On considère $t \in \mathbb{R}_+$ dans cette partie.

T7. Quelle est la loi du nombre d'atomes N_t encore présents à l'instant t ? En déduire un estimateur sans biais de $e^{-\theta t}$, et un estimateur $\hat{\theta}_{3,t}$ de θ.

T8. Donner un intervalle de confiance pour $e^{-\theta t}$. Comment a t'on intérêt à choisir t dans ce cadre ?

S7. Un tel estimateur résulte-il en un estimateur sans biais de θ ? Calculer une estimation de l'erreur quadratique moyenne ($\mathbb{E}[|\hat{\theta} - \theta|^2]$) de l'estimateur et représenter son évolution en fonction du temps. Que pensez vous du comportement d'un tel estimateur ? Quelle information n'a pas été utilisée ?

On observe également les instants de désintégration $T_{(1)}, \dots, T_{(D_t)}$, avec $D_t = N - N_t$.
Pour $d \in \{0, \dots, n\}$, on note $S_d := \{(t_1, \dots, t_d) \in \mathbb{R}^d : 0 \leq t_1 \leq \dots \leq t_d \leq t\}$. On considère λ_d la mesure de Lebesgue sur S_d. L'ensemble S_0 est par convention un singleton, et λ_0 un Dirac en ce

singleton. La variable aléatoire $(T_{(1)}, \dots, T_{(D_t)})$ est à valeurs dans $\bigcup_{d=0}^{n} S_d$.

T9. Déterminer sa densité par rapport à la mesure $\lambda_0 + \cdots + \lambda_n$ (pour un d dans $\{0, \dots, n\}$ et pour B un borélien de S_d, on calculera d'abord $\mathbb{P}(T_{(1)}, \dots, T_{(D_t)} \in B)$). Montrer que l'estimateur du maximum de vraisemblance vérifie alors :

$$\frac{1}{\hat{\theta}_{4,t}} = \frac{1}{D_t}(T_{(1)} + \cdots + T_{(D_t)} + (n - D_t)t) = \frac{1}{D_t}\sum_{k=1}^{n} \min(T_k, t). \tag{3}$$

Quelle est la probabilité que D_t soit nul ? Comment a t'on intérêt à choisir t ?

S8. Simuler le nouvel estimateur proposé. Représenter graphiquement les estimations de $1/\theta$ et θ. Estimer le risque quadratique de cet estimateur en fonction de t.

S9. Dans le cadre d'une procédure de datation, on cherche à estimer le temps t. On connait la proportion initiale d'atomes radioactifs de carbone ^{14}C, $(1,2 \times 10^{-12})$. On compte initialement 1000 atomes de carbone, représentant une proportion de 3×10^{-13}.
A partir des observations des temps de désintégration "futurs", à télécharger (Fichier Datation_obs, https://www.dropbox.com/sh/sqtwnonv0pg1s69/AACv1_Xtf0sM9Gv-TLUbRNySa?dl=0), proposer un estimateur du temps t et estimer une datation de l'objet considéré.
Reproduire cette situation, et proposer un intervalle de confiance empirique pour t.

5 Comparaison des estimateurs

S10. Proposer une illustration des résultats précédents permettant de comparer les différents estimateurs.

4 | Modèle de ségrégation de Schelling

proposé par Aymeric Dieuleveut, aymeric.dieuleveut@gmail.com

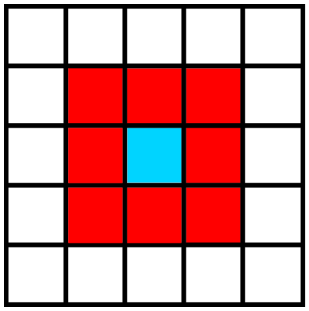
On considère dans ce problème deux modèles, qui pourront être abordés indépendamment. La seconde partie est facultative.
Pour tout $u \in \mathbb{N}$, on utilise la notation $[u] = \{1, \dots, u\}$.

1 Modèle de ségrégation de Schelling

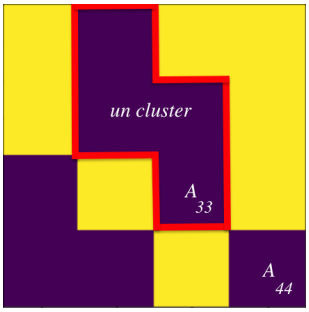
Introduction : Le modèle de ségrégation de Schelling est un des premiers modèles de sciences sociales à illustrer comment les comportements et décisions individuelles d'agents peuvent aboutir collectivement à une ségrégation.
Thomas Schelling est un économiste américain. Il a reçu le prix Nobel d'économie en 2005, pour avoir : « amélioré notre compréhension des mécanismes de conflit et de coopération par l'analyse de la théorie des jeux.»
Soit $M \geq 1$. On considère un "échiquier" de taille $M \times M$. Sur certaines cases (ou *sites*), qu'on appelle *occupées*, se trouve un agent. Un agent peut être de deux couleurs différentes. Les couleurs définissent deux groupes, qu'on note G_{-1}, G_1. Une partie des sites sont *libres*. Les agents peuvent se déplacer sur ces sites. Pour tout site $(i, j) \in M \times M$, on note $A_{i,j}$ son état : $A_{i,j} = k$ si l'agent appartient au groupe $k \in \{-1, 1\}$, et $A_{i,j} = 0$ si la case est vide.

Thomas Schelling

On définit les cases *voisines* d'une case comme ses 8 cases adjacentes, voir Figure 4.1a.
On considère une proportion $L \in [0; 1]$, qu'on appelle la *tolérance*. Si la fraction des voisins d'un agent qui appartiennent à son groupe est inférieure à L, l'agent est *insatisfait*. Sa *satisfaction* est 0. Inversement, si la fraction de ses voisins qui appartiennent à son groupe est supérieure à L, l'agent est *satisfait*. Sa *satisfaction* est 1.



(a) Les voisins de la case bleue sont les 8 cases rouges.



(b) Exemple de la définition d'un cluster. En particulier A_{33} et A_{44} ne sont pas voisins au sens des clusters, mais sont voisins pour la définition de la satisfaction/tolérance.

T1. On note $V_{i,j}$ les voisins de la case i, j. On note $p_{i,j}$ la proportion des voisins du même groupe que $A_{i,j}$. Donner une expression de $p_{i,j}$.

On part d'une configuration initiale aléatoire, avec une proportion de cases vides qu'on note ρ : pour chaque case, on tire indépendamment sa configuration, de sorte que, pour tout site $(i, j) \in M \times M$,

$$\mathbb{P}(A_{i,j} = 0) = \rho$$
$$\mathbb{P}(A_{i,j} = +1) = \mathbb{P}(A_{i,j} = -1) = \tfrac{1-\rho}{2}.$$

T2. Quelle est la loi du nombre total de cases occupées N_{tot} ? Comment se comporte ce nombre asymptotiquement quand $M \to \infty$.

A chaque instant t, on tire un agent au hasard, satisfait ou insatisfait et il se déplace **sur une case satisfaisante** (uniformément parmi les cases sur lesquelles il sera satisfait). On répète cette dynamique jusqu'à ce que tous les agents soient satisfaits, ou jusqu'à ce que la proportion d'agents satisfaits n'évolue plus, et on s'intéresse à l'évolution de l'état de la grille.

T3. Justifier que seules certaines valeurs de L sont significatives. Quelles sont ces valeurs ?

Dans les simulations, on définit un *pas de temps* comme le temps au bout duquel **chaque agent** a en moyenne bougé (ou essayé de bouger) une fois.

1.1 Simulations

S1. Générer la configuration initiale, pour une grille de taille $M = 100$, et $L = 0.5, \rho = 5\%$, et afficher cet état initial.
(Indication : on représentera la grille par une matrice de taille $M \times M$, à coefficients dans $\{-1, 0, 1\}$. On pourra utiliser la fonction `plt.imshow(X)` en python pour afficher la matrice.)

S2. Coder l'algorithme décrit ci-dessus, pour des paramètres ρ, L.

S3. Pour une grille de taille $M = 100$, et $\rho = 5\%$, représenter l'état de la grille après $T = \{1, 10, 100, 400\}$ pas de temps (remarquer que cela représente environ $(1 - \rho)M^2$ fois plus d'itérations de l'algorithme, voir la définition d'un pas de temps ci-dessus).

Dans la suite, on se place dans la configuration $M = 50$.

S4. Représenter l'état final ($T = 400$), pour les 20 configurations possibles de $\rho \in \{2\%, 6\%, 12\%, 18\%\}$ et $L \in \{1/5;\ 2/7;\ 1/3;\ 5/7;\ 3/4\}$.
Commenter en détail les résultats. En particulier, comment évolue l'aspect de l'état final en fonction de L ?

1.2 Une mesure de ségrégation

On définit un cluster comme un groupe connexe d'agents d'un même groupe. La définition formelle de connexité est donnée ci-après : les points ne sont "connectés" qu'à leurs 4 plus proches voisins, voir la Figure 4.1b pour un exemple.
Formellement, on dit que le groupe est connexe si quels que soient deux points $(i_0, j_0), (i_T, j_T)$ du groupe, il existe un chemin entre ces deux points, c'est-à-dire une séquence $(i_k, j_k)_{k=0,\dots,T}$, telle que pour tout $k \in \{0, \dots, T-1\}$, $A_{i_k,j_k} = A_{i_0,j_0}$ (même couleur), et $|i_k - i_{k+1}| + |j_k - j_{k+1}| = 1$ (4-voisins).
On note C l'ensemble des clusters. Pour tout cluster $c \in C$, on définit la masse n_c d'un cluster comme le nombre de points qu'il contient. On a bien sûr que $\sum_{c \in C} n_c = N_{\text{tot}}$ le nombre total de case occupées.
On définit le coefficient de ségrégation $S \in [0; 1]$ comme la moyenne pondérée des poids des clusters, renormalisée. Pour une configuration donnée, avec $p_c = n_c / N_{\text{tot}}$, soit

$$S = \frac{2}{N_{\text{tot}}} \sum_{c \in C} n_c p_c = \frac{2}{N_{\text{tot}}^2} \sum_{c \in C} n_c^2. \tag{1}$$

T4. Justifier le choix de la renormalisation.

On s'intéresse à la valeur asymptotique (quand $T \to \infty$) moyenne (on effectuera $n = 20$ tentatives indépendantes pour chaque configuration de ρ, M, L afin d'obtenir une valeur moyennée de s et une

estimation de la variation. On représentera en général les résultats avec des diagrammes en boite).

S5. Pour une configuration donnée à un temps t, coder un algorithme qui permet de déterminer les clusters, et représenter le résultat sur une grille de taille $M = 10$. On pourra utiliser et adapter du code disponible [1].

S6. Représenter l'évolution du coefficient de ségrégation en fonction de $T = \{1, 10, 100, 400\}$ pour une grille de taille $M = 50$, et $L = 0.5, \rho = 5\%$.

S7. On considère $M = 50$. Pour chaque valeur de $\rho \in \{2\%, 6\%, 12\%, 18\%, 25\%\}$, représenter le coefficient de ségrégation moyenné final ($T = 400$) en fonction de L (on considèrera une grille de 16 valeurs pour L : $L \in \{0.2 + k/20, k \in [20]\}$). Représenter également la satisfaction moyenne des agents à l'état final.

Commenter l'ensemble des résultats. En particulier, montrer qu'il existe une ou deux valeurs critiques de L auxquelles le coefficient de ségrégation varie rapidement. On appelle ce phénomène une *transition de phase*.

1.3 Variante

Dans cette variante, seuls les agents **insatisfaits** se déplacent (toujours sur une case libre ou ils se trouveraient satisfaits). Plus précisément, on considère l'ensemble des agents du groupe $k \in \{-1, 1\}$ qui sont insatisfaits, et l'ensemble des cases satisfaisantes pour les agents du groupe k. A chaque instant t, on tire uniformément un agent parmi les agents insatisfaits, et cet agent se déplace sur une case aléatoire parmi les cases satisfaisantes. Si aucune telle case n'existe, il se déplace sur une case aléatoire. On définit l'énergie du modèle par :

$$E_S = -\sum_{i,j\in[M]\times[M]} \sum_{(k,l)\in V_{i,j}} A_{ij}A_{kl} - K \sum_{i,j\in[M]\times[M]} \sum_{(k,l)\in V_{i,j}} A_{ij}^2 A_{kl}^2 \tag{2}$$

T5. Montrer que dans la variante ci-dessus, l'énergie du modèle est décroissante avec le temps, pour un choix adapté de K. A quoi K correspond-il ?

S8. Représenter l'évolution de l'énergie du modèle en fonction du temps, dans le cas général et dans la variante. Commenter.

Le modèle de ségrégation de Schelling décrit ci-dessus peut être comparé à un modèle de Blume-Emery-Griffiths, qu'on étudiera ci-dessous, dans lequel on contraindrait la proportion d'agents de chaque type à être fixe.

2 Facultatif : Modèle de champ magnétique

On considère une grille de taille $M \times M$, indexée par $(i, j) \in [M] \times [M]$. En chacun des points de ce réseau, on considère une variable aléatoire $\sigma_{i,j} \in \{-1, 1\}$, qu'on appelle "spin".

2.1 Modèle d'Ising

L'énergie du modèle (dans le cas du champ externe nul) dans une configuration $\sigma = (\sigma_{i,j})_{i,j}$, est donnée par la fonction suivante :

$$H(\sigma) = -J \sum_{\langle i\ j\rangle} \sigma_i \sigma_j \tag{3}$$

qu'on appelle le Hamiltonien du système. J est une constante. La notation de somme sur $\langle i\ j\rangle$ correspond à sommer sur toutes les paires de sites adjacents (deux sites $(i, j), (i', j')$ sont adjacents si $|i - i'| + |j - j'| = 1$, donc au sens de la Figure 4.1b).

1. https://github.com/6arlos6/Percolation-Theory---Adaptation-of-Hoshen-Kopelman-algorithm-for-cluster-labeling-

La probabilité d'une configuration suit une distribution de Boltzmann $\mathbb{B}$ avec une température inverse β :

$$\mathbb{B}(\sigma) = \frac{e^{-\beta H(\sigma)}}{\sum_\sigma e^{-\beta H(\sigma)}}, \tag{4}$$

où le dénominateur correspond à la renormalisation nécessaire pour avoir une distribution de probabilité.

T6. Combien de termes comporte la somme sur l'ensemble des configurations possibles ? Vous semble-t-il possible de calculer la constante de renormalisation ?

Pour estimer le comportement du système sous cette distribution, on utilise l'algorithme de Métropolis Hasting.

Algorithme de Métropolis

Introduction. [Source : wikipedia] "The Metropolis-Hastings algorithm can draw samples from any probability distribution $P(x)$, provided that the value of a function $f(x)$ proportional to the density of P can be calculated. The requirement that $f(x)$ must only be proportional to the density, rather than exactly equal to it, makes the Metropolis-Hastings algorithm particularly useful, because calculating the necessary normalization factor is often extremely difficult in practice.
The Metropolis-Hastings algorithm works by generating a sequence of sample values in such a way that, as more and more sample values are produced, the distribution of values more closely approximates the desired distribution $P(x)$. These sample values are produced iteratively, with the distribution of the next sample being dependent only on the current sample value (thus making the sequence of samples into a Markov chain). Specifically, at each iteration, the algorithm picks a candidate for the next sample value based on the current sample value. Then, with some probability, the candidate is either accepted (in which case the candidate value is used in the next iteration) or rejected (in which case the candidate value is discarded, and current value is reused in the next iteration) –the probability of acceptance is determined by comparing the values of the function $f(x)$ of the current and candidate sample values with respect to the desired distribution $P(x)$."

En général, l'algorithme fonctionne comme suit : si on connait une fonction f proportionnelle à la densité, à partir d'un état initial aléatoire, on construit itérativement des configurations comme suit :

1. On tire un candidat x' selon une distribution $g(x'|x_t)$ appelée probabilité de transition.
2. On calcule le ratio d'acceptation $\alpha = \frac{f(x')}{f(x_t)} = \frac{P(x')}{P(x_t)}$ (remarquer qu'on n'a pas besoin de calculer la constante de renormalisation)
3. on accepte le candidat avec probabilité $\min(\alpha, 1)$: pour cela, on tire $u \sim \mathcal{U}[0;1]$ et :
   — si $u \leq \alpha$, on accepte le candidat : $x_{t+1} = x'$
   — si $u > \alpha$, on rejette le candidat : $x_{t+1} = x_t$.

Simulation pour le modèle d'Ising Dans le cadre du modèle d'Ising, l'algorithme est ainsi spécifié :

1. On considère un état de départ aléatoire dans lequel les spins sont tirés de façon i.i.d., de sorte que $\mathbb{P}(\sigma_{i,j} = \pm 1) = 1/2$ pour tout i, j.
2. À partir d'une configuration $\sigma = (\sigma_{i,j})_{i,j}$ donnée, on considère une probabilité de transition uniforme sur l'ensemble des M^2 configurations σ' que l'on peut obtenir en inversant **un unique spin**. Formellement

$$g(\sigma'|\sigma) = \frac{1}{M^2} \mathbb{1}\{\exists (k,l) \quad \sigma'_{(k,l)} = -\sigma_{i,j} \text{ et } \forall (i,j)/(i,j) \neq (k,l), \sigma'_{(k,l)} = \sigma_{i,j}\}. \tag{5}$$

3. A chaque instant $t \geq 0$, on dispose d'une configuration σ^t. On sélectionne une configuration candidate σ' en inversant 1 spin. On calcule l'énergie des deux configurations $H(\sigma^t)$, $H(\sigma')$, et on accepte la nouvelle configuration avec probabilité :

$$\begin{cases} 1 & \text{si } H(\sigma^t) > H(\sigma') \\ e^{-\beta(H(\sigma') - H(\sigma_t))} & \text{sinon.} \end{cases} \tag{6}$$

S9. Simuler l'algorithme de Métropolis Hasting à partir d'une configuration de départ, pour $t = T \times M^2$ étapes.
On considèrera $T = 100, M = 40, J = 1, \beta = 1$. Représenter l'état final.
Faire de même pour $T = 100, M = 40, J = 1, \beta = 1/4$.

Remarque : pour la sélection du candidat, plutôt que de tirer à chaque temps $t \in [TM^2]$ une position i, j uniformément sur $[M] \times [M]$ où inverser le spin, on pourra tirer à chaque temps $t = (k-1)M^2 + 1$, $k \in [T]$, une permutation aléatoire de $[M] \times [M]$ et utiliser les indices de cette permutation dans l'ordre entre $t = (k-1)M^2 + 1$ et $t = kM^2$.

On s'intéresse au champ magnétique moyen Σ à l'état final (i.e., à la configuration σ^T) :

$$\Sigma(\sigma^T) = \left| \frac{\sum_{i,j \in [M] \times [M]} \sigma^T_{i,j}}{M^2} \right|. \tag{7}$$

T7. Justifier que Σ^T est une variable aléatoire. De quelles variables aléatoires dépend Σ^T ? En réalité, on veut estimer $\mathbb{E}_{\sigma \sim \mathbb{B}}[\Sigma(\sigma)]$. Proposer une méthode pour estimer cette quantité.

S10. On fixe $T = 100, M = 40, J = 1$. On définit $T = 1/\beta$. Représenter l'évolution du champ magnétique moyen sous la distribution de Boltzmann, en fonction de la température, pour $T \in \{0.1 + 0.1k, \quad k \in [50]\}$.
Montrer qu'il existe une transition de phase à une température critique de $T_c \simeq 2,3J$.

Remarque. On peut démontrer un tel résultat : dans un modèle de dimension 2, la température critique est exactement $T_c = \frac{2J}{\ln(1+\sqrt{2})k_B}$ (k_B est une constante physique, qu'on n'a pas mentionnée précédemment pour alléger les notations).

Conclusion : Modèle de Blume-Emery-Griffiths Le modèle de Blume-Emery-Griffiths est une variance du modèle d'Ising, qui permet de modéliser l'état d'un mélange d'atomes d'Hélium, qui peuvent exister sous deux formes distinctes, ^{3}He (deux protons et un neutron) et ^{4}He (deux protons et deux neutrons).
Sur une grille de taille $[M] \times [M]$, avec un $\sigma_{i,j}$ en chaque site : $\sigma_{i,j}$ peut prendre les valeurs $\{-1, 0, 1\}$. Les atomes ^{3}He correspondent aux sites où $\sigma_{i,j} = 0$, et les atomes ^{4}He correspondent aux sites où $\sigma_{i,j} = \pm 1$.

L'énergie du système est décrite par le Hamiltonien suivant :

$$H_{BEG}(\sigma) = -J \sum_{\langle i \ j \rangle} \sigma_i \sigma_j + D \sum_{i,j \in [M] \times [M]} \sigma_i^2 + \mathcal{K} \sum_{\langle i \ j \rangle} \sigma_i^2 \sigma_j^2. \tag{8}$$

On remarque que ce modèle permet de faire le lien entre le modèle d'Ising et le modèle de Schelling. Pour $D = \mathcal{K} = 0$, on retrouve le modèle d'Ising, alors que pour une concentration constante en atomes de ^{3}He, (i.e. $\sum_{i,j \in [M] \times [M]} \sigma_i^2$ est constante), on retrouve l' "énergie" définie à l'équation (2).
On pourrait établir un lien entre la température T (ou $\beta = T^{-1}$) du modèle d'Ising et le "bruit" ajouté par le mouvement des agents **satisfaits** dans la partie 1.2 (comparer avec la partie 1.3).

5 | Ruine et casinos

proposé par Aymeric Dieuleveut, aymeric.dieuleveut@gmail.com

On s'intéresse à l'évolution de la fortune d'un gérant de casino. On dénote $(Y_t)_{t\geq 0}$ le capital du gérant du jeu au temps t. Le but est de modéliser l'évolution du capital, et d'estimer la probabilité que le gérant soit ruiné durant le processus.

Le modèle

Les rentrées d'argent du gérant (c'est-à-dire les dépenses des clients) sont déterministes et gouvernées par un paramètre $\alpha > 0$ de la façon suivante : entre les instants 0 et t, le gérant gagne αt.
Les gains des joueurs, quant à eux, ont lieu aux instants de saut d'un processus de Poisson $(N_t)_{t\geq 0}$ de paramètre 1 : on considère $(\xi_k)_{k\geq 0}$ une suite de variables aléatoires indépendantes identiquement distribuées de loi exponentielle de paramètre 1 et on pose : $T_0 = 0$,

$$\forall i \in \mathbb{N}^*, T_i = \sum_{k=1}^{i} \xi_k \tag{1}$$

et

$$\forall t \in \mathbb{R}^+, N_t = \max\{i \in \mathbb{N}, T_i \leq t\}. \tag{2}$$

Les gains des joueurs sont donnés par une suite de v.a. positives $(X_i)_{i\geq 0}$ i.i.d., indépendantes du processus $(N_t)_{t\geq 0}$. A l'instant T_1, un joueur gagne X_1, à l'instant T_2, un joueur gagne X_2, etc. A tout instant t, N_t est donc le nombre de joueurs qui ont eu un gain avant l'instant t.
On suppose que X_1 admet un moment d'ordre 2 et on définit :

$$\mu = \mathbb{E}[X_1], \sigma^2 = \mathrm{Var}[X_1]. \tag{3}$$

Soit Y_0 le capital du gérant à l'instant initial. Le capital du gérant à l'instant t est donc :

$$Y_t = Y_0 + \alpha t - \sum_{i=1}^{N_t} X_i. \tag{4}$$

On définit la probabilité de ruine du gérant comme une fonction du capital initial :

$$r(y) = \mathbb{P}(\exists t \in \mathbb{R}_+, Y_t < 0 | Y_0 = y). \tag{5}$$

T1. Montrer que pour tout $y > 0$, $r(y) > 0$.

S1. Représenter 5 réalisations du processus $(N_t)_{t\in[0;100]}$. Pour $\alpha = 1$, et X_1 de loi de Poisson d'espérance 1, représenter 5 réalisations le processus $(Y_t)_{t\geq 0}$.

Cas $\alpha < \mu$

T2. Montrer que les variables $N_i - N_{i-1}$ sont des variables aléatoires de Poisson de paramètre 1. En déduire que

$$\frac{N_t}{t} \to 1 \quad p.s. \tag{6}$$

On rappelle la définition de la fonction de répartition empirique : pour un échantillon $(X_1, \dots, X_m)$, $\hat{F}_m(x) = \frac{1}{m}\sum_{i=1}^m \mathbb{1}_{X_i \leq x}$.

S2. En simulant $m = 100$ répétitions indépendantes de l'expérience, représenter un histogramme empirique de la variable $N_i - N_{i-1}$, pour un i fixé. Calculer et représenter la fonction de répartition empirique. Que remarque-t-on ? Quel résultat du cours retrouve-t-on ?

T3. Montrer que $\frac{Y_t}{t} \to \alpha - \mu$ ps. En déduire que si $\alpha < \mu$, alors pour tout $y \in \mathbb{R}_+$, $r(y) = 1$. Commenter ce résultat.

S3. Pour $\alpha = 1$ et $\mu = 2$ (on utilisera une loi de Poisson à nouveau), représenter 10 réalisation de trajectoires du processus $(Y_t)_{t\geq 0}, Y_0 = 10$.

S4. Déterminer le temps moyen de ruine du gérant $\bar{T}(\mu)$, pour $\alpha = 1$, et $\mu \in]1; 2]$ (on utilisera une grille de valeurs pour μ). Pour ce faire, on répètera l'expérience à de multiples reprises et reportera les résultats sous forme de diagramme en boite.

Cas $\alpha = \mu$

T4. On définit, pour tout $n \in \mathbb{N}$,

$$S_n = \sum_{i=1}^{n} (X_i - \alpha\xi_i). \tag{7}$$

1. Quel est le lien entre Y_{T_n} et S_n ?

2. Montrer que l'évènement $\limsup_{n\to\infty}\{\frac{S_n}{\sqrt{n}} > 1\}$ appartient à la tribu asymptotique de la suite $(X_i - \alpha\xi_i)_{i\geq 1}$. Montrer que $\limsup_{n\to\infty} S_n = +\infty$.

3. Montrer que pour tout $y \in \mathbb{R}_+$, $r(y) = 1$. Commenter ce résultat.

S5. Simuler différentes trajectoires du processus $(S_n)_{1\leq n\leq 10000}$. Déterminer le temps moyen de ruine du gérant $\bar{T}(\alpha)$, en fonction de $y \in \{1, 2, 5, 10\}$... Commenter les résultats.

En pratique, que pensez-vous de cette modélisation ? Quel phénomène a-t-on ignoré ?

Cas $\alpha > \mu$

Le gérant a évidemment intérêt à choisir $\alpha > \mu$. On se place à présent dans cette configuration et on fait l'hypothèse suivante sur le modèle :

$$(H_A) \qquad \exists A > 0 \text{ tel que } \exp(AX_1) \text{ soit intégrable et } \mathbb{E}[\exp(A(X_1 - \alpha T_1))] = 1. \tag{8}$$

T5. Vérifier que l'équation (8) admet **au plus** une solution strictement positive en la variable A.

Rappeler l'expression de la fonction génératrice des moments d'une variable aléatoire exponentielle de paramètre λ. Si X_1 suit une loi exponentielle de paramètre $\lambda > \alpha^{-1}$, quelle valeur de A satisfait (8) ?

S6. On considère dans cette question X_1 de loi gaussienne (on ignorera le fait que X_1 puisse avoir des valeurs négatives), de moyenne $\mu = \alpha/2$, et de variance $\sigma = \mu/2$. Prouver l'existence d'une solution et déterminer numériquement la valeur de la solution pour différentes valeurs de α.

T6. On veut montrer que sous l'hypothèse (H_A), la probabilité de ruine vérifie $r(y) \leq \exp(-Ay)$ pour tout $y \geq 0$. On utilise à nouveau la marche aléatoire S_n définie en (7). On définit $r_n(y) := \mathbb{P}(\max_{1\leq k\leq n} S_k > y)$.

— Montrer que $r(y) = \mathbb{P}(\max_{n\geq 1} S_n > y) = \lim_{n\to\infty} \mathbb{P}(\max_{1\leq k\leq n} S_k > y)$.

— Montrer par récurrence que pour tout $n > 0$, pour tout $y \geq 0$, $r_n(y) \leq \exp(-Ay)$. (indication : pour l'hérédité, on décomposera l'évènement selon que $S_1 > y$ ou non).

— Conclure.

S7. Illustrer ce résultat par une simulation.

Cas des gains exponentiels

Dans le cas des gains exponentiels, on a vu ci-dessus qu'il était possible de déterminer la valeur exacte de la constante A. On montre à présent qu'il est possible de donner une forme explicite de la fonction $r(y)$.

T7. Soit $g(y) = 1 - r(y) = \mathbb{P}(\max_{1\leq n} S_n \leq y)$. Montrer que g satisfait :

$$g(y) = \frac{e^{y/\alpha}}{\alpha} \int_y^\infty \left(\int_0^z g(v)\lambda e^{\lambda v} dv \right) e^{-(\lambda+1/\alpha)z} dz. \tag{9}$$

En dérivant, montrer que $r(y) = \frac{e^{-(\lambda-1/\alpha)y}}{\alpha y}$ vérifie l'équation précédente et conclure.

S8. On souhaite vérifier si les gains des joueurs sont distribués selon une loi exponentielle et en estimer les paramètres.
Simuler un échantillon $(X_1, \ldots, X_n)$ de variables aléatoires exponentielles de paramètre λ de votre choix. Rappeler l'expression et calculer l'estimateur du maximum de vraisemblance.

On introduit la statistique suivante :

$$H_n = \sup_{x\geq 0} \left| \hat{F}_n(x) - \left(1 - e^{-x/\bar{X}_n}\right) \right|, \tag{10}$$

avec $\bar{X}_n$ la moyenne empirique des $(X_i)_{1\leq i\leq n}$, et $\hat{F}_n$ la fonction de répartition empirique.

T8. Montrer que sous l'hypothèse que les variables $(X_1, \ldots, X_n)$ suivent en effet des lois exponentielles i.i.d., $H_n \to_{n\to\infty} 0$ p.s.
Que dire dans le cas contraire ?

S9. Pour différentes valeurs de n et λ, représenter la distribution empirique de H_n et formuler une conjecture. Prouver cette conjecture.

Facultatif. Proposer un test pour vérifier si la loi des observations est exponentielle. En particulier, estimer la valeur de la région de rejet pour un test de niveau 5%.

6 | Variables de Cauchy

proposé par Aymeric Dieuleveut, aymeric.dieuleveut@gmail.com

On considère dans ce problème des variables aléatoires de loi $\mathcal{C}(\theta, \sigma)$, de densité

$$f_{(\theta,\sigma)}(x) = \frac{1}{\pi\sigma(1 + (\frac{x-\theta}{\sigma})^2)}$$

sur $\mathbb{R}$, avec $(\theta, \sigma) \in \mathbb{R} \times \mathbb{R}_+$. On appelle θ le "paramètre de position", et σ le "paramètre d'échelle". Dans ce problème, on se propose d'étudier certaines propriétés de ces variables, ainsi que l'estimation statistique des paramètres.
Les distributions de Cauchy interviennent dans de nombreux phénomènes physiques : par exemple, si une source ponctuelle lumineuse est placée au dessus d'un sol horizontal, et émet des rayons de façon uniforme, alors la quantité de lumière projetée au sol suit une distribution de Cauchy.

Préliminaires

Déterminer la fonction de répartition de la loi $\mathcal{C}(\theta, \sigma)$.
Déterminer la fonction caractéristique de la loi $\mathcal{C}(\theta, \sigma)$,

$$\varphi_{(\theta,\sigma)}(t) = \int_{-\infty}^{\infty} \exp(ixt) f_{(\theta,\sigma)}(x) dx.$$

Que peut-on dire de la moyenne et de la variance de la loi $\mathcal{C}(\theta, \sigma)$? Que peut-on dire de $\varphi_{(\theta,\sigma)}$ en 0 ?
Soit $X \sim \mathcal{C}(0, 1)$. Comment obtenir une variable aléatoire de loi $\mathcal{C}(\theta, \sigma)$ (on démontrera ce résultat) ?

Simulation d'une loi de Cauchy

S1. A partir de la fonction de répartition, proposer une méthode pour simuler une variable aléatoire de loi $\mathcal{C}(\theta, \sigma)$. Justifier l'interprétation physique donnée dans l'introduction. Simuler un échantillon de 1000 variables i.i.d. de loi $\mathcal{C}(\theta, \sigma)$, et représenter son histogramme. Quelle observation peut-on faire ? Comment rendre cet histogramme plus lisible ?

T1. On considère Y_1, Y_2 deux variables aléatoires indépendantes de loi normale $\mathcal{N}(0, 1)$. Calculer la loi de $\frac{Y_1}{Y_2}$. A partir de cette observation, comment simuler une variable $\mathcal{C}(\theta, \sigma)$?

S2. Simuler un échantillon de 1000 variables aléatoires $\mathcal{C}(\theta, \sigma)$ à partir de la seconde méthode. Tracer la fonction de répartition empirique. Qu'observe-t-on ? A quel résultat théorique cela correspond-il ?

Inverse d'une loi de Cauchy

Soit $n \in \mathbb{N}$ et $X_1, \ldots, X_n$ des variables aléatoires indépendantes de loi $\mathcal{C}(0, \sigma)$.

T2. Déterminer la loi de $1/X_1$ à partir de la fonction de répartition. Comment aurait-on pu anticiper ce résultat ?

T3. On définit la moyenne empirique (ou arithmétique) Z_n et la moyenne harmonique W_n :

$$Z_n = \frac{X_1 + \cdots + X_n}{n}, \quad W_n = \frac{n}{1/X_1 + \cdots + 1/X_n}.$$

Quelles sont les distributions de W_n, Z_n ?

S3. Comment simuler des variables aléatoires $X_1, \dots, X_n$ de densité $f_{(\theta,\sigma)}(x) = \frac{1}{\pi\sigma(1+(x-\frac{\theta}{\sigma})^2)} \mathbb{1}_{x\geq 0}$?
Que peut-on alors dire de Z_n et W_n ?
Quelles phénomènes "surprenants" observe-t-on (il y en a au moins 2) ?
Pour $n = 10$, simuler 100 tirages des variables Z_{10}, W_{10} et représenter les histogrammes et fonction de répartition empiriques. Commenter.

Loi des grands nombres, dispersion

T4. Des variables i.i.d. de Cauchy satisfont-elles la loi des grands nombres et le TCL ? Vérifier empiriquement cette assertion ? Que peut-on en conclure quant aux hypothèses de la LGN et du TCL ?

S4. On considère des variables aléatoires i.i.d de loi $\mathcal{N}(1,1)$, $\mathcal{C}(1,1)$. Représenter la somme des n premières variables aléatoires, pour n entre 1 et 10^5. Répéter cette expérience 5 fois (on fera 5 expériences numériques avec des Gaussiennes, puis on fera 5 autres expériences avec des Cauchy). Commenter les résultats.

Inférence statistique pour le paramètre θ, $\sigma = 1$.

On considère un $n-$échantillon de loi de Cauchy de densité $f_{(\theta,1)}(x) = \frac{1}{\pi(1+(x-\theta)^2)}$.

Maximum de vraisemblance.

T5. Calculer la fonction de vraisemblance de cet échantillon.
On note $\hat{\theta}_{\mathrm{MV}}$ la valeur du paramètre estimée, à partir de cet échantillon par la méthode du maximum de vraisemblance. Que vérifie $\hat{\theta}_{\mathrm{MV}}$? Pourquoi existe-t-il ?

S5. Application numérique :

a) Générer un échantillon de taille $n = 200$ suivant une loi de Cauchy de paramètre $\theta = 10$.
On considère à présent que la valeur du paramètre θ nous est inconnue, et on souhaite l'estimer à partir du $n-$échantillon précédemment simulé.

b) Méthode de la grille (I) :
Tracer la fonction de log vraisemblance pour 100 valeurs du paramètre θ choisies entre $\theta = 1$ et $\theta = 100$. Puis, proposer une méthode de résolution graphique puis numérique pour calculer une valeur approchée de $\hat{\theta}_{\mathrm{MV}}$.

c) Méthode du gradient (II) :
Rappeler la méthode de Newton-Raphson puis utiliser cette méthode pour calculer une autre approximation de l'estimateur $\hat{\theta}_{\mathrm{MV}}$. Quelle méthode vous semble la plus pertinente ?

Méthode alternative (1).

T6. À quoi correspond le paramètre θ d'une Cauchy ? Proposer un nouvel estimateur $\hat{\theta}_{\mathrm{med}}$ de θ basé sur cet remarque.

Méthode alternative (2).

On dénote $X_{(1)}, \dots, X_{(n)}$ les statistiques d'ordre de $X_1, \dots, X_n$, c'est à dire les variables aléatoires telles que

$$\{X_{(1)}, \dots, X_{(n)}\} = \{X_1, \dots, X_n\}, \quad \text{et } X_{(1)} \leq X_{(2)} \cdots \leq X_{(n)}.$$

T7. Dans cette question, on considère $n = 3$, et $\theta = 0$. Montrer que $X_{(1)}$ et $X_{(3)}$ n'ont pas de moment d'ordre 1. (On remarquera que $X_1 \mathbb{1}_{X_1>0} \leq X_{(3)} \mathbb{1}_{X_1>0}$ p.s.).

Montrer que la densité de $X_{(2)}$ est :

$$f_{(2)}(x) = \frac{6}{\pi^3}\frac{1}{1+x^2}\left(\frac{\pi}{2}+\arctan(x)\right)\left(\frac{\pi}{2}-\arctan(x)\right).$$

Montrer que $X_{(2)}$ admet un moment d'ordre 1 et le calculer. La variance de $X_{(2)}$ est elle finie ?
Pour $n > 3$, conjecturez quelles statistiques d'ordre admettent des moments d'ordre p, $p \geq 1$.
Pour $n = 3$, vérifier empiriquement si une suite de variables aléatoires i.i.d. de même loi que $X_{(2)}$ satisfait le TCL.
Pour $n = 5$, vérifier empiriquement si une suite de variables aléatoires i.i.d. de même loi que $X_{(3)}$ satisfait le TCL.

T8. On considère $q \in]0;1[$, et $(Y_1, \ldots, Y_{n_q})$ telles que

$$(Y_1, \ldots, Y_{n_q}) = (X_{(\lceil \frac{1-q}{2} n \rceil)}, \ldots X_{(\lfloor \frac{1+q}{2} n \rfloor)})$$

Soit

$$\hat{\theta}_q := \mathbb{E}\left[\frac{Y_1 + \cdots + Y_{n_q}}{nq}\right]$$

Montrer que $\hat{\theta}_q$ existe, et converge quand $n \to \infty$. Déterminer sa limite.

S6. Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes de loi $\mathcal{C}(0,1)$. Déterminer une approximation de $\mathbb{E}[X_{(n-1)}]$, pour $n \in \{3, 10, 100, 1000\}$.

Application numérique :

S7.

a) Générer $m = 100$ échantillons (de taille $n \in \{10, 100, 1000, 10000\}$) d'une loi de Cauchy de paramètres de votre choix. Pour chacun de ces échantillons, calculer les différents estimateurs proposés (pour $\hat{\theta}_q$, on utilisera $q \in \{0.1, 0.25, 0.50, 0.75\}$). On définit le risque quadratique moyen d'un estimateur comme :

$$R(\hat{\theta}) = \mathbb{E}_{X_1,\ldots,X_n \sim \mathcal{C}(\theta)}[|\hat{\theta} - \theta|^2] \tag{1}$$

Calculer une approximation du risque moyen de chaque estimateur proposé. Quel estimateur vous semble le plus adapté ?

b) Déterminer la valeur optimale q_n^* du paramètre q, pour $n \in \{10, 100, 1000, 10000\}$.

c) Télécharger les échantillons "Dataset 1,2,3,4" disponibles à l'adresse suivante https://www.dropbox.com/sh/54c6qtorl1hvgmz/AAAU8ohchNJp8cIWrrFLQWFMa?dl=0. Pour chacun d'entre eux, estimer les paramètres θ par les différentes méthodes proposées ci-dessus, et commenter.

Facultatif : Inférence statistique pour le paramètre σ, $\theta = 0$.

On considère un $n-$échantillon de loi de Cauchy de densité $f_{(1,\sigma)}(x) = \frac{1}{\pi\sigma(1+(\frac{x}{\sigma})^2)}$.

T9. Soit X une variable aléatoire de loi $\mathcal{C}(0,\sigma)$. Calculer $\mathbb{P}(|X| \leq 2\sigma)$. Quelle est l'interprétation de ce résultat ? Proposer un estimateur de σ.

S8. Proposer une méthode pour calculer l'estimateur du maximum de vraisemblance de σ.

Application numérique :

a) Générer un échantillon de taille $n = 200$ suivant une loi de Cauchy de paramètre $\theta = 0$, $\sigma = 1$. Pour les deux estimateurs ci-dessus, calculer une approximation de l'erreur quadratique moyenne d'estimation de σ.

b) Télécharger les échantillons "Dataset 5,6 " disponibles à l'adresse suivante https://www.dropbox.com/sh/54c6qtorl1hvgmz/AAAU8ohchNJp8cIWrrFLQWFMa?dl=0. Pour chacun d'entre eux, estimer les paramètres θ par les différentes méthodes proposées ci-dessus, et commenter.

7 | Dimensionnement d'un accueil de banque

proposé par Arnaud Guillin, guillin@math.univ-bpclermont.fr

On s'intéresse ici au problème suivant : une banque souhaite dimensionner son accueil pour les opérations journalières de ses clients, c'est-à-dire qu'elle veut évaluer le nombre de places à prévoir pour que

1. la banque ne soit pas complètement pleine (et que des clients ne puissent pas être servis),
2. la banque ne soit pas globalement vide (la salle est trop grande, perte d'argent, ...)
3. le temps d'attente d'un client soit raisonnable.

1 Cas d'un guichet

On va s'intéresser dans un premier temps au cas où il n'y a qu'un seul guichet et où la capacité d'accueil est supposée infinie, et nous allons raisonner en temps discret : à chaque unité de temps, les évènements suivant peuvent arriver

— s'il y a un client au guichet, il est servi avec probabilité $0 < p < 1$, sinon il reste au guichet

— un client arrive dans la banque avec probabilité $0 < q < 1$

et ces deux évènements surviennent de manière indépendante.

T1. Soit $(D_i)_{i\geq 1}$ une suite de v.a.i.i.d. telles que $\mathbb{P}(D_i = 1) = p = 1 - \mathbb{P}(U_i = 0)$ et $(A_i)_{i\geq 1}$ une suite de v.a.i.i.d. telles que $\mathbb{P}(A_i = 1) = q = 1 - \mathbb{P}(A_i = 0)$ indépendantes des (D_i). On note $(X_i)_{i\geq 0}$ la longueur de la file d'attente au temps i, que l'on définit de manière récursive pour $i \geq 1$ par

$$X_i = \max(X_{i-1} - D_i, 0) + A_i$$

et X_0 une variable aléatoire de $\mathbb{N}$ (indépendante des $(A_i), (D_i)$). Montrer que (X_i) décrit bien la dynamique de la file d'attente.

S1. Simulez des trajectoires de la longueur de la file d'attente pendant une période de 1000 pour $(p, q) = (0.1, 0.9)$, ou $(0.9, 0.1)$ ou $(0.45, 0.55)$ ou $(0.9, 0.8)$. Qu'observez-vous ?

T2. Calculer pour tout $i > 0$ et tout x, y et $x_0, ..., x_{i-1}$ de $\mathbb{N}$

$$\mathbb{P}(X_{i+1} = y \,|\, X_i = x, X_{i-1} = x_{i-1}, ..., X_0 = x_0) := P_{xy}$$

et montrer que

$$\mathbb{P}(X_{i+1} = y \,|\, X_i = x, X_{i-1} = x_{i-1}, ..., X_0 = x_0) = \mathbb{P}(X_{i+1} = y \; X_i = x).$$

On appelle un tel processus une chaîne de Markov homogène de matrice de transition P. Montrer également que pour tout x, y de $\mathbb{N}$ il existe $l = l(x, y)$ tel que $\mathbb{P}(X_l = y \,|\, X_0 = x) > 0$. On dit qu'elle est irréductible.

S2. On note $T^0 := \inf\{n \geq 1\,;\, X_n = 0\}$. On suppose $X_0 = x$. Evaluer par simulation pour les 4 cas de **S1** la probabilité $\mathbb{P}(T_0 < \infty \,|\, X_0 = x)$.

T3. En utilisant la Loi des Grands Nombres pour la suite i.i.d. $(A_i - D_i)_i$, montrer que si $p > q$ alors pour tout $x > 0$, $\mathbb{P}(T_0 < \infty \,|\, X_0 = x) = 1$ et on est donc sûr de servir tous les clients. Il est donc

souhaitable pour la banque que son guichet aille plus vite pour servir que les clients n'arrivent, ce qui semble logique ! !

T4. Montrer que si $p > q$ alors il existe une probabilité π telle que pour tout x, y de $\mathbb{N}$ on a

$$\pi P = \pi.$$

On appelle π probabilité stationnaire ou invariante car si $X_0 \sim \pi$ alors pour tout $i \geq 1$, $X_i \sim \pi$.

S3. On note $T^x := \inf\{n \geq 1\,;\, X_n = x\}$. Evaluer par simulation $\mathbb{E}(T^x \,|\, X_0 = x)$ et comparer π_x avec $\frac{1}{\mathbb{E}(T^x \,|\, X_0 = x)}$.

S4. Montrer par simulation que quelque soit X_0, si et dans les cas $(p, q) = (0.9, 0.1)$ ou $(0.9, 0.8)$, pour n grand on a X_n à peu près de même loi que π.

T5. La banque choisit finalement que la taille de sa file d'attente sera de l'ordre de $L = \mathbb{E}(X_0) + \sqrt{2Var(X_0)}$ avec $X_0 \sim \pi$. Expliquer ce choix puis calculer L.

2 Cas multi-guichet

On va maintenant supposer qu'il y a N guichets indépendants travaillant comme dans la section précédente mais que les arrivées suivent une loi de Poisson de paramètre θ. La banque considère qu'lle veut fonctionner comme avant. C'est-à-dire fixer une taille de salle $L = \mathbb{E}(X_0) + \sqrt{2Var(X_0)}$ avec $X_0 \sim \pi$, mais bien évidemment π est nettement plus dur à calculer (si elle existe).

S5. Pouvez vous donner une condition intuive pour que la probabilité invariante puisse exister. Vous pourrez essayer de vérifier que votre condition est cohérente par simulation (en fixant $N = 5$ par exemple et en jouant sur p et θ).

Nous supposerons dans la suite que N, p et θ sont tels qu'il existe une (unique) probabilité invariante. Si π n'est pas calculable, nous allons essayer de l'approcher ! Pour cela, nous allons admettre le résultat observé dans la section précédente :

$$\pi_x = \frac{1}{\mathbb{E}(T^x \,|\, X_0 = x)}$$

et pour tout x de $\mathbb{N}$, $\pi_x > 0$.

T6. Le but de cette longue question est de montrer la Loi des Grands Nombres pour les chaînes de Markov dans notre cas : si F est telle que $\sum_x \pi(x)|F(x)| < \infty$ alors

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} F(X_k) = \sum_x \pi(x)F(x) := \pi(F). \qquad \mathbb{P} - ps$$

Nous allons procéder en plusieurs étapes : on supposera que $X_0 = 0$

1. On note par récurrence avec $T_1^0 = T^0$, pour tout $k > 1$, $T_k^0 := \inf\{n > T_{k-1}^0\,;\, X_n = 0\}$. Montrer que les variables pour $i > 1$, $U_i = T_i^0 - T_{i-1}^0$ sont des variables i.i.d. de loi T^0. Montrer également que les variables
$$Y_i = \sum_{k=T_i^0}^{T_{i+1}^0 - 1} F(X_k)$$
sont i.i.d..
2. Calculer $\mathbb{E}(Y_i)$.
3. Montrer ensuite que
$$\frac{1}{k} \sum_{l=0}^{T_{k+1}^0 - 1} F(X_l) \overset{\mathbb{P}-p.s}{\longrightarrow} \frac{\pi(F)}{\pi_0}$$
et en déduire que
$$\frac{1}{k} T_{k+1}^0 \to \frac{1}{\pi_0}.$$

4. En déduire finalement
$$\frac{1}{T_k^0}\sum_{l=0}^{T_k^0-1} F(X_l) \overset{\mathbb{P}-p.s}{\longrightarrow} \pi(F).$$
5. On note $N_n = \sum_{l=1}^{n} 1_{X_l=x}$. Montrer que $N_n \to \infty$ $p.s.$. Montrer que $T^0_{N_n} \leq N < T^0_{N_n+1}$.
6. En déduire le résultat (Attention au dernier bout...).

S6. Par simulation, estimer la longueur de la file d'attente dans le cas $N = 5$, $p = 0.8$, $\theta = 1$, puis $N = 10$, $p = 0.2$, $\theta = 1.5$

S7. Par simulation, évaluer le temps moyen d'attente d'un client arrivant (dans les deux cas précédant) en régime stationnaire.

8 | Modèle du votant

proposé par Arnaud Guillin, guillin@math.univ-bpclermont.fr

1 Modélisation

On souhaite modéliser la contagion concurrente de deux opinions dans une population donnée finie (taille N).
Bien évidemment, nous allons nous placer dans un cadre relativement restrictif mais il existe des généralisations nombreuses (mais malheureusement inabordables à ce niveau) et pertinentes.
On suppose donc que les individus sont immobiles et en interaction avec certains autres individus de la population. Il y a donc un graphe sous-jacent de connections entre les individus. Les partisans de l'opinion A, resp. opinion B, seront affublés du numéro 1, resp. du numéro 0.
Dans toute la suite, on fait les hypothèses suivantes :

1. chaque individu influence directement ses voisins et eux seulement,
2. l'influence est positive : un individu a tendance à rallier son voisin à son opinion,
3. tous les individus ont le même pouvoir d'influence,
4. chacun des N individus est en communication directe avec au plus deux autres individus,
5. deux individus donnés sont en interaction au moins indirecte,
6. un seul individu change d'opinion par unité de temps.

Ces conditions, encore une fois très restrictives, ont une cohérence qui va nous garantir une écriture d'un modèle simple.

2 Définitions et modèle

Les individus sont numérotés de 1 à N et on leur affecte la valeur 1 si ils sont proA, et 0 si ils sont proB. L'opinion de la population à un instant donné s'identifie donc à un élément de $E = \{0,1\}^S$ avec $S = \{1, ..., N\}$. Pour modéliser les changements possibles, on va munir E d'une structure de graphe non orienté.
Un graphe non orienté est un couple (S, A) où les éléments de S sont les sommets du graphe et ceux de A les arêtes reliant ces sommets. Deux points x et y de S sont dits voisins si une arête les relie.
Identifions E à l'ensemble des applications de S dans $\{0,1\}$: pour $\eta \in E$,

$$\forall x \in S, \eta(x) \in \{0,1\}.$$

Les éléments η de E sont appelés *configurations*.

On veut construire un processus aléatoire à valeurs dans E qui va décrire les changements d'opinion à un instant donné : le processus ne peut passer d'une configuration η à une configuration η' que si celle-ci ne diffère de η qu'en exactement **un site**.
Pour $\eta \in E$ et $x \in S$, on note η_x la configuration changée au site x : $\forall y$

$$\eta_x(y) = \eta(y)\, 1_{\{y \neq x\}} + (1 - \eta(x))\, 1_{\{y=x\}}.$$

Nous avons encore la lattitude de spécifier pour toute configuration η et tout site x les probabilités de transition de η vers η_x.

3 Modèle cyclique

Nous avons décidé en introduction que chaque individu peut influencer exactement deux autres voisins. Le graphe correspondant à cette situation est naturellement le suivant : soit S l'ensemble $\{0, ..., N-1\}$ identifié à $\mathbb{Z}/N\mathbb{Z}$. On relie alors les points avec des arêtes de proche en proche, soit finalement le cercle discret à N points. Les points x et y sont donc voisins si $|x - y| = 1\, mod(N)$.

3.1 Une chaîne de Markov naturelle

Une chaîne de Markov est (grossièrement) un processus qui pour évoluer n'a besoin de connaître de son passé que le présent : le processus (X_n) est donc tel que X_{n+1} est une variable aléatoire dépendant de l'état de X_n (mais pas de $X_{n-1}, ..., X_0$). C'est ce processus X_n qui va modéliser l'évolution de l'opinion dans la population.

T1. Si à un instant donné n tous les individus ont le même avis, que se passe-t-il ensuite ?

Décrivons plus précisément les transitions : considérant une configuration η différente de $(0, ..., 0)$ et $(1, ..., 1)$, sachant que $X_n = \eta$, $X_{n+1} = \eta_x$ où x est choisi aléatoirement (et indépendamment du passé) avec une probabilité proportionnelle au nombre de voisins de x d'opinion différente.

T2. Préciser justement cette probabilité en fonction de η.

T3. Pouvez vous trouver une fonction f telle que $X_{n+1} = f(X_n, u_{n+1})$ où la suite u_n est i.i.d. de loi uniforme sur $[0, 1]$? (cela justifie notamment que l'on construit bien une chaîne de Markov.)

S1. Soit $N = 10$, partant d'un état initial quelconque ; faire une représentation graphique de quelques évolutions aléatoires de l'opinion.

3.2 Evolution en temps long

Notons pour tout $n \in \mathbb{N}$, $S_n = \sum_{x \in S} X_n(x)$.

T4. Que représente S_n ? Montrer que $\mathbb{E}(S_n|(X_{n-1}, ...X_0)) = S_{n-1}$. En déduire que le nombre moyen de sites marqués 1 est constant au cours du temps.

S2. Retrouvez par simulation ce phénomène.

S3. Montrer par simulation que si $\mathbb{E}(S_0) = Np$ alors la loi de X_n converge vers

$$(1 - p)\delta_{(0,...,0)} + p\delta_{(1,...,1)}.$$

3.3 Composantes connexes

Si η est une configuration, on appelle composante connexe de η un ensemble maximal, au sens de l'inclusion, de sites voisins sur lequel η est constante.

S4. Montrer par simulation que le nombre de composantes connexes de X_n est décroissant en fonction de n. (Vous pouvez essayer de le montrer théoriquement aussi...)

3.4 Evolution du nombre de proA

En fait S_n est une chaîne de Markov à valeurs dans $\{0, ..., N\}$ dont on pourrait d'ailleurs déterminer la matrice de transition. On va s'intéresser ici au phénomène d'absorption. Pour tout $i = 0, ..., N$ on note

$$a_i(n) = \mathbb{P}(S_n = 0|S_0 = i) \qquad b_i(n) = \mathbb{P}(S_n = N|S_0 = i).$$

T5. Montrer que pour tout $i = 1, ..., N - 1$ on a

$$a_i(n+1) = \frac{1}{2}a_{i-1}(n) + \frac{1}{2}a_{i+1}(n) \qquad b_i(n+1) = \frac{1}{2}b_{i-1}(n) + \frac{1}{2}b_{i+1}(n).$$

Soit U le nombre de sauts de de la chaîne avant absorption $U = \inf\{n\,;\, S_n \in \{0, N\}\}$. On pose $u_i = \mathbb{E}(U|S_0 = i)$.

S5. Pour N donné, donner une évaluation Monte Carlo de u_i.

T6. Montrer que les u_i sont solution de

$$u_i = \frac{1}{2}u_{i-1} + \frac{1}{2}u_{i+1} + 1, \qquad u_0 = u_N = 0$$

et résoudre cette équation (on peut chercher les solutions sous la forme $u_i = a + bi + ci^2$).

4 Une généralisation (Facultatif)

Il y a maintes généralisations possibles. On s'intéresse ici au cas suivant : on rajoute deux individus aux opinions indéfectibles : $S = \{-1, ..., N+1\}$ avec -1 et $N+1$ qui ne communiquent pas (et le reste est inchangé). On ne considére que les configurations où $\eta(-1) = 0$ et $\eta(N+1) = 1$. Le reste de la dynamqie est le même.

S6. Simuler ce nouveau processus. Lorsque le temps est suffisament grand, combien observez vous de composantes connexes ?

T7. On suppose à présent que X_0 a exactement deux composantes connexes. On note $Y_n = \max\{x \in S; X_n(x) = 0\}$. Quelle est la dynamique de Y_n ?

S7. Quel est le comportement de Y_n quand n tend vers l'infini ?

9 | Étude des graphons

proposé par Pierre Latouche, pierre.latouche@polytechnique.edu

Ce projet s'intéresse au modèle de W-graphe caractérisé par une fonction appelée *graphon*. Dans un premier temps, nous analyserons ce modèle et construirons des fonctions de simulation. La seconde partie établira le lien entre ce modèle et des modèles plus simples de graphe aléatoire. Ces liens permettront enfin de proposer une stratégie d'inférence sur données réelles.

1 Modèle

Nous considérons un modèle de graphe aléatoire appelé modèle de W-graphe. D'un point de vue théorique, ce modèle est défini comme l'objet mathématique caractérisant la limite d'une série de graphes denses lorsque le nombre de sommets augmente. Dans ce projet, un graphe sera dit *dense* si son nombre d'arêtes est quadratique en son nombre de sommets.
Chaque sommet est d'abord associé à une variable latente tirée à partir d'une loi uniforme sur l'intervalle $[0, 1]$. La probabilité d'apparition d'une arête entre deux sommets est alors supposée dépendre de leurs variables latentes respectives. Ainsi, pour chaque sommet $i \in \{1, \dots, n\}$ du graphe, une variable U_i est tirée à partir d'une loi $\mathcal{U}([0, 1])$ et la présence d'une arête (i, j) dans le graphe est générée à partir d'une loi de Bernoulli de probabilité $W(U_i, U_j)$. La fonction $W(\cdot, \cdot)$:

$$W : [0, 1]^2 \to [0, 1],$$

est appelée *graphon*. Toutes les variables U_i sont tirées aléatoirement de manière *iid*. Les arêtes sont elles indépendantes, conditionnellement. En résumé :

- $U_i \sim \mathcal{U}([0, 1]), \forall i \in \{1, \dots, n\}$ (iid)
- $X_{ij}|(U_i, U_j) \sim \mathcal{B}(W(U_i, U_j)), \forall (i, j), i \neq j$

Ainsi, $X_{ij} = 1$ s'il existe une arête (i, j) dans le graphe et 0 sinon. Par la suite, nous noterons $X \in \mathcal{M}_{n\times n}(\{0, 1\})$ la matrice d'adjacence associée au graphe et contenant les éléments X_{ij}. De plus, nous fixerons $U = (U_i)_i$, l'ensemble des variables U_i. Les graphes considérés sont supposés non-orientés et sans auto-boucle, donc la matrice X est symétrique. Nous avons également $X_{ii} = 0, \forall i \in \{1, \dots, n\}$.
Le modèle de W-graphe généralise la plupart des modèles de graphe aléatoire existant. Nous verrons notamment que le modèle d'Erdös et Rényi de 1959 ainsi que le modèle à blocs stochastiques (Holland) de 1983 sont des cas particuliers de W-graphe. L'unique objet permettant de simuler selon ce modèle est la fonction $W(\cdot, \cdot)$. Cela pose des difficultés dans la phase d'estimation sur données réelles.

T1. Donnez la loi marginale de X_{ij}.

T2. Montrez que les variables U_i de U ne sont *pas* indépendantes, connaissant la matrice d'adjacence X et la fonction graphon $W(\cdot, \cdot)$. *Montrez que la loi de U sachant X et $W(\cdot, \cdot)$ ne s'écrit pas comme un produit de facteurs en U_i.*

S1. Ecrivez une fonction *simu* prenant en entrée n ainsi qu'une fonction graphon $W(\cdot, \cdot)$, et simulant un graphe aléatoire selon un modèle de W-graphe.

T3. Montrez que ce modèle n'est pas identifiable. *Considérez 2 fonctions graphons W_1 et $W_2 = f(W_1)$ où f est à caractériser, et montrez que la loi de X est la même sous W_1 et W_2.*

T4. Montrez que ce modèle est identifiable si on se restreint aux graphons $W(\cdot, \cdot)$ tels que $g(y) = \int_0^1 W(x, y)dx$ est une fonction croissante.

S2. Simulez un graphe à $n = 100$ sommets où $W(\cdot, \cdot)$ respecte les conditions d'identifiabilité.

T5. Montrez qu'un graphe simulé selon un modèle de W-graphe est nécessairement dense ou vide.

S3. Illustrez graphiquement la question précédente en utilisant la fonction *simu* et en étudiant les variations du nombre d'arêtes des simulations, en fonction de n.

2 Liens avec d'autres modèles de graphe aléatoire

L'objectif de cette section est d'établir les liens existant entre le modèle de W-graphe et le modèle d'Erdös-Renyi de 1959, ainsi que le modèle à blocs stochastiques de 1983.

T6. Le modèle d'Erdös-Rényi est le modèle le plus simple de graphe aléatoire. Il fait l'hypothèse que toutes les arêtes sont tirées de manière *iid* selon une loi de Bernoullli de paramètre p :

$$X_{ij} \sim \mathcal{B}(p), \forall i \neq j (iid).$$

Montrez que le modèle d'Erdös-Rényi est un cas particulier de W-graphe. *Considérez une fonction* $W(\cdot, \cdot)$ *particulière et montrez que* $\mathbb{P}(X_{ij} = 1) = p, \forall i \neq j$ *selon ce modèle*

S4. Ecrivez une fonction *simuER* prenant en entrée n ainsi qu'une probabilité p, et simulant un graphe selon un modèle d'Erdös-Rényi.

T7. Le modèle à blocs stochastiques de 1983 est une généralisation du modèle précédent où les blocs ainsi que les arêtes entre blocs sont générés à partir de modèles d'Erdös-Rényi avec des paramètres différents. Ainsi, pour chaque sommet $i \in \{1, \dots, n\}$ du graphe, un vecteur Z_i est tiré à partir d'une loi multinomiale de paramètre $(1, \pi)$ où $\pi^\intercal = (\pi_1, \dots, \pi_K)$ tel que $\pi_k \in]0, 1[, \forall k \in \{1, \dots, K\}$ et $\sum_{k=1}^{K} \pi_k = 1$. Ainsi, π est dans un $(K-1)$simplexe où K désigne le nombre de blocs. Par définition, $\mathbb{P}(Z_{ik} = 1) = \pi_k$ et $\sum_{k=1}^{K} Z_{ik} = 1, \forall i \in \{1, \dots, n\}$. Ensuite, si le sommet i est dans le bloc k et le sommet j dans le bloc l, la présence d'une arête (i, j) dans le graphe est tirée à partir d'une loi de Bernoulli de paramètre μ_{kl}. Tous les vecteurs Z_i sont tirés aléatoirement de manière *iid*. Les arêtes sont elles indépendantes, conditionnellement. En résumé :

— $Z_i \sim \mathcal{M}(1, \pi), \forall i \in \{1, \dots, n\}$ (iid)

— $X_{ij} | (Z_{ik} = 1, Z_{jl} = 1) \sim \mathcal{B}(\mu_{kl}), \forall (i, j), i \neq j$

Par la suite, nous noterons $\mu \in \mathcal{M}_{n \times n}([0, 1])$ la matrice dont les éléments sont les μ_{kl}. Montrez que le modèle à blocs stochastiques est un cas particulier de W-graphe. *En définissant* $\sigma_0 = 0$ *et* $\sigma_l = \sum_{k=1}^{l} \pi_k, \forall l \in \{1, \dots, K\}$, *vous construirez une grille sur l'intervalle* $[0, 1]$ *et une fonction* $W(\cdot, \cdot)$ *par blocs à l'aide des probabilités* μ_{kl}.

S5. Ecrivez une fonction *simuBlocs* prenant en entrée n ainsi que les paramètres π et μ, et simulant un graphe aléatoire selon le modèle à blocs stochastiques.

3 Estimation

Nous nous intéressons maintenant au problème de l'estimation des paramètres de la fonction graphon d'un W-graphe. Ce problème est particulièrement difficile et nécessite des outils que vous verrez en partie l'année prochaine. Nous proposons donc ici de nous appuyer sur le résultat de la question 3 de l'exercice précédent. Le modèle à blocs stochastiques étant un cas particulier de W-graphe, l'idée est d'estimer les paramètres des blocs, de construire le graphon associé, et de voir ce graphon comme un estimateur. Pour $Z = (Z_i)_i$ fixé, les estimateurs des paramètres du modèle à blocs stochastiques sont donnés par :

$$\hat{\pi_k} = \frac{n_k}{n}, \forall k \in \{1, \dots, K\},$$

où $n_k = \sum_{i=1}^{n} Z_{ik}$ et

$$\mu_{kk} = \frac{1}{n_k(n_k - 1)} \sum_{i \neq j}^{n} Z_{ik} Z_{jk} X_{ij}, \forall k \in \{1, \dots, K\},$$

$$\mu_{kl} = \frac{1}{n_k n_l} \sum_{i \neq j}^{n} Z_{ik} Z_{jl} X_{ij}, \forall k \neq l.$$

S6. Chargez les données blog.txt. Construisez Z à partir des partis politiques fournis. Estimez π ainsi que μ.

S7. Construisez et affichez le graphon par blocs associé. Discutez des résultats.

10 | Modèle de graphe aléatoire par blocs et composantes géantes

proposé par Pierre Latouche, pierre.latouche@polytechnique.edu

Ce projet s'intéresse à un modèle de graphe aléatoire par blocs qui est une généralisation du modèle d'Erdös-Rényi (1959). Dans un premier temps, nous analyserons ce modèle et définirons des fonctions de simulation. La seconde partie se concentrera sur l'estimation des paramètres sur données réelles. Enfin, les conditions d'apparition d'une composante connexe *géante* dans le graphe selon ce modèle seront étudiées. Ci-dessous, nous noterons (S) les questions faisant intervenir des simulations et des développements informatiques. Les questions (T) sont elles théoriques.

1 Modèle

Nous considérons un modèle de graphe aléatoire par blocs noté $\boldsymbol{\mathcal{M}}$ ci-dessous. Chaque sommet est supposé appartenir à un groupe particulier. Dès lors, la probabilité d'apparition d'une arête entre deux sommets dépend de leurs groupes respectifs. Ainsi, pour chaque sommet $i \in \{1, \dots, n\}$ du graphe, un vecteur Z_i est tiré à partir d'une loi multinomiale de paramètre $(1, \pi)$ où $\pi^\intercal = (\pi_1, \dots, \pi_K)$ tel que $\pi_k \in]0, 1[, \forall k \in \{1, \dots, K\}$ et $\sum_{k=1}^K \pi_k = 1$. Ainsi, π est dans un $(K-1)$simplexe où K désigne le nombre de groupes. Par définition, $\mathbb{P}(Z_{ik} = 1) = \pi_k$ et $\sum_{k=1}^K Z_{ik} = 1, \forall i \in \{1, \dots, n\}$. Ensuite, si le sommet i est dans le groupe k et le sommet j dans le groupe l, la présence d'une arête (i, j) dans le graphe est tirée à partir d'une loi de Bernoulli de paramètre μ_{kl}. Tous les vecteurs Z_i sont tirés aléatoirement de manière *iid*. Les arêtes sont elles indépendantes, conditionnellement. En résumé :

— $Z_i \sim \mathcal{M}(1, \pi), \forall i \in \{1, \dots, n\}$ (iid)

— $X_{ij} | (Z_{ik} = 1, Z_{jl} = 1) \sim \mathcal{B}(\mu_{kl}), \forall (i, j), i \neq j$

Ainsi $X_{ij} = 1$ s'il existe une arête (i, j) dans le graphe et 0 sinon. Par la suite, nous noterons $X \in \mathcal{M}_{n \times n}(\{0, 1\})$ la matrice d'adjacence associée au graphe et contenant les éléments X_{ij}. De plus, nous fixerons $Z = (Z_i)_i$, l'ensemble des vecteurs Z_i et $\mu \in \mathcal{M}_{K \times K}([0, 1])$ la matrice des μ_{kl}. Les graphes considérés sont supposés non-orientés et sans auto-boucle, donc les matrices X et π sont symétriques. Nous avons également $X_{ii} = 0, \forall i \in \{1, \dots, n\}$.

T1. Donnez la loi marginale de X_{ij}.

T2. Montrez que les vecteurs Z_i de Z ne sont *pas* indépendants, connaissant la matrice d'adjacence X et les paramètres π ainsi que μ. *Montrez que la loi de Z sachant X, π, et μ ne s'écrit pas comme un produit de facteurs en Z_i.*

S1. Ecrivez une fonction *simu* prenant en entrée n ainsi que les paramètres π et μ, et simulant un graphe aléatoire selon $\boldsymbol{\mathcal{M}}$.

S2. Simulez un graphe à $n = 100$ sommets, $K = 4$ clusters et une matrice π telle que $\pi_{kk} = 0.8, \forall k \in \{1, \dots, K\}$ et $\pi_{kl} = 0.1, \forall k \neq l$.

S3. Illustrez graphiquement le coût algorithmique de la fonction *simu* en étudiant les variations en temps d'exécution des simulations, en fonction de n.

S4. Ecrivez une fonction *simu2* dont les entrées et sorties sont similaires à *simu* mais ayant un temps d'exécution sensiblement inférieur. Vous simulerez directement les arêtes à l'intérieur et entre des paires de blocs.

S5. Reprenez la question S3. et illustrez l'amélioration proposée à la question S4.

2 Estimation

Nous nous intéressons maintenant au problème de l'estimation des paramètres de $\boldsymbol{\mathcal{M}}$ pour une matrice d'adjacence X observée.

T3. Montrez que la log-vraisemblance des données s'écrit :

$$\ell(\pi,\mu) = \log\left\{\sum_{Z\in\mathcal{H}}\left(\prod_{i<j}^{n}\prod_{k=1}^{K}\prod_{l=1}^{K}\left(\mu_{kl}^{X_{ij}}(1-\mu_{kl})^{1-X_{ij}}\right)^{Z_{ik}Z_{jl}}\prod_{i=1}^{n}\prod_{k=1}^{K}\pi_k^{Z_{ik}}\right)\right\},$$

où $\mathcal{H}$ désigne tous les ensembles Z de Z_i possibles.

T4. Est-il selon vous possible de maximiser $\ell(\pi,\mu)$ par rapport à π et μ ? Si oui, proposez une solution.

T5. Nous supposons Z connu, donnez les estimateurs de π et μ maximisant :

$$\ell_Z(\pi,\mu) = \sum_{i<j}^{n}\sum_{k=1}^{K}\sum_{l=1}^{K} Z_{ik}Z_{jl}\left(X_{ij}\log(\mu_{kl}) + (1-X_{ij})\log(1-\mu_{kl})\right) + \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\log(\pi_k)$$

S6. Chargez les données blog.txt. Construisez Z à partir des partis politiques fournis. Estimez π ainsi que μ et discutez des résultats.

3 Transition de phase

Dans cette section, nous étudions l'apparition d'une composante *géante* dans le graphe en fonction du nombre de sommets et des paramètres de $\boldsymbol{\mathcal{M}}$. Une composante connexe d'un graphe sera ici définie comme *géante* si asymptotiquement (où l'asymptotique est en n), le nombre de sommets dans cette composante est de l'ordre de n.
Nous considérons l'opérateur $T_{\kappa,\pi}$ qui à tout vecteur $v \in \mathbb{R}^K$ associe un vecteur de $\mathbb{R}^K$ tels que :

$$(T_{\kappa,\pi}v)_k = \sum_{l=1}^{K}\kappa(k,l)v_l\pi_l, \forall k \in \{1,\dots,K\}.$$

Notons que $T_{\kappa,\pi}$ est construit à partir d'un vecteur π de $\boldsymbol{\mathcal{M}}$ et d'un noyau $\kappa(x,y)$ sur un espace métrique fini $S = \{1,\dots,K\}$. De plus, nous nous intéressons aux normes suivantes :

$$||v||_2 = \left(\sum_{k=1}^{K}\pi_k v_k^2\right)^{\frac{1}{2}}, \forall v \in \mathbb{R}^K,$$

et

$$||T_{\kappa,\pi}|| = \sup\{||T_{\kappa,\pi}v||_2 : v_k \geq 0, \forall k, ||v||_2 \leq 1\} < \infty.$$

En notant C le nombre de sommets (taille) dans la plus grande composante connexe du graphe, et $\mu_{kl} = \kappa(k,l)/n, \forall (k,l) \in S^2$, il est possible de montrer théoriquement que $C = \Theta(\log(n))$ si $||T_{\kappa,\pi}|| < 1$ et $C = \Theta(n)$ si $||T_{\kappa,\pi}|| > 1$. Dans cet exercice, nous nous intéressons à la transition de phase dans $\boldsymbol{\mathcal{M}}$, c'est-à-dire aux conditions de changement du régime $||T_{\kappa,\pi}|| < 1$ à $||T_{\kappa,\pi}|| > 1$.

T6. Montrez que sous l'hypothèse $\mu_{kl} = \kappa(k,l)/n, \forall (k,l) \in S^2$, l'espérance du nombre d'arêtes dans le graphe est en $\Theta(n)$.

T7. Notons H la matrice de taille $K \times K$ telle que $H_{kl} = \kappa(k,l)$ et $\mathrm{diag}(\pi)$ la matrice diagonale dont la diagonale est le vecteur π. Montrez que la transition de phase a lieu lorsque $\max_k|\lambda_k| \geq 1$ où λ_k désigne une valeur propre de la matrice $H\mathrm{diag}(\pi)$.

S7. A l'aide de la fonction *simu2* définie à l'exercice précédent, illustrez la transition de phase.

11 | Quantization

proposé par Tony Lelièvre, lelievre@cermics.enpc.fr

On considère une variable aléatoire à densité X à valeurs dans $\mathbb{R}^d$, telle que $\mathbb{E}(\|X\|^2) < \infty$, et de support $\mathbb{R}^d$. La notation $\|\cdot\|$ désigne la norme euclidienne dans $\mathbb{R}^d$. On note μ la loi de X. On pourra par exemple penser à un vecteur gaussien de moyenne nulle et de covariance l'identité. L'objectif de ce projet est d'approcher au mieux (en un sens à préciser) la variable aléatoire continue X par une variable aléatoire discrète pouvant prendre N valeurs. L'ensemble de ces N valeurs s'appelle la grille de quantization et est noté

$$\Gamma = \{x_1, \dots, x_N\} \in (\mathbb{R}^d)^N.$$

On note

$$\mathcal{G}_N(x_1, \dots, x_N) = \mathbb{E}\left(\min_{1\leq i\leq N} \|X - x_i\|^2\right)$$

la fonction de distorsion qui mesure l'erreur par rapport à la meilleure variable aléatoire (au sens L^2) à valeurs dans Γ. On considère dans ce projet le problème consistant à minimiser $\mathcal{G}_N$ sur l'ensemble des grilles $\Gamma = \{x_1, \dots, x_N\}$ possibles.

Remarque : Les techniques vues dans ce projet sont utiles pour réaliser des formules d'intégration numérique (cf. Question 4), mais aussi pour diviser un gros jeu de données en N paquets (algorithmes de *clustering* utilisés en apprentissage), ce qui correspond à un cas où la mesure initiale μ est discrète.

T1. On considère tout d'abord que $\Gamma = \{x_1, \dots, x_N\}$ est fixé et de cardinal N. Vérifier que

$$\min_{1\leq i\leq N} \|X - x_i\| = \|X - \pi_\Gamma(X)\|$$

où pour tout $y \in \mathbb{R}^d$,

$$\pi_\Gamma(y) = \sum_{i=1}^{N} x_i \, 1_{C_i(\Gamma)}(y)$$

avec $C_i(\Gamma)$ la i-ème cellule de Voronoi définie par

$$C_i(\Gamma) = \{y \in \mathbb{R}^d, \|y - x_i\| < \min_{j\neq i} \|y - x_j\|\}.$$

On rappelle que pour tout i, $C_i(\Gamma)$ est un polyèdre ouvert de $\mathbb{R}^d$, dont les bords sont des hyperplans médians entre des couples de points (x_i, x_j), pour $j \neq i$. L'ensemble des cellules de Voronoi (ou plutôt leurs fermetures) forme une partition de l'espace $\mathbb{R}^d$.

T2. Nous allons montrer que la fonction $\mathcal{G}_N$ atteint son minimum en une grille optimale $\Gamma^{(N)} = \left\{x_1^{(N)}, \dots, x_N^{(N)}\right\}$ de N points distincts.

1. *Question facultative, à admettre en première lecture.* Montrer que $\sqrt{\mathcal{G}_N} : (\mathbb{R}^d)^N \to \mathbb{R}$ est une fonction 1-Lipschitz, l'espace $(\mathbb{R}^d)^N$ étant muni de la norme $\|(x_1, \dots, x_N)\|_\infty = \max_i \|x_i\|$. En déduire que $\mathcal{G}_N$ est une fonction continue et localement lipschitzienne.
2. Soit $N = 1$. Montrer que $\mathcal{G}_1$ atteint son minimum en $x_1^{(1)} = \mathbb{E}(X)$.
3. On raisonne par récurrence, et on suppose donc qu'il existe $\Gamma^{(N)}$ qui réalise le minimum de $\mathcal{G}_N$ avec $\Gamma^{(N)} = \left\{x_1^{(N)}, \dots, x_N^{(N)}\right\}$. Soit x un point de $\mathbb{R}^d$ tel que $x \notin \Gamma^{(N)}$. On considère $\Gamma^* = \Gamma^{(N)} \cup \{x\}$. Montrer que
$$\mathbb{E}\|X - \pi_{\Gamma^*}(X)\|^2 < \mathbb{E}\|X - \pi_{\Gamma^{(N)}}(X)\|^2.$$

En déduire que $\text{card}(\Gamma^{(N)}) = N$.

4. On introduit l'ensemble

$$K_{N+1} = \left\{ x \in (\mathbb{R}^d)^{N+1},\, \mathcal{G}_{N+1}(x) \leq \mathbb{E}\|X - \pi_{\Gamma^*}(X)\|^2 \right\}.$$

Montrer que K_{N+1} est un ensemble fermé, non vide et borné. *Indication : on pourra prouver la bornitude par l'absurde.*

5. Conclure sur l'existence d'un minimum pour $\mathcal{G}_{N+1}$.
6. Montrer que

$$\lim_{N\to\infty} \mathcal{G}_N(\Gamma^{(N)}) = 0.$$

T3. On va caractériser dans cette question le minimum de $\mathcal{G}^{(N)}$ en utilisant l'équation d'Euler associée au problème de minimisation étudié dans la question précédente.

1. Soit $x = (x_1, \ldots, x_N) \in (\mathbb{R}^d)^N$ N points distincts de $\mathbb{R}^d$. Montrer que $\mathcal{G}_N$ est différentiable en x et que

$$\nabla \mathcal{G}_N(x) = 2\left(\mathbb{E}\left(1_{X\in C_i(\Gamma)}(x_i - X)\right)\right)_{1\leq i\leq N}$$

où $\Gamma = \{x_1, \ldots, x_N\}$.

2. En déduire que

$$\mathbb{E}\left(X|\pi_{\Gamma^{(N)}}(X)\right) = \pi_{\Gamma^{(N)}}(X).$$

Indication : montrer qu'il suffit de vérifier que pour tout i, $x_i^{(N)} = \mathbb{E}(X|X \in C_i(\Gamma^{(N)}))$.

On note

$$\hat{\mu}_N = \sum_{i=1}^N w_i^{(N)} \delta_{x_i^{(N)}} \text{ où } w_i^{(N)} = \mathbb{P}\left(X \in C_i(\Gamma^{(N)})\right)$$

la loi de la variable aléatoire quantizée $\hat{X}^{\Gamma^{(N)}} = \pi_{\Gamma^{(N)}}(X)$.

T4.

1. Soit $F : \mathbb{R}^d \to \mathbb{R}$ une fonction bornée et Lipschitz. Montrer que

$$\left|\mathbb{E}(F(X)) - \mathbb{E}\left(F(\hat{X}^{\Gamma^{(N)}})\right)\right| \leq [F]_{lip}\sqrt{\mathcal{G}_N(\Gamma^{(N)})}.$$

2. En déduire que $\hat{\mu}_N$ converge en loi vers μ, dans la limite $N \to \infty$.
3. *Question facultative.* Soit $F : \mathbb{R}^d \to \mathbb{R}$ une fonction bornée et telle que ∇F est Lipschitz. Montrer que

$$\left|\mathbb{E}(F(X)) - \mathbb{E}\left(F(\hat{X}^{\Gamma^{(N)}})\right)\right| \leq [\nabla F]_{lip}\,\mathcal{G}_N(\Gamma^{(N)}).$$

Indication : utiliser le résultat de la question 3.2.

On se propose dans la suite de comparer deux types d'algorithme pour calculer numériquement la quantization optimale d'une loi gaussienne en dimension $d = 1$ et $d = 2$, centrée et de covariance l'identité. On présentera pour chacune des simulations les résultats obtenus, ainsi que des courbes de convergence en fonction du nombre d'itérations. On pourra également faire varier N.

S1. On considère tout d'abord un algorithme de gradient :

$$x(k+1) = x(k) - \gamma \nabla \mathcal{G}_N(x(k))$$

qui permet de construire une suite $x(k) \in (\mathbb{R}^d)^N$ dont on peut espérer qu'elle converge vers un minimum (local) de $\mathcal{G}_N$.
Implémenter cet algorithme pour le cas de la gaussienne en dimension 1. Pourquoi est-ce compliqué d'utiliser cet algorithme en dimension plus grande ?

S2. On considère maintenant une version stochastique de l'algorithme précédent :

$$X(k+1) = X(k) - \gamma_k \left(1_{Y(k+1)\in C_i(X(k))}(X_i(k) - Y(k+1))\right)_{1\le i\le N}$$

où γ_k est une suite de pas à choisir, et $(Y(k))_{k\ge 1}$ désigne une suite i.i.d. de loi μ. Expliquer le lien entre cet algorithme et celui de la question précédente. On peut montrer que pour espérer une convergence, il faut choisir γ_k tel que $\sum_{k\ge 1}\gamma_k = \infty$ et $\sum_{k\ge 1}\gamma_k^2 < \infty$. Implémenter cet algorithme pour le cas de la gaussienne en dimension 1 et 2, en prenant $\gamma_k = 1/k$. On suivra également la convergence des poids : pour $i \in \{1,\dots,N\}$,

$$w_i(k+1) = (1-\gamma_{k+1})w_i(k) + \gamma_{k+1}1_{I_{k+1}=i}$$

où $I_{k+1} \in \arg\min_i \|X_i(k) - Y(k+1)\|$. Le poids $w_i(k)$ doit converger vers $\mu(C_i(X(\infty)))$. Discuter les difficultés numériques quand la dimension d ou le nombre de points N deviennent grands.

S3. En utilisant la relation $\mathbb{E}(X|\pi_{\Gamma^{(N)}}(X)) = \pi_{\Gamma^{(N)}}(X)$, un autre algorithme naturel est le suivant : étant donné $X(k)$, construire $X(k+1)$ en considérant : pour $i \in \{1,\dots,N\}$,

$$X_i(k+1) = \mathbb{E}(X|\pi_{\Gamma(k)}(X) = X_i(k))$$

où $\Gamma(k) = \{X_1(k),\dots,X_N(k)\}$. Implémenter cet algorithme pour le cas de la gaussienne en dimension 1. Pourquoi est-ce compliqué d'utiliser cet algorithme en dimension plus grande ?

S4. Une version stochastique de l'algorithme précédent, adaptée à la dimension grande, est la suivante : pour $i \in \{1,\dots,N\}$,

$$X_i(k+1) = \frac{\sum_{m=1}^M Y(m)1_{Y(m)\in C_i(\Gamma(k))}}{\sum_{m=1}^M 1_{Y(m)\in C_i(\Gamma(k))}}$$

où $(Y(m))_{1\le m\le M}$ désigne une suite i.i.d. de loi μ. Expliquer le lien entre cet algorithme et celui de la question précédente. Implémenter cet algorithme pour le cas de la gaussienne en dimension 1 et 2.

S5. (facultatif) Illustrer les résultats asymptotiques $N \to \infty$ de la Question T4. sur des fonctions F de votre choix.

12 | Simulation d'évènements rares

proposé par Tony Lelièvre, lelievre@cermics.enpc.fr

Soit X une variable aléatoire à valeurs dans $\mathbb{R}$, de densité $q : \mathbb{R} \to \mathbb{R}_+$ et $a \in \mathbb{R}$ un réel fixé. On s'intéresse à l'estimation de la probabilité :

$$p = \mathbb{P}(X > a)$$

que l'on suppose très petite. En pratique, a un seuil et p est une probabilité de défaillance.
Remarque : L'algorithme étudié dans ce projet est utilisé en pratique pour étudier des évènements rares, c'est-à-dire des évènements dont la probabilité est très petite (typiquement inférieure à 10^{-6}). L'algorithme permet à la fois de calculer la probabilité de l'évènement rare, mais aussi de générer des réalisations de cet évènement rare.

1 Monte Carlo

On propose d'utiliser dans un premier temps une méthode de Monte Carlo très simple, basée sur l'estimateur :

$$P_n = \frac{1}{n} \sum_{i=1}^{n} 1_{X^i > a}$$

où $(X^i)_{i \geq 1}$ est une suite i.i.d. de variables aléatoires de loi $q(x)\,dx$.

T1. Que vaut la limite presque sûre de P_n quand $n \to \infty$? Calculer la variance $\mathrm{Var}(P_n)$ en fontion de p et n. Comment se comporte l'erreur relative $\frac{\sqrt{\mathrm{Var}(P_n)}}{p}$ dans la limite $p \to 0$? Quelles sont les conséquences pratiques de ce comportement asymptotique ?

S1. Illustrer numériquement les difficultés associées à cette approche naïve dans les cas où $X \sim \mathcal{E}(2)$ et $X \sim \mathcal{N}(0,1)$ avec $a \in \{3,4,5,6\}$.

2 Méthode de splitting (fixe)

Dans cette partie, on introduit une méthode de *splitting* pour estimer p.

T2. Soit $0 = a_0 < a_1 < \ldots < a_m = a$ une subdivision de l'intervalle $[0,a]$. Montrer que

$$p = \prod_{j=1}^{m} p_j$$

où $p_j = \mathbb{P}(X > a_j | X > a_{j-1})$.

T3. Pour $j \in \{1,\ldots,m\}$, on suppose que l'on sait simuler une suite i.i.d. de variables aléatoires $(X_j^i)_{i \geq 1}$ de loi $q_{a_{j-1}}(x)\,dx$ avec, pour tout $b \in \mathbb{R}$,

$$q_b(x) = \frac{q(x) 1_{x>b}}{\int q(x) 1_{x>b}\,dx}. \tag{1}$$

On suppose de plus que les variables aléatoires $(X_j^i)_{i\geq 1, j\in\{1,\dots,m\}}$ sont indépendantes. On considère alors l'estimateur

$$\bar{P}_n = \prod_{j=1}^{m} \frac{1}{n} \sum_{i=1}^{n} 1_{X_j^i > a_j}.$$

Montrer que $\mathbb{E}(\bar{P}_n) = p$ et que, presque sûrement, $\lim_{n\to\infty} \bar{P}_n = p$. Prouver que

$$\lim_{n\to\infty} n\mathrm{Var}(\bar{P}_n) = mp^2 \left(-1 + \frac{1}{m} \sum_{j=1}^{m} \frac{1}{p_j} \right).$$

T4. Montrer que pour tout $p_1, \dots, p_m$ strictement positifs et tels que $\prod_{j=1}^m p_j = p$,

$$\frac{1}{m} \sum_{j=1}^{m} \frac{1}{p_j} \geq \frac{1}{p^{1/m}},$$

et que le minimum est atteint dans le cas $p_1 = \ldots = p_m = p^{1/m}$. Dans la suite, on note $\alpha = p^{1/m}$.

T5. Discuter les difficultés liées à l'implémentation pratique de cette méthode.

3 Splitting adaptatif

L'objectif de cette section est d'étudier un algorithme qui permet de générer les niveaux a_j de manière à réaliser approximativement $p_1 = \ldots = p_m = \alpha = 1 - 1/n$ (et donc le nombre de niveaux m sera tel que, approximativement, $(1-1/n)^m = p$). Plus précisément, on considère l'algorithme de *splitting* adaptatif suivant, qui fait évoluer un ensemble de n variables aléatoires :

- **Initialisation** : A l'itération $j = 0$, générer n variables aléatoires i.i.d. $(X_0^i)_{1\leq i\leq n}$ de loi $q(x)\,dx$ et considérer :
$$A_0 = \min_{i\in\{1,\dots,n\}} X_0^i \text{ et } I_0 = \arg\min_{i\in\{1,\dots,n\}} X_0^i.$$
- **Itérations** : A l'itération $j \geq 1$, les variables aléatoires $(X_j^i)_{1\leq i\leq n}$ sont obtenues à partir de la configuration précédente $(X_{j-1}^i)_{1\leq i\leq n}$ de la façon suivante :
  - On remplace la particule I_{j-1} par une nouvelle variable aléatoire de loi $q_{A_{j-1}}(x)\,dx$ (définie par l'équation (1)), et on ne modifie pas les autres :
$$X_j^{I_{j-1}} \sim q_{A_{j-1}}(x)\,dx \text{ et } X_j^i = X_{j-1}^i \text{ pour } i \neq I_{j-1}.$$
  - On introduit :
$$A_j = \min_{i\in\{1,\dots,n\}} X_j^i \text{ et } I_j = \arg\min_{i\in\{1,\dots,n\}} X_j^i.$$
- **Critère d'arrêt** : L'algorithme est arrêté dès que $A_j > a$. On note $\hat{J}_n = \min\{j \geq 0, A_j > a\}$ et $\hat{P}_n = \left(1 - \frac{1}{n}\right)^{\hat{J}_n}$.

Noter que la suite des niveaux $(A_j)_{j\geq 1}$ est une suite aléatoire, de même que le nombre de niveaux $\hat{J}_n$.

T6. Soit $F(y) = \mathbb{P}(X \leq y)$ la fonction de répartition de la variable aléatoire X. Soit $\Lambda(y) = -\ln(1 - F(y))$. Vérifier que Λ est une fonction continue, croissante et positive, $\lim_{y\to-\infty} \Lambda(y) = 0$ et $\Lambda(a) = -\ln p$.

T7. On note, pour $u \in (0,1]$, $F^{-1}(u) = \inf\{x \in \mathbb{R}, F(x) \geq u\}$ l'inverse généralisé de F. Montrer que $\{x, F(x) \geq u\} = [F^{-1}(u), \infty)$, $F(F^{-1}(u)) = u$. En déduire que $F(X)$ est de loi uniforme sur $(0,1)$, et que $\Lambda(X)$ est une variable aléatoire exponentielle de paramètre 1.
Remarque : L'hypothèse F continue permet de s'assurer que les I_j sont bien définis, car dans ce cas les $\arg\min_{i\in\{1,\dots,n\}} X_j^i$ *contiennent bien un unique élément presque sûrement.*

T8. Soit $b \in \mathbb{R}$ et Y une variable aléatoire de loi $q_b(y)\,dy$ (définie par l'équation (1)). Montrer que pour tout $z \in \mathbb{R}$,

$$\mathbb{P}(\Lambda(Y) > z) = \exp(\Lambda(b) - \max(z, \Lambda(b))).$$

T9. Rappeler pourquoi si X et Y sont deux variables aléatoires indépendantes, alors, pour toute fonction mesurable bornée f,

$$\mathbb{E}(f(X,Y)) = \mathbb{E}(g(X)) \text{ avec } g(x) = \mathbb{E}(f(x,Y)).$$

T10. Montrer que les variables aléatoires $(\Lambda(X_1^i) - \Lambda(A_0))_{1\le i\le n}$ sont i.i.d. de loi exponentielle de paramètre 1, et sont indépendantes de $\Lambda(A_0)$, une variable aléatoire de loi exponentielle de paramètre n.
Indication : on pourra par exemple prouver que pour tout $z_1, \dots, z_n, z$ des réels positifs,

$$\mathbb{P}(\Lambda(X_1^1) - \Lambda(A_0) > z_1, \dots, \Lambda(X_1^n) - \Lambda(A_0) > z_n, \Lambda(A_0) > z) = \exp(-(z_1 + \ldots + z_n + nz)).$$

T11. Montrer que pour tout $j \ge 1$, les variables aléatoires $(\Lambda(X_j^i) - \Lambda(A_{j-1}))_{1\le i\le n}$ sont i.i.d. de loi exponentielle de paramètre 1, et sont indépendantes de $(\Lambda(A_{k-1}) - \Lambda(A_{k-2}))_{1\le k\le j}$, des variables aléatoires indépendantes de loi exponentielle de paramètre n (avec la convention $A_{-1} = 0$).
Indication : on pourra raisonner par récurrence sur j.

T12. En déduire la loi de $\hat{J}_n$ et montrer que

$$\mathbb{E}(\hat{P}_n) = p.$$

T13. Calculer la variance de $\hat{P}_n$. Comment se comporte l'erreur relative $\frac{\sqrt{\mathrm{Var}(\hat{P}_n)}}{p}$ dans la limite $p \to 0$. Commenter ce résultat, ainsi que les difficultés liées à l'implémentation pratique de cet algorithme.

Implémentation

L'objectif de cette question est maintenant d'implémenter l'algorithme de splitting adaptatif.

S2. On suppose que $X \sim \mathcal{E}(2)$. Vérifier que pour tout b, $X + b$ a pour loi $q_b(x)\,dx$. Implémenter l'algorithme de splitting adaptatif et discuter son efficacité en comparaison de l'algorithme Monte Carlo naïf de la Section 1.

S3. On suppose que $X \sim \mathcal{N}(0,1)$. Dans ce cas, il n'y a pas de manière simple pour échantillonner $q_b(x)\,dx$. On introduit alors l'algorithme suivant, pour un paramètre $\alpha \in (0,1)$ à choisir : pour $(Y_0, \dots, Y_m)$ donnés, on construit Y_{m+1} en considérant :

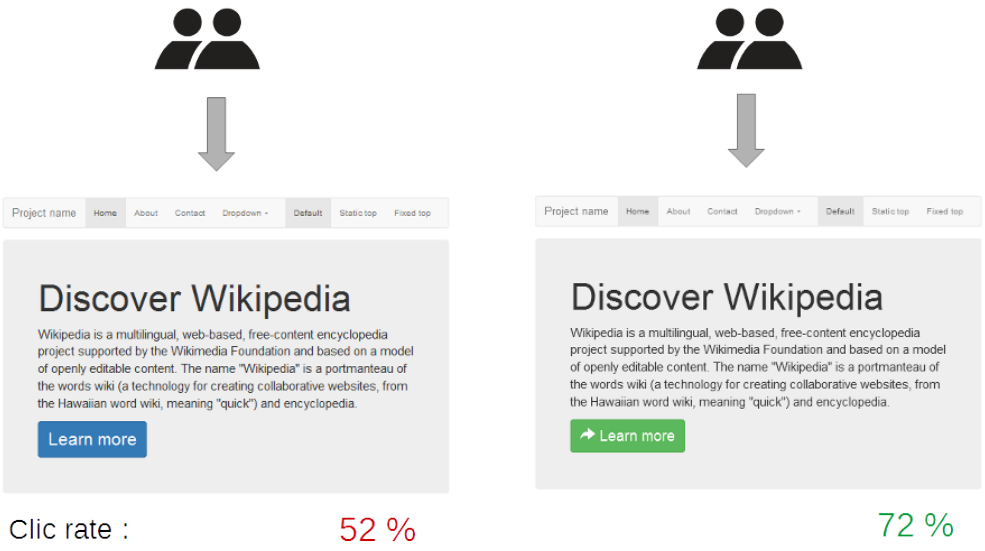
— $\tilde{Y}_{m+1} = \alpha Y_m + \sqrt{1-\alpha^2} G_m$ où $G_m \sim \mathcal{N}(0,1)$;

— $Y_{m+1} = \tilde{Y}_{m+1} 1_{\tilde{Y}_{m+1} > b} + Y_m 1_{\tilde{Y}_m \le b}$.

Montrer que si $Y_m \sim q_b(x)\,dx$, alors $Y_{m+1} \sim q_b(x)\,dx$. On admettra qu'en itérant l'algorithme précédent un nombre suffisant de fois, on obtient un échantillon Y_m "presque indépendant" de Y_0, et distribué suivant $q_b(x)\,dx$. En utilisant cet algorithme pour échantillonner $q_b(x)\,dx$ (en prenant, à l'itération j de l'algorithme, la condition initiale $Y_0 = X_{j-1}^i$ pour un des $i \ne I_{j-1}$), implémenter l'algorithme de splitting adaptatif et discuter son efficacité en comparaison de l'algorithme naïf de la Section 1 (attention au choix du paramètre α et du nombre d'itérations de la chaîne $(\tilde{Y}_m)_{m\ge 0}$ pour échantillonner $q_b(x)\,dx$.)

13 | AB Testing

proposé par Erwan Le Pennec, erwan.le-pennec@polytechnique.edu



L'objectif de ce projet est d'étudier un problème classique en marketing digital : comment choisir entre une version de référence d'un site, dite version A, et une nouvelle version, dite version B ? Cela peut correspondre, par exemple, à un nouveau design du site ou à une nouvelle politique tarifaire.
On s'intéresse ici à un choix basé sur la proportion d'utilisateur ayant un certain comportement dans la version A et dans la version B. Dans l'exemple ci-dessus, on s'intéresse à la proportion d'utilisateur cliquant sur le bouton « Learn more ». On fera dans la suite l'hypothèse qu'on souhaite choisir la version ayant la plus grande proportion.

1 Loi de la proportion observée d'utilisateurs

On suppose ici que l'on a observé le comportement de n utilisateurs d'une des versions et l'on souhaite modéliser la proportion observée chez les utilisateurs ayant eu le comportement souhaité.
On fera l'hypothèse que tous les utilisateurs se comportent de manière indépendante les uns des autres.

T1. On mesure le comportement d'un utilisateur par une variable X qui vaut 1 si l'utilisateur a le comportement souhaité et 0 sinon. Montrer que l'on peut modéliser le comportement d'un utilisateur par une variable de Bernoulli X de paramètre p,

$$\mathbb{P}(X = 1) = p \quad \text{et} \quad \mathbb{P}(X = 0) = 1 - p.$$

T2. Si l'on note X_i le comportement du $i$ème utilisateur, vérifier que

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

est une variable aléatoire calculable à partir des observations des utilisateurs de moyenne p.

T3. Montrer qu'il ne s'agit de rien d'autre que de la proportion de comportement souhaité parmi les utilisateurs observés.

T4. Vérifier que

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to p$$

S1. Illustrer de manière numérique ce comportement. Pour les illustrations numériques, on prendra $p = .72$.

2 Vitesse de convergence de la proportion observée

Dans la section précédente, on a vérifié que $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ s tendait vers p lorsque n tendait vers l'infini. On souhaite préciser un peu plus cette convergence.

T5. Vérifier que

$$\sqrt{n}\,(\hat{p} - p) \to \mathcal{N}(0, p(1-p))$$

et en déduire que

$$\mathbb{P}\left(\hat{p} \geq p + \frac{\epsilon}{\sqrt{n}}\right) \to \mathbb{P}\left(N \geq \frac{\epsilon}{\sqrt{p(1-p)}}\right) \quad \text{et} \quad \mathbb{P}\left(|\hat{p} - p| \geq \frac{\epsilon}{\sqrt{n}}\right) \to \mathbb{P}\left(|N| \geq \frac{\epsilon}{\sqrt{p(1-p)}}\right)$$

où N est une variable gaussienne centrée réduite.

S2. Illustrer de manière numérique ces deux convergences.

T6. Démontrer que

$$\mathbb{P}(N \geq \epsilon) \leq e^{-\frac{\epsilon^2}{2}} \quad \text{et} \quad \mathbb{P}(|N| \geq \epsilon) \leq 2e^{-\frac{\epsilon^2}{2}}$$

et donc que

$$\lim \mathbb{P}\left(\hat{p} \geq p + \frac{\epsilon}{\sqrt{n}}\right) \leq e^{-\frac{\epsilon^2}{2p(1-p)}} \quad \text{et} \quad \lim \mathbb{P}\left(|\hat{p} - p| \geq \frac{\epsilon}{\sqrt{n}}\right) \leq 2e^{-\frac{\epsilon^2}{2p(1-p)}}$$

T7. En utilisant l'inégalité de Bienaymé-Tchebychev, montrer que

$$\mathbb{P}\left(|\hat{p} - p| \geq \frac{\epsilon}{\sqrt{n}}\right) \leq \frac{p(1-p)}{\epsilon^2}$$

S3. Vérifier numériquement que cette inégalité est respectée. Pourquoi semble-t-elle améliorable ?

T8. Le théorème de Hoeffding nous dit que si $Z_1, \ldots, Z_n$ sont des variables aléatoires indépendantes tels que $Z_i \in [b_i, a_i]$ alors

$$\mathbb{P}(\sum_{i=1}^{n}(Z_i - \mathbb{E}(Z_i)) > \epsilon) \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n}(a_i - b_i)^2}}$$

En déduire que

$$\mathbb{P}\left(\hat{p} > p + \frac{\epsilon}{\sqrt{n}}\right) \leq e^{-2\epsilon^2} \quad \text{et} \quad \mathbb{P}\left(|\hat{p} - p| \geq \frac{\epsilon}{\sqrt{n}}\right) \leq e^{-2\epsilon^2}$$

T9. Le théorème de Bernstein dit que si $\text{var}(Z_i) \leq \sigma_i^2$ alors

$$\mathbb{P}\left(\sum_{i=1}^{n}(Z_i - \mathbb{E}(Z_i)) > \epsilon\right) \leq e^{-\frac{\epsilon^2}{\sum_{i=1}^n \sigma_i^2 + 2B\epsilon/(3\sqrt{n}}}$$

où $B = \max_i \max(b_i - \mathbb{E}(Z_i), \mathbb{E}(Z_i) - a_i) \leq \max_i(b_i - a_i)$. En déduire que

$$\mathbb{P}\left(\hat{p} - p \geq \frac{\epsilon}{\sqrt{n}}\right) \leq e^{-\frac{\epsilon^2}{2p(1-p)+2/3\frac{\epsilon}{\sqrt{n}}\max(p,1-p)}} \quad \text{et} \quad \mathbb{P}\left(|\hat{p} - p| \geq \frac{\epsilon}{\sqrt{n}}\right) \leq e^{-\frac{\epsilon^2}{2p(1-p)+2/3\frac{\epsilon}{\sqrt{n}}\max(p,1-p)}}$$

S4. Vérifier numériquement que ces comportements sont vérifiés.

T10. Quel est l'avantage pratique de l'inégalité de Hoeffding par rapport à celle de Bernstein ?

S5. Comparer ces inégalités avec celles obtenues via la limite gaussienne.

T11. (facultatif) Démontrer ces deux théorèmes.

3 Comparaison entre deux proportions

Le problème originel n'est pas exactement d'estimer les proportions p_A et p_B des utilisateurs ayant le comportement souhaité dans les versions A et B mais plutôt de voir si p_B est *vraiment* plus grand que p_A. En pratique, le risque d'un changement n'étant pas négligeable, on souhaite conserver le modèle A lorsque $p_B \leq p_A + \delta$ et basculer vers B dans le cas contraire.
On note n_A le nombre d'utilisateurs ayant vu la version A et n_B celui ayant vu la version B et on propose de choisir entre les versions A et B de la manière suivante :

— si $\hat{p}_B \leq \hat{p}_A + \delta + \sqrt{\frac{-\log\gamma}{2}}\left(\frac{1}{\sqrt{n_A}} + \frac{1}{\sqrt{n_B}}\right)$ alors on conserve A

— sinon on choisit B

où $\gamma \in (0,1)$ est un paramètre à préciser.

T12. Montrer que

$$\mathbb{P}\left(\hat{p}_B - \hat{p}_A \geq p_B - p_A + \epsilon\left(\frac{1}{\sqrt{n_A}} + \frac{1}{\sqrt{n_B}}\right)\right) \leq e^{-2\epsilon^2}$$
$$\text{et} \quad \mathbb{P}\left(\hat{p}_B - \hat{p}_A \leq p_B - p_A - \epsilon\left(\frac{1}{\sqrt{n_A}} + \frac{1}{\sqrt{n_B}}\right)\right) \leq e^{-2\epsilon^2}.$$

T13. En déduire que si $p_B \leq p_A + \delta$ alors

$$\mathbb{P}\left(\hat{p}_B - \hat{p}_A \geq \delta + \epsilon\left(\frac{1}{\sqrt{n_A}} + \frac{1}{\sqrt{n_B}}\right)\right) \leq e^{-2\epsilon^2}.$$

et que si $p_B \geq p_A + \delta + \Delta$ alors

$$\mathbb{P}\left(\hat{p}_B - \hat{p}_A \leq \delta + \Delta - \epsilon\left(\frac{1}{\sqrt{n_A}} + \frac{1}{\sqrt{n_B}}\right)\right) \leq e^{-2\epsilon^2}.$$

T14. Vérifier que si $p_B \leq p_A + \delta$ alors la probabilité de choisir B est inférieure à γ.

T15. Montrer que si $\Delta \geq \sqrt{\frac{-\log\gamma}{2}}\left(\frac{1}{\sqrt{n_A}} + \frac{1}{\sqrt{n_B}}\right)$ la probabilité de choisir B est inférieure à

$$e^{-2\left(\Delta/\left(\frac{1}{\sqrt{n_A}}+\frac{1}{\sqrt{n_B}}\right) - \sqrt{-\log(\gamma)/2}\right)^2}.$$

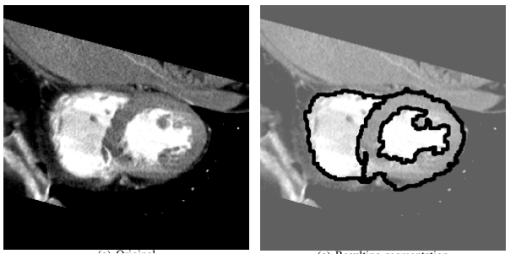
S6. Vérifier numériquement ces comportements. On pourra prendre $p_A = .52$ et $p_B = .72$ et faire l'hypothèse que $\alpha = n_B/(n_A + n_B)$ où α est un paramètre pris entre 0 et 1 que l'on garde constant en faisant varier $n = n_A + n_B$.

T16. Supposons que l'on pense que le gain de version B est supérieur de 10% à la version A , qu'on souhaite s'assurer que le gain soit au moins de 1% et qu'on a choisit $\gamma = 5\%$, quelle est le nombre minimal d'observations nécessaires si l'on affecte 30% des observations à la version B pour que l'on puisse choisir la version B avec une probabilité supérieur à 10% ?

S7. Implémenter une fonction donnant ce nombre en fonction des différents paramètres.

14 | Marche aléatoire sur une grille et segmentation

proposé par Erwan Le Pennec, erwan.le-pennec@polytechnique.edu



L'objectif de ce projet est d'étudier quelques propriétés des marches aléatoires sur les grilles régulières (un cas particulier de graphe) et d'utiliser ces résultats dans un algorithme de segmentation d'image à partie de *graines* fixées par l'utilisateur.

1 Grille et graphe

La grille $\mathcal{G}$ à 8 voisins de taille $n_1 \times n_2$ est définie comme le graphe non orienté dont les sommets sont les points (x_1, x_2) avec x_1 un entier compris entre 1 et n_1 et x_2 un entier compris entre 1 et n_2 et les arrêtes relient les *voisins*, i.e. les couples de sommets distincts (x_1, x_2) et (x'_1, x'_2) tels que $|x_1 - x'_1| \leq 1$ et $|x_2 - x'_2| \leq 1$.

T1. Montrer que ce graphe possède $n_1 \times n_2$ sommets.

T2. Combien de sommets (en fonction de n_1 et n_2) ont, respectivement, 8, 5 et 3 voisins ?

T3. En déduire que le nombre total d'arrêtes est de $4n_1n_2 - 3n_1 - 3n_2 + 2$.

T4. Démontrer que ce graphe est connexe, i.e. qu'on peut relier par un chemin suivant des arrêtes tout couple de sommets.

T5. On dit qu'un graphe est bipartite si l'on peut partitionner ses sommets en deux groupes de sorte que toutes les arrêtes relient des sommets dans des groupes différents. Montrer que la grille considérée n'est pas bipartite mais qu'elle l'est si on enlève les arrêtes diagonales (i.e. pour lesquels $|x_1 - x'_1| = |x_2 - x'_2| = 1$).

2 Marche aléatoire sur un graphe

Soit $\mathcal{G}$ un graphe fini non orienté connexe, on note S_i ses n sommets, $V(S_i)$ les voisins du sommet S_i et $A(S_i, S_j) = A(S_j, S_i)$ l'arrête reliant les sommets S_i et S_j.
On définit une marche aléatoire sur $\mathcal{G}$ partant d'un sommet X_0 par la dynamique suivante :

$$\forall t \in \mathbb{N}, \quad \mathbb{P}(X_{t+1} = A | X_t) = \frac{W(A, X_t)}{\sum_{A' \in V(X_t)} W(A', X_t)}$$

où $W(A_i, A_i) = W(A_j, A_i)$ est un poids strictement positif si il y a une arrête entre A_i et A_j et égal à 0 sinon.

T6. Vérifier que $\mathbb{P}(X_{t+1} = A|X_t) = 0$ si $A \notin V(X_t)$.

S1. Proposer un algorithme de simulation dans le cas d'une grille $n_1 \times n_2$ avec des poids W, uniformes, i.e. égaux à 1 si il y a une arrête et 0 sinon.

S2. Modifier cet algorithme pour prendre en compte des poids quelconques.

3 Loi limite

On s'intéresse au comportement limite de cette marche aléatoire, i.e. si elle existe

$$\lim_{t\to\infty} \mathbb{P}(X_t = A|X_0).$$

La matrice de transition P de la marche aléatoire est définie par

$$P_{i,j} = \mathbb{P}(X_{t+1} = A_j|X_t = A_i) = \frac{W(A_i, A_j)}{\sum_{A'\in V(A_i)} W(A_i, A')}$$

T7. Démontrer que

$$\mathbb{P}(X_t = A|X_0) = P^t e_{X_0}$$

où P^t est la matrice P élevée à la puissance t et e_{X_0} est un vecteur de taille n avec des coefficients nuls sauf en X_0 où le coefficient vaut 1.

S3. À l'aide d'une simulation numérique, vérifier que, sur une grille de taille 8×8 avec des poids uniformes, la loi limite semble exister et qu'elle ne semble pas dépendre du point initial.

T8. Montrer que P est

1. une matrice stochastique, i.e. dont toutes les lignes contiennent des coefficients positifs se sommant à 1,
2. une matrice irréductible, i.e. telle que pour tout couple (i, j), il existe k tel $(P^k)_{i,j} > 0$

T9. En utilisant le théorème de Perron-Frobenius, déduire qu'il existe un unique vecteur π positif sommant à 1 tel que $P\pi = \pi$.

S4. Comment le déterminer numériquement ?

T10. En admettant que sous une hypothèse supplémentaire sur P, on obtient

$$\lim_{t\to\infty} (P^t)_{i,j} = \pi_i$$

démonter que sous cette hypothèse, valide dans le cas de la grille $\mathcal{G}$ à 8 voisins,

$$\mathbb{P}(X_t = A_i|X_0) \to \pi_i$$

T11. (facultatif) Démontrer le résultat sur la limite de P^t.

T12. (facultatif) Démontrer que sous l'hypothèse de convergence pour P^t, le nombre $N_t(i)$ de visite de A_i avant le temps t vérifie

$$\mathbb{E}(N_t(i)/t) = \pi_i.$$

4 Temps d'accès

On définit le temps de retour $R(A_i)$ comme la moyenne du nombre d'étapes pour qu'une marche aléatoire partant de A_i revienne en A_i et le temps d'accès $H(A_i, A_j)$ comme la moyenne du nombre d'étapes pour qu'une marche aléatoire partant de A_i arrive en A_j.

T13. Démontrer que pour tout couple (A_i, A_j), il existe un chemin de taille inférieur à n reliant A_i à A_j. En déduire qu'il existe $\epsilon > 0$ tel que $\forall k \in \mathbb{N}$

$$\mathbb{P}(\exists t \in [kn+1, k(n+1)], X_t = A_j | X_0 = A_i) \geq \epsilon$$

T14. Montrer que cela implique

$$\mathbb{P}(H(A_i, A_j) \geq kn) \leq (1-\epsilon)^k$$

puis en déduire

$$H(A_i, A_j) \leq \frac{n}{\epsilon}$$

T15. On note $T_n(i)$ l'instant de la $n$ième visite en i. Montrer que $T_{n+1}(i) - T_n(i)$ est d'espérance $R(A_i)$ et que ces différences sont i.i.d.. En déduire que

$$\frac{T_n(i)}{n} \xrightarrow{P} R(A_i).$$

T16. Vérifier que

$$T_n(i) \leq t \Rightarrow N_{t'}(i) \geq n, \forall t' \geq t. \quad \text{et que} \quad N_t(i) \leq n \Rightarrow T_{n'}(i) \geq t, \forall n' \geq n.$$

T17. En combinant ces résultats avec les convergences en probabilités de $T_n(i)/n$ et $N_t(i)/n$ vers respectivement $R(A_i)$ et π_i, démontrer que $R(A_i) = \frac{1}{\pi_i}$.

S5. Par simulation numérique, estimer $H(A_i, A_j)$ pour une grille de taille $8x8$ avec des poids uniformes.

T18. Vérifier que

$$H(A_i, A_j) = \begin{cases} 0 & \text{si } i = j \\ 1 + \sum_{k, A_k \in V(A_i)} P_{i,k} H(A_k, A_j) & \text{sinon} \end{cases}$$

T19. En déduire que $H_j = H(A_i, A_j)$ satisfait le système d'équation linéaire

$$\begin{aligned} ((I-P)H_j)_i &= 1 \quad (i \neq j) \\ H_{j,j} &= 0 \end{aligned}$$

T20. En notant que $R(A_i) = 1 + \sum_{k, A_k \in V(A_i)} P_{i,k} H(A_k, A_j)$, vérifier que

$$(I-P)H_j = 1 - R(A_j)e_j$$

S6. Comparer les solutions obtenus par simulation et par résolution des systèmes linéaires en terme de précision et de vitesse.

5 Segmentation

Leo Grady a proposé un algorithme de segmentation d'image en K groupes à partir de graines, au moins un pixel marqué par l'utilisateur dans chacun des groupes, dont le principe est le suivant :

a) Associer à une image I de taille $n_1 \times n_2$ la grille $\mathcal{G}$ à 8 voisins de taille $n_1 \times n_2$ avec des poids

$$W((x_1,x_2),(x_1',x_2')) = \begin{cases} 0 & \text{si } (x_1',x_2') \notin V((x_1,x_2)) \\ e^{-\kappa(I(x_1,x_2)-I(x_1',x_2'))^2} & \text{sinon} \end{cases}$$

où κ est un paramètre à régler.

b) Laisser l'utilisateur *marquer* au moins un pixel pour chacune des groupes d'intérêt.

c) Calculer le temps d'accès de chacun des points non marqués à chacun des points marqués.

d) Associer à chaque point non marqué son point marqué le plus proche en terme de temps d'accès et lui attribuer le groupe correspondant.

T21. Montrer qu'on peut déterminer le temps d'accès à chacun des points marqués en résolvant autant de systèmes linéaires que de points marqués.

S7. Implémenter cet algorithme en utilisant les résultats de la section précédente sur des images de taille 64×64.

T22. Montrer qu'on peut diminuer le nombre de systèmes linéaire à résoudre au nombre de groupes (indépendamment du nombre de points marqués dans chaque groupe).

S8. Implémenter cette version améliorée et comparer avec la précédente en terme de vitesse.

T23. Comment choisir κ ?

15 | Composante géante des graphes aléatoires

proposé par Laurent Massoulié, laurent.massoulie@inria.fr

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.
Dans le cas où $p = \lambda/n$ pour une constante $\lambda > 0$ fixe, un résultat dû à Erdős et Rényi établit que si $\lambda < 1$, pour une constante $f(\lambda) > 0$, alors

$$\lim_{n\to\infty} \mathbb{P}(X_1 \leq f(\lambda)\ln(n)) = 1,$$

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

$$\lim_{n\to\infty} \mathbb{P}(X_1/n > \epsilon) = 0, \tag{1}$$

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 $\epsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}(|X_1/n - [1 - p_{ext}(\lambda)]| \leq \epsilon) = 1 \tag{2}$$

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 $g(\lambda) > 0$,

$$\lim_{n\to\infty} \mathbb{P}(X_2 \leq g(\lambda)\ln(n)) = 1 \tag{3}$$

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

$$z = \phi(z), \tag{4}$$

où $\phi(z) = e^{-\lambda(1-z)}$ est la fonction caractéristique de la variable de Poisson de paramètre λ. La propriété (1) est en fait un cas particulier de (2), puisque $p_{ext}(\lambda) = 1$ pour $\lambda \leq 1$.

S1. Ecrire un programme qui, étant donné un graphe G non orienté avec ensemble de sommets $[n]$, décrit par sa matrice d'adjacence A, calcule la taille X_1 de sa plus grande composante connexe.

S2. Ecrire un programme qui produit une approximation de $p_{ext}(\lambda)$ pour $\lambda > 1$ donné.

S3. Pour n pas trop grand (on commencera pour $n = 100$, et selon la vitesse d'exécution on pourra tester de plus grands multiples de 100), 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 en utilisant les programmes des questions précédentes la moyenne empirique des tailles observées pour X_1/n à la valeur prédite en théorie, soit $1-p_{ext}(\lambda)$. 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 :

$$\forall \sigma \in \{+, -\}, \quad p_{\sigma,ext} \in [0, 1],\ p_{\sigma,ext} = e^{-\lambda_{\sigma+}(1-p_{+,ext})} e^{-\lambda_{\sigma-}(1-p_{-,ext})}. \tag{5}$$

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 :

$$\lim_{n\to\infty} \frac{1}{n} X_+ = r(1 - p_{+,ext}),\ \lim_{n\to\infty} \frac{1}{n} X_- = (1 - r)(1 - p_{-,ext}). \tag{6}$$

S4. Ecrire un programme pour obtenir par itération, en démarrant avec le vecteur nul, une solution approchée de l'équation de point fixe (5).

S5. Ecrire un programme qui, à partir d'un graphe G décrit par sa matrice d'adjacence, et le vecteur $\{\sigma_u\}$ des types de ses sommets, supposés appartenir à $\{+, -\}$, retourne les nombres X_+, X_- de sommets de chaque type dans la plus grande composante connexe du graphe.

S6. Pour $r = 0.3$, $M_{++} = M_{--} = 4$, $M_{+-} = M_{-+} = 1$, et pour n pas trop grand (on commencera pour $n = 100$, et selon la vitesse d'exécution on pourra tester de plus grands multiples de 100), simuler une vingtaine de répliques du graphe aléatoire multi-types décrit ci-dessus. Comparer, en utilisant les programmes des questions précédentes, les moyennes empiriques des tailles observées pour X_+/n, X_-/n aux valeurs prédites en théorie données en (6). La prédiction théorique est-elle précise pour les paramètres considérés ?

16 | Connectivité des graphes aléatoires

proposé par Laurent Massoulié, laurent.massoulie@inria.fr

On s'intéresse à des 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.
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 :

$$\lim_{n\to\infty} \mathbb{P}\left(\mathcal{G}(n,(\ln(n)+c)/n) \text{ connecté}\right) = e^{-e^{-c}}. \tag{1}$$

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 $Z_u = \mathbf{1}_{u \text{ sommet isolé de } G}$. 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

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

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

$$\lim_{n\to\infty} \mathbb{E}(X^{\underline{k}}) = e^{-kc}. \tag{2}$$

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

$$\lim_{n\to\infty} \mathbb{P}(X=0) = e^{-e^{-c}}. \tag{3}$$

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 :

$$\binom{n}{k}\mathbb{P}([k] \text{ composante connexe de } \mathcal{G}(n,p)) = \binom{n}{k}\mathbb{P}(\mathcal{G}(k,p) \text{ connecté})(1-p)^{k(n-k)}. \tag{4}$$

On peut alors majorer grossièrement $\mathbb{P}(\mathcal{G}(k,p) \text{ connecté})$ 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 (4) 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

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

T4. Etablir que cette dernière propriété, combinée avec le résultat (3) de la première question, implique le résultat d'Erdős-Rényi (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 (1) d'Erdős-Rényi fournit-il une bonne approximation de $\mathbb{P}(\mathcal{G}(n,p) \text{ connecté})$ 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

$$\lim_{n\to\infty} \mathbb{P}(\mathcal{G}'(n,p) \text{ connexe}) = \lim_{n\to\infty} \mathbb{P}(X_{out} = X_{in} = 0) = e^{-2e^{-c}}. \tag{5}$$

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 (5) en simulant une vingtaine de graphes $\mathcal{G}'(n,p)$ pour les paramètres correspondants.

17 | Comment se prémunir contre les aléas malheureux ?

proposé par Thibaut Mastrolia, thibaut.mastrolia@polytechnique.edu

La théorie de la ruine est considérée comme une branche de la gestion des risques en assurance. Ceci correspond à l'étude de la probabilité qu'un évènement défavorable ait lieu pour un assuré sous contrat d'assurance. Ces accidents impliquent donc des risques que l'assureur doit évaluer afin de verser des indémnité à l'assuré.

Nous modélisons typiquement un accident comme un saut intervenant (négativement) dans la richesse de l'assureur. Dans une première partie, nous analysons la modélisation d'accidents assimilés à des processus à sauts. Puis, nous étendrons cette études à la modélisation du coût total lié à ces accidents. Enfin, nous verrons comment calibrer la réserve initiale de l'assurance afin de minimiser sa probabilité de ruine, *i.e.* la probabilité que sa richesse devienne négative.

1 De la modélisation d'accidents aléatoires

La modélisation de l'arrivée d'évèments aléatoires tels que des incidents associés à un projet risqué, des tremblements de terre ou encore l'arrivée de clients dans un magasin, doit respecter un certains nombre de critères afin d'être le plus réalistes que possible. Les processus de Poisson, que nous allons étudier dans cette section, sont un des moyens de répondre à cette modélisation.

On modélise la survenance d'évènements aléatoire par la suite $(\tau_n)_n$ de variables aléatoires i.i.d. de loi exponentielle de paramètre $\lambda > 0$ fixé. On rappelle que la loi exponentielle de paramètre λ est de densité f_λ définie par

$$\gamma_\lambda(x) = \lambda e^{-\lambda x}\mathbf{1}_{x\geq 0},\ x \in \mathbb{R}.$$

On note $T_1 = \tau_1$ la date de survenance du premier évènement et $T_n := \sum_{i=1}^n \tau_n$, $n > 0$ l'instant auxquel le $n$ième évènement intervient.

T1. Montrer que T_n admet une densité donnée par

$$\gamma_{\lambda,n}(x) = \frac{\lambda^n}{(n-1)!}x^{n-1}e^{-\lambda x}\mathbf{1}_{x\geq 0},\ x \in \mathbb{R}.$$

T2. On définit un processus de Poisson N comme une variable aléatoire en tout temps t définie par

$$N_0 = 0,\ N_t = \sum_{i=1}^{+\infty} \mathbf{1}_{T_n\leq t}.$$

En remarquant que $N_t = \sup\{n \in \mathbb{N},\ T_n \leq t\}$, donner l'interprétation de N_t.

T3. Donner une condition nécéssaire et suffisante telle que

$$\mathbb{P}(N \text{ ne fait que des sauts de taille } 1) > 0.$$

En déduire

$$\mathbb{P}(N \text{ ne fait que des sauts de taille } 1) = 1.$$

Indication : remarquer que

$$\{N \text{ ne fait que des sauts de taille } 1\} = \{\forall i \geq 1, T_i < T_{i+1}\} = \cap_{i\geq 1}\{\tau_i > 0\}$$

T4. Montrer que N_t admet une loi de Poisson de paramètre λt.
Indication : On remarque que $\{N_t = n\} = \{T_n \leq t < T_{n+1}\}$.

On admet que $(N_t)_{t\geq 0}$ est un processus homogène à accroissements indépendants et stationnaires :

1. (stationarité) Si $0 \leq s \leq t$, alors $N_t - N_s$ a même loi que N_{t-s}. Ainsi $N_t - N_s$ suit la loi de Poisson de paramètre $\lambda(t-s)$.
2. (indépendance)Soit $0 < t_1 < t_2 < \cdots t_n$ alors les variables aléatoires $N_{t_1}, N_{t_2} - N_{t_1}, \cdots, N_{t_n} - N_{t_{n-1}}$ sont indépendantes.

T5. Soit $n \in \mathbb{N}$, $t > 0$ et $U_1, U_2, \cdots, U_n$ une suite de n variables aléatoires indépendantes de loi uniforme sur $[0,t]$. On note $U_{(1)}, U_{(2)}, \cdots, U_{(n)}$ les statistiques d'ordre associées à $(U_1, \cdots, U_n)$, c'est à dire que $(U_{(1)}, U_{(2)}, \cdots, U_{(n)})$ est le vecteur $(U_1, U_2, \cdots, U_n)$ ré-ordonné dans l'ordre croissant. En particulier

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

Montrer que le vecteur aléatoire $(U_{(1)}, U_{(2)}, \cdots, U_{(n)})$ a une densité que l'on calculera.

T6. Montrer que la loi conditionnelle de $(T_1, \cdots, T_n)$ sachant $N_t = n$ est la même que celle de $(U_{(1)}, U_{(2)}, \cdots, U_{(n)})$.

S1. En utilisant la question précédente, simuler un processus de Poisson avec $\lambda = 5$ et $t = 1$.

2 Coût des indémnités

On se donne un processus de Poisson de paramètre λ comme défini précédemment. Soit $(J_n)_n$ une suite de v.a. i.i.d. et indépendantes de toutes les variables aléatoires N_t, $t \geq 0$.

On considère le processus suivant, dit *processus de Poisson composé* :

$$\forall t \geq 0,\ C_t = \sum_{i=1}^{N_t} J_i.$$

On admet que ce processus (C_t) est aussi un processus à accroissements indépendants et stationnaires.

T7. Que représentent J_i et C_t ?

T8. Montrer que C est un processus de Poisson (simple) si et seulement si

$$J_1 \sim \mathcal{B}er(p),\ p \in (0,1),\ t \geq 0,$$

où $\mathcal{B}er(p)$ est une loi de Bernoulli de paramètre p.
Indication : autrement dit, $\mathbb{P}[C$ *ne fait que des sauts de tailles* 0 *ou* $1] = 1$.

T9. On suppose que J_1 a une variance finie et on pose :

$$\mu = E[J_1] \text{ et } \sigma^2 = \mathrm{Var}(J_1).$$

Soit $t \geq 0$. Calculer l'espérance et la variance de C_t en fonction de μ, σ^2, λ et t.

S2. En s'inspirant de la question 7, simuler une trajectoire de C_t avec $t = 1$, $\lambda = 5$ en supposant que les J_i ont pour loi la loi normale centrée de variance 5.

3 Probabilité de ruine

Les processus précédemment introduit sont particulièrement pertinents pour une compagnie d'assurance qui assure un certain type de risques. On suppose que les accidents qu'elle assure surviennent selon un processus de Poisson N_t de paramètre $\lambda > 0$ (une date initiale $t = 0$ étant choisie) et J_i représente le montant de l'indemnité du i-ème accident que doit verser la compagnie. On suppose que les (J_i) sont des v.a. positives i.i.d. et indépendantes du processus (N_t) si bien que la somme totale des indemnités versées par la compagnie entre les instants 0 et t est donnée par le processus de Poisson composé :

$$C_t = \sum_{i=1}^{N_t} J_i.$$

On appelle R_t la richesse (ou réserve) de la compagnie d'assurance à l'instant t. On suppose qu'elle est de la forme

$$R_t = u + ct - C_t = u + ct - \sum_{i=1}^{N_t} J_i.$$

$u > 0$ représente la richesse initiale et le terme ct correspond à une rentrée régulière d'argent (on suppose donc $c > 0$).
La question naturelle pour la compagnie d'assurance est de savoir comment choisir u de façon à ce que la *probabilité de ruine* soit en dessous d'un seuil donné.
Pour cela on introduit :

- $\varphi(u;t) = \mathbb{P}(\exists s \in [0,t], R_s < 0)$, la probabilité de ruine avant l'instant t associée à une richesse initiale u.
- $\varphi(u) = \mathbb{P}(\exists s > 0, R_s < 0)$, le probabilité de ruine totale avec richesse initiale u.

T10. Montrer que $\varphi(u) = \lim_{t\to+\infty} \varphi(u,t)$.

T11. On suppose de plus que la condition de "chargement de sécurité" ("safety-load condition") est vérifiée :

$$c - \lambda\mu > 0.$$

En appliquant la loi forte des grands nombres, montrer qu'alors

$$\lim_{t\to+\infty} R_t = +\infty \ p.s.$$

Interpréter ce résultat.

S3. Inversement lorsque $c - \lambda\mu < 0$, le processus R_t converge vers $-\infty$ presque sûrement. En prenant maintenant des J_i de loi normale réduite et d'espérance $\mu = 1$, $\lambda = 5$ et $c = 7$ simuler une trajectoire de R_t et donner le moment où la ruine a lieu.

Référence : Soren Asmussen, *Ruin Probabilities.*

18 | Microcrédit de projet d'investissement risqué

proposé par Thibaut Mastrolia, thibaut.mastrolia@polytechnique.edu

Dans ce projet nous étudions un modèle de microcrédit introduit par G. Tedeshi dont l'obtention d'un prêt dépend de la situation financière du projet d'investissement. La première partie de ce projet présente un modèle très simple. La deuxième partie étand ce premier cas à un modèle avec exclusion de crédit en cas de refus de prêt.

1 Un exemple de microcrédit par changement d'état

Commençons avec un exemple très simple où un agent financier peut se trouver dans deux états en tout temps t :

— candidat à un prêt, nous notons alors cet état A comme "applicant"

— receveur d'un prêt, nous notons cet état B comme "beneficiary".

Nous supposons que si l'agent financier bénéficie d'un prêt au temps k, il rembourse son crédit en cas de succès du projet avec probabilité $\beta \in [0,1]$. Son prêt est alors automatiquement reconduit la période suivante $k+1$, donc reste dans l'état B. Si l'agent fait faillite (avec probabilité $1-\beta$) et n'est pas en mesure de rembourser son prêt, il perd sa reconduction automatique et devient alors demandeur (donc dans l'état A) en $k+1$.

En temps que candidat à un prêt (c'est à dire si l'agent est dans l'état A), il obtient un prêt avec probabilité $\alpha \in [0,1]$.

On note $(E_k)_{k\in\mathbb{N}}$ une suite de variables aléatoires à valeurs dans $\{A,B\}$.

T1. On définit pour tout entier $k \geq 0$

$$P_{1,1} = \mathbb{P}(E_{k+1} = A | E_k = A),\ P_{2,1} = \mathbb{P}(E_{k+1} = A | E_k = B),$$

$$P_{1,2} = \mathbb{P}(E_{k+1} = B | E_k = A),\ P_{2,2} = \mathbb{P}(E_{k+1} = B | E_k = B).$$

Calculer $P_{i,j}, i,j \in \{1,2\}$ en fonction de α, β et écrire la matrice P. Que représente-t-elle ?

S1. On attribue la valeur 0 lorsque l'agent est dans l'état A et 1 à l'état B. On suppose que $\mathbb{P}(E_0 = A) = \frac{1}{2}$. Soit $k = 1000$, $\alpha = \beta = \frac{1}{2}$ tracer une trajectoire[1] possible de $(E_k)_{k\geq 0}$. Calculer $\frac{1}{1000}\sum_{k=0}^{1000} E_k$ sur cette trajectoire. Ce résultat était-il prévisible ? (on montrera en particulier que dans ce cas, $(E_k)_{k\geq 0}$ est une suite de variables aléatoires identiquement distribuées et indépendantes.

T2. On considère maintenant π_0 le vecteur ligne $(\mathbb{P}(E_0 = A), \mathbb{P}(E_0 = B))$ supposé connu. On pose

$$\pi_k = (\mathbb{P}(E_k = A), \mathbb{P}(E_k = B)).$$

Que signifie π_k ?

1. On appelle trajectoire associée à une suite $(X_n)_{n\geq 0}$ de variable aléatoire l'application $n \longmapsto X_n(\omega)$ où $\omega \in \Omega$ est fixé. Ici, cela correspond à tirer aléatoirement E_0 puis les états suivants en fonction des règles de changement d'état données en énoncé.

On considère maintenant $\beta = \frac{3}{4}$ et $\alpha = \frac{1}{2}$.

T3. On suppose $\pi_0 = (\frac{1}{2}, \frac{1}{2})$. Tracer 100 trajectoire de $(E_k)_{1\leq k\leq 500}$ en attribuant toujours la valeur 0 à l'état A et 1 à l'état B. Comment approcheriez vous π_{500} dans ce cas ? Qu'obtenez vous ?

S2. En utilisant la méthode précédente, donner le graphe de π_{500} en fonction de π_0 quelconque dans $[0,1]^2$ tel que la somme des composantes vaut 1. Que constatez vous ?

T4. Montrer que

$$\pi_k = \pi_0 P^k.$$

T5. Calculer $\pi_* = (\pi_*^1, \pi_*^2)$ à valeurs dans $[0,1]^2$ tel que

$$\pi_* = \pi_* P.$$

T6. Etudier la convergence des composantes de π_k.

T7. Montrer qu'il existe une constante positive C et une constante $\eta \in [0,1)$ à déterminer telles que pour toute loi initiale π_0 nous avons

$$|\mathbb{P}(E_k = \varepsilon_i) - \pi_*^i| \leq C\eta^k,\ i \in \{1,2\},\ \varepsilon_1 = A,\ \varepsilon_2 = B.$$

Qu'en déduisez vous sur la convergence de la variable E_k ?

T8. On suppose maintenant qu'un agent financier démarre en 0 avec un état E_0 suivant une loi de Bernoulli de paramètre $p_0 \in [0,1]$ (en supposant toujours que la valeur 1 correspond à l'état B et 0 l'état A). La probabilité de succés β de son projet est connu, mais il n'a pas accès à sa probabilité α d'obtenir le prêt en temps que candidat à l'état E_1. Après 500 tentatives, l'agent a obtenu le prêt 60% du temps. En remarquant que E_1 est une loi de Bernoulli dont on rappellera le paramètre en fonction des données, donner un intervalle de confiance de α en fonction de β et p_0 à 95%.

2 Extension à un modèle d'exclusion sur période fixe

On suppose maintenant que si la banque refuse son prêt à un emprunteur, ce dernier entre dans une période d'exclusion de longueur N, c'est à dire qu'il ne pourra pas obtenir de prêt pendant N années. Ainsi :

- La probabilité que l'agent bénéficiant d'un crédit en k voit son prêt reconduit de k à $k+1$ ne dépend que du succés de son projet et vaut donc α,
- si l'agent est bénéficiaire d'un prêt en k et fait faillite avec probabilité $1-\alpha$, il se retrouve dans un état d'exclusion pendant les N prochaines années,
- si l'agent est dans sa dernière année d'exclusion, la probabilité qu'il obtienne un prêt l'année d'après est β, sinon il reste en suspens de prêt pendant encore une année avant de recandidater.

On note E_k l'état dans lequel l'agent financier se trouve [2] l'année k qui peut être B s'il est bénéficiaire du prêt, A^1 s'il est candidat au prêt ou bien $A^j,\ j \in \{2,\dots,N\}$ s'il est exclu du prêt pour une période de j années restante.

Nous allons maintenant formaliser ce qui a été fait dans le premier exercice afin d'étudier la convergence en temps long de la loi associée à l'état de l'agent emprunteur.

On dit qu'une matrice Q réelle de taille $(N+1)\times(N+1)$ est une matrice de transition sur un espace fini $E = \{x_1,\dots,x_{N+1}\}$ si

2. Notons qu'ici, E_k ne dépend que de l'état de l'agent en $k-1$ et non de tous ces états passés. Nous disons que $(E_k)_{k\geq 1}$ est une chaîne de Markov sur l'espace des états $\{B, A^1,\dots,A^N\}$. Ce type d'objet sera étudié plus en détail en 2A. Nous donnons ici quelques résultats associés à cette théorie pour illustrer notre modélisation.

— $Q(x_i, x_j) \geq 0$ pour tout $(i, j) \in \{1, \ldots, N+1\}^2$,
— $\forall x_i \in E$, $\sum_{y \in E} Q(x_i, y) = 1$.

Il en découle que $Q(x_i, x_j) \in [0, 1]$. On note $Q_{i,j} := Q(x_i, x_j)$.

On dit que Q est la matrice de transition associée â $(E_k)_{k \geq 0}$ si

$$Q(x_i, x_j) = \mathbb{P}(E_{k+1} = x_j | E_k = x_i).$$

T9. Calculer

$$\begin{aligned}
P_{1,1} &= \mathbb{P}(E_{k+1} = B | E_k = B), \\
P_{1,N+1} &= \mathbb{P}(E_{k+1} = A^N | E_k = B), \\
P_{1,j} &= \mathbb{P}(E_{k+1} = A^j | E_k = B),\ j \in \{2, \ldots, N-1\}, \\
P_{2,1} &= \mathbb{P}(E_{k+1} = B | E_k = A^1)\ P_{2,2} = \mathbb{P}(E_{k+1} = A^1 | E_k = A^1), \\
P_{i+1,i} &= \mathbb{P}(E_{k+1} = A^{i-1} | E_k = A^i),\ i \in \{2, \ldots, N\} \\
P_{2,j+1} &= \mathbb{P}(E_{k+1} = A^j | E_k = A^1),\ j \in \{2, \ldots, N\}, \\
P_{i,j} &= \mathbb{P}(E_{k+1} = A^{j-1} | E_k = A^{i-1}),\ i \in \{3, \ldots, N+1\},\ i \neq j+1,
\end{aligned}$$

et écrire la matrice P dans $\mathcal{M}_{N+1 \times N+1}([0, 1])$.
En identifiant $x_1 = B$ et $x_i = A^{i-1}$ pour $2 \leq i \leq N+1$, montrer que P est la matrice de transition associée à $(E_k)_{k \geq 1}$.

T10. Résoudre $\pi_\star P = \pi_\star$ d'inconnue le vecteur ligne $\pi_\star$ de taille $N+1$ à valeur dans $[0, 1]^{N+1}$ tel que la somme de ces composantes soit égale à 1.

T11. On note $\pi_k := (\mathbb{P}(E_k = B), \mathbb{P}(E_k = A^1), \ldots, \mathbb{P}(E_k = A^N))$. Donner la relation entre π_k et π_{k+1} puis déduire π_k en fonction de π_0 et P.

S3. Soit $N = 5$, $\alpha = \frac{1}{3}$ et $\beta = \frac{3}{4}$, calculer numériquement chaque composante de $\pi_0 P^n$ où $n = 1000$ et $\pi_0 := (\frac{1}{N+1}, \ldots, \frac{1}{N+1})$ de taille $N+1$. Que constatez vous en vous aidant des deux questions précédentes ? Tester ce résultat en fonction d'autre vecteur π_0 de taille $N+1$. Que constatez vous ?

19 | Équation de la chaleur

proposé par Clément Rey, clement.rey@polytechnique.edu

1 Objectif

On s'intéresse ici au mouvement Brownien. Il s'agit d'un processus Gaussien, qu'on note $(W_t)_{t\geqslant 0}$, centré de fonction de variance/covariance $\mathbf{Cov}(W_s, W_t) = \inf\{s,t\}$ pour tout $t, s \geqslant 0$ (avec $W_0 = 0$). Le but de ce projet est est de calculer $\mathbb{E}[f(x+W_T)]$, où $T > 0$ est fixé, et f (qui est connu) appartient à une classe de fonctions que l'on précisera plus tard. Il s'agit de la solution de l'équation de la chaleur. Pour cela, on met en œuvre une méthode de type Monte Carlo en supposant simplement des variables aléatoires de Bernoulli.

2 Préliminaires

On appelle processus Gaussien, un ensemble de variables aléatoires $\{G_t, t \geqslant 0\}$ tel que pour tout $m \in \mathbb{N}^*$, et tout $t_i \geqslant 0$, $i \in \{1, \dots, m\}$, le vecteur $(G_{t_1}, \dots, G_{t_m})$ est gaussien.
On appelle mouvement Brownien et on note $(W_t)_{t\geqslant 0}$, un processus Gaussien centré, *i.e.* $\forall t \geqslant 0$, $\mathbb{E}[W_t] = 0$ et tel que $\mathbf{Cov}(W_s, W_t) = \inf\{s,t\}$ pour tous $s, t \geqslant 0$.

T1. Soit $f \in \mathcal{C}^2(\mathbb{R})$. On défini $u : \mathbb{R}_+ \times \mathbb{R}$, $(t,x) \mapsto u(t,x) := \mathbb{E}[f(x+W_t)]$. Montrer que u vérifie l'équation de la chaleur

$$\begin{cases} \partial_t u(t,x) = \partial^2_{x,x} u(t,x) \\ u(0,x) = f(x) \end{cases}$$

Le projet consiste donc à calculer une solution de cette équation en un point (T,x), c'est à dire $\mathbb{E}[f(x+W_T)]$. On l'appelle l'équation de la chaleur. On dit également que f est une fonction test.

T2. Supposons que $\mathbb{E}[|f(x+W_T)|] < +\infty$. Proposer une méthode de calcul numérique pour cette espérance basée sur la simulation d'un échantillon *i.i.d.* de loi normale centrée réduite de taille M.

T3. On pose $f(x) = \exp(x)$. Calculer $\mathbb{E}[f(x+W_T)]$ de façon analytique. Vérifier que l'on peut calculer l'espérance avec la méthode numérique porposée à la question précédente.

S1. Implémenter cette méthode numérique et représenter l'erreur de calcul pour différentes valeurs de M et de $T \leqslant 3$.

3 Approximation de moment Brownien

T4. Soit $t \geqslant 0$. A partir d'une pièce équilibrée (c'est à dire une loi de Bernoulli de paramètre $1/2$), proposer une construction pour une variable aléatoire U_t sous la forme $\phi(t)Z$, où ϕ est une fonction déterministe et Z est une variable aléatoire admettant un moment d'ordre 4, qui ne dépend pas de t, et telle que $\mathbb{E}[|U_t|^q] = \mathbb{E}[|\phi(t)Z|^q] = \mathbb{E}[|\mathcal{N}(0,t)|^q]$ pour $q \in \{1,2,3\}$ et $\mathcal{N}(0,t)$ une variable aléatoire de loi normale centré de variance t.

S2. Implémenter l'algorithme de simulation de U_t. Comment vérifier numériquement que $U_t = \phi(t)Z$ a bien les mêmes moments d'ordre q pour tout $q \in \{1, \dots, 5\}$, que la loi normale centrée de variance t ? Implémenter cette vérification.

L'objectif est maintenant de calculer $\mathbb{E}[f(x + W_T)]$ en utilisant simplement une pièce équilibrée c'est à dire une loi de Bernoulli de paramêtre $1/2$.

4 Résolution numérique de l'équation de la chaleur

4.1 Fonctions test régulières

On suppose dans cette partie que f est une fonction de $\mathcal{C}_b^4(\mathbb{R})$ (fonctions dérivables 4 fois qui ont toutes leurs dérivées d'ordre inférieur ou égal à 4 continues et bornées).

On fixe $T > 0$ et on note $n \in \mathbb{N}^*$, le nombre de points d'observations de la trajectoire. Pour $k \in \mathbb{N}$ on définit $t_k^n = kT/n$ et on introduit la grille de temps $\pi_{T,n} = \{t_k^n = kT/n, k \in \{0, \dots, n\}\}$. Afin de calculer $\mathbb{E}[f(x + W_T)]$, nous allons introduire une approximation de $(x + W_{t_k^n})_{k\in\{0,\dots,n\}}$ sur la grille de temps $\pi_{T,n}$. On pose

$$\begin{cases} X^n_{t^n_{k+1}}(x) = X^n_{t^n_k}(x) + \phi(T/n)Z_k \\ X_0(x) = x \end{cases}$$

où $(Z_k)_{k\in\{1,\dots,n\}}$ est une suite de variables aléatoires *i.i.d* suivant la loi de Z. L'objectif est ici de montrer que pour tout $f \in \mathcal{C}_b^4(\mathbb{R})$ on a

$$\sup_{x\in\mathbb{R}} |\mathbb{E}[f(x + W_T)] - \mathbb{E}[f(X_T^n(x))]| \leqslant C/n$$

On introduit $P_t f(x) = \mathbb{E}[f(x + W_t)]$ et $Q_t^n f(x) = \mathbb{E}[f(X_t^n(x))]$ pour $t \in \pi_{T,n}$ et $f \in \mathcal{C}_b(\mathbb{R})$.

On admet que pour tout $f \in \mathcal{C}_b^4(\mathbb{R})$ et tout $t \geqslant 0$, alors $P_t f$ et $Q_t^n f$ sont également des fonctions de $\mathcal{C}_b^4(\mathbb{R})$.

T5. (facultatif) Montrer que le mouvement Brownien est à accroissements indépendants. En déduire la relation de semigroupe : $P_{t+s} = P_t P_s$. Montrer également qu'on a la relation de semigroupe sur Q^n : $Q^n_{t+s} = Q^n_t Q^n_s$.

T6. (facultatif) En utilisant la relation de semigroupe montrer que

$$P_T f(x) - Q_T^n f(x) = \sum_{k=0}^{m-1} P_{t_k^n} \Delta^n Q^n_{T-t^n_{k+1}} f(x).$$

avec $\Delta^n = P_{T/n} - Q^n_{T/n}$.

T7. Montrer que pour tout $f \in \mathcal{C}_b^4(\mathbb{R})$, alors $\sup_{x\in\mathbb{R}} |\Delta^n f(x)| \leqslant C/n^2$. En déduire que

$$\sup_{x\in\mathbb{R}} |\mathbb{E}[f(x + W_T)] - \mathbb{E}[f(X_T^n(x))]| \leqslant C/n$$

S3. On pose $f = \cos$. Proposer un algorithme de calcul de $u(T, x) = \mathbb{E}[\cos(x + W_T)]$ basé sur la simulation de M trajectoires indépendantes de $(X_t^n)_{t\in\pi_{T,n}}$, chacune de ces trajectoires étant simulée à l'aide n réalisations indépendantes de Z. Implémenter cette méthode dite de Monte Carlo. En utilisant le Théorème Central Limite, déterminer un intervalle asymptotique [1] $I_{n,M}$ tel que $\mathbb{P}(\mathbb{E}[\cos(X_t^n)] \in I_{n,M}) \approx 95\%$. Donner cet intervalle pour $M = 10^4, 10^6$ et $n = 20$.

1. On entend ici par asymptotique le fait que la loi limite (Normale) dans le TCL est supposée atteinte pour M assez grand. On pourra également remplacer la variance par la variance empirique dans le TCL.

S4. Discuter des choix de n et M pour obtenir une précision souhaitée ϵ. Choisir les valeurs $x = 0.5$ et $T = 1$. Représenter l'erreur commise en fonction de n. Observer l'ordre de l'erreur obtenu en traçant $\log(|\mathbb{E}[f(x+W_T)] - \mathbb{E}[f(X_T^n(x))]|)$ (ou plus exactement son approximation en approchant les espérance par des méthodes de Monte Carlo).

4.2 Pour aller plus loin - Fonctions test singulières

On se propose maintenant d'étudier des cas qui sorte du cadre précédent. Plus particulièrement on va s'intéresser à des classes plus large de fonctions test f que les classe des fonctions $\mathcal{C}_b^4(\mathbb{R})$. On ne fera pas d'étude théorique de l'erreur ici. En revanche, il est attendu une analyse numérique de l'erreur de convergence en fonction de n.

S5. Elargir les expérimentations numériques précédentes au cas de fonctions f simplement mesurable à croissance polynomiale. Que constatez-vous ?

T8. Montrer que si f est dérivable partout sauf sur une ensemble dénombrable de points, $\mathbb{E}[f'(W_T)] = \mathbb{E}[f(W_T)W_T/T]$.

S6. En déduire deux algorithmes pour le calcul de la fonction de répartition de W_T à l'aide d'un pièce équilibrée. L'implanter et calculer les quantiles à 95% du Brownien pour $T = 1$.

S7. Elargir ces expérimentations numériques au cas de fonctions dépendants de la trajectoire du Brownien. Par exemple, $f((W_t)_{t\in[0,T]}) = \sup_{t\in[0,T]} W_t$ ou encore $f((W_t)_{t\in[0,T]}) = \inf\{t \in [0,T], W_t \geqslant a\}$ pour $a \in \mathbb{R}$. Que constatez-vous ?

20 | Occurrence d'un mot

proposé par Clément Rey, clement.rey@polytechnique.edu

1 Objectif

L'objectif de ce projet est d'étudier l'apparation d'un mot binaire (mot dont les lettres sont soit 0 soit 1) dans une suite aléatoire de variable aléatoire de Bernoulli de paramêtre 1/2. Dans un second temps on étudiera un jeu : Pour 2 mots binaires donnés, on étudie la probabilité que l'un apparaisse avant l'autre. Cette étude mène au paradoxe de Penney.

2 Préliminaires

Soit $l \in \mathbb{N}$ la longueur d'un mot. Soit $A := (a_i)_{i\in\{1,\dots,l\}} \in \{0,1\}^l$ un mot binaire. Pour $k \in \{1,\dots,l\}$ on notera $A_k := (a_i)_{i\in\{1,\dots,k\}}$. En particulier $A_l = A$. On considère maintenant une suite de tirages indépendants d'une pièce équilibrée, c'est à dire une suite de variable aléatoires $(\epsilon_k)_{k\in\mathbb{N}^*}$ *i.i.d.* de loi de Bernoulli de paramètre 1/2. On définit le processus aléatoire $(X_n^A)_{n\in\mathbb{N}}$ de la façon suivante : $X_0^A = 0$ et pour tout $n \in \mathbb{N}^*$,

$$X_n^A = \begin{cases} 0 & \text{si } \forall k \in \{1,\dots l\}, (\epsilon_{n-k+1},\dots \epsilon_n) \neq A_k \\ k & \text{si } (\epsilon_{n-k+1},\dots \epsilon_n) = A_k \text{ et } (\epsilon_{n-k},\dots \epsilon_n) \neq A_{k+1} \\ l & \text{si } (\epsilon_{n-l+1},\dots \epsilon_n) = A_l = A. \end{cases}$$

T1. Que décrit le processus $(X_n^A)_{n\in\mathbb{N}}$? Montrer que pour tout $n \in \mathbb{N}$ et $(i_j)_{j\in\{1,\dots,n+1\}} \in \{0,\dots,l\}^{n+1}$ on a

$$\mathbb{P}(X_{n+1}^A = i_{n+1} | X_n^A = i_n, \dots X_1^A = i_1) = \mathbb{P}(X_{n+1}^A = i_{n+1} | X_n^A = i_n).$$

On dit alors que le processus $(X_n^A)_{n\in\mathbb{N}}$ est une chaîne de Markov.

On note maintenant, pour $i, j \in \{0,\dots,l\}$ et $n \in \mathbb{N}$, $p_{i,j,n} = \mathbb{P}(X_{n+1}^A = j | X_n^A = i)$.

T2. Montrer que $p_{i,j,n}$ ne dépend pas de n. On dit alors que le processus est stationnaire et on note simplement $p_{i,j}$ au lieu de $p_{i,j,n}$. $(p_{i,j})_{i,j\in\{0,\dots,l\}}$ est appelé la matrice de transition du processus de Markov.

3 Première occurrence d'un mot

On s'intéresse maintenant à l'instant moyen d'apparition d'un mot binaire.

Méthode de calcul par simulation.

S1. Proposer et implémenter un algorithme pour la simulation d'une trajectoire de $(X_n^A)_{n\in\{0,\dots,N\}}$, pour $N \in \mathbb{N}^*$ fixé.
On note I^A l'indice de la première occurrence du mot A dans la suite de variables aléatoire $(\epsilon_n)_{n\in\mathbb{N}^*}$.

S2. Soit $N_{tronc} \in \{0,\dots,N\}$ un paramètre de troncature de la trajectoire, c'est à dire un nombre maximum du nombre de lettres observées pour trouver le mot A. Proposer un algorithme pour le

calcul de $\mathbb{E}[\mathbf{1}_{\{0,\dots,L\}}(I^A)I^A]$ basé sur M simulations indépendantes de la trajectoire de $(X_n^A)_{n\in\{0,\dots,N\}}$. Implémenter cet algorithme pour $A = (0,0,1),(1,1,0,1)$, $N_{tronc} = N = 15$. Justifier la convergence (avec M) de cet algorithme.

S3. Ici, on cherche numériquement une valeur N_{tronc} pour laquelle (lorsque M est assez grand fixé) choisir $N^*_{tronc} \geqslant N_{tronc}$ n'a qu'une influence négligeable sur la précision du résultat [1]. Cela signifie qu'on observe alors l'erreur Monte Carlo (de la Loi des Grands Nombres) et non l'erreur de troncature. On considère alors que $\mathbb{E}[\mathbf{1}_{\{0,\dots,N_{tronc}\}}(I^A)I^A]$ est une bonne approximation de $\mathbb{E}[I^A]$. Que vaut N_{tronc} lorsque $M = 10^4$? Représenter N_{tronc} en fonction de M.

S4. Déterminer numériquement l'ordre de convergence en fonction de la valeur de N_{tronc}. On pourra utiliser comme valeur de référence $\mathbb{E}[\mathbf{1}_{\{0,\dots,N_{sup}\}}(I^A)I^A]$, où N_{sup} est choisi grand devant les valeurs d'étude N_{tronc}, et tracer l'erreur en échelle logarithmique.

Méthode de calcul récursive.

On étudie, pour $k \in \{0,\dots,l-1\}$ et $n,m \in \mathbb{N}$ avec $n > m$,

$$q_k(m,n) := \mathbb{P}(X^A_{m+1} \neq l,\dots,X^A_{m+n-1} \neq l, X^A_{m+n} = l | X^A_m = k)$$

T3. Montrer que $q_k(m,n)$ ne dépend pas de m. On notera simplement $q_k(n)$ dans la suite.

T4. Montrer que pour tout $k \in \{0,\dots,l-1\}$ et $n \in \mathbb{N}^*$

$$q_k(n) = \begin{cases} p_{k,l} & \text{si } n = 1 \\ \sum_{j=0}^{l-1} p_{k,j} q_j(n-1) & \text{si } n > 1. \end{cases}$$

En déduire que pour tout $k \in \{0,\dots,l-1\}$, alors q_k est une loi de probabilité sur $\mathbb{N}^*$.

On note maintenant I^A_k l'indice de la première occurrence du mot A dans la suite de variables aléatoire $(\epsilon_n)_{n\in\mathbb{N}^*}$ en partant de l'état k.

S5. Soit $N_{tronc} \in \mathbb{N}$. Déduire de la question précédente un algorithme récursif pour le calcul de $\mathbb{E}[I^A_0 \mathbf{1}_{\{0,\dots,N_{tronc}\}}(I^A_0)] (= \mathbb{E}[I^A \mathbf{1}_{\{0,\dots,N_{tronc}\}}(I^A)])$. Comparer ces résultats à ceux du calcul par simulation (pour $A = (0,0,1),(1,1,0,1)$, $N_{tronc} = N = 15$).

S6. Etudier empiriqement l'ordre de convergence de cette méthode pour le calcul de $\mathbb{E}[I^A_0]$ par l'expression $\mathbb{E}[I^A_0 \mathbf{1}_{\{0,\dots,N\}}(I^A_0)]$ en fonction du choix de N. Comparer avec les résultats obtenus pour la méthode par simulation.

4 Quel sera le premier mot à apparaître

4.1 Le paradoxe de Penney - Cas particulier

On considère deux mots : $A = 110$ et $B = 011$. On cherche à savoir si l'un de ces deux mots à une plus grande chance d'apparaîttre avant l'autre.

T5. Montrer que $\mathbb{P}(\epsilon_1,\epsilon_2,\epsilon_3 = A) = \mathbb{P}(\epsilon_1,\epsilon_2,\epsilon_3 = B)$.

S7. Implémenter ce jeu. En particulier, on prendra en entrée les mots A et B et une suite suite de lettres de taille N et en sortie, on donnera le vainqueur (s'il y en a un). En vous inspirant des implémentations précédentes, proposer et implémenter une méthode basée sur la simulation de M trajectoires (chacune de taille N) $(X^A_n)_{n\in\{0,\dots,N\}}$ et $(X^B_n)_{n\in\{0,\dots,N\}}$ (et prendre $N = N_{tronc}$) pour calculer la probabilité que le mot A apparaisse avant le mot B. Que constatez vous ? Etudier l'influence des choix de N et M sur la qualité du résultat.

1. $\mathbb{E}[\mathbf{1}_{\{0,\dots,N_{tronc}\}}(I^A)I^A] \approx \mathbb{E}[\mathbf{1}_{\{0,\dots,N^*_{tronc}\}}.(I^A)I^A]$ pour tout $N^*_{tronc} \geqslant N_{tronc}$

4.2 Pour aller plus loin - Étude du cas général

Soit A et B deux mots binaire de longueur $l \in \mathbb{N}^*$. L'objectif est ici de déterminer quel mot va apparaître en premier et avec quelle probabilité.

T6. Montrer que $(X_n^A, X_n^B)_{n\in\mathbb{N}}$ est une chaîne de Markov stationnaire. Pour $n \in \mathbb{N}^*$ fixé, les variables aléatoires X_n^A et X_n^B sont-elles indépendantes ? On notera $p^{A,B}_{(i,i'),(j,j')}$ sa matrice de transition.

De façon similaire à la partie précédente, on introduit, pour $k, j \in \{0, \dots, l-1\}$ et $n, m \in \mathbb{N}$ avec $n > m$,

$$q^A_{k,j}(n,m) := \mathbb{P}(X^A_{m+1} \neq l, X^B_{m+1} \neq l, \dots, X^A_{m+n-1} \neq l, X^B_{m+n-1} \neq l, X^A_{m+n} = l | X^A_m = k, X^B_m = j)$$

et

$$q^B_{k,j}(n,m) := \mathbb{P}(X^A_{m+1} \neq l, X^B_{m+1} \neq l, \dots, X^A_{m+n-1} \neq l, X^B_{m+n-1} \neq l, X^B_{m+n} = l | X^A_m = k, X^B_m = j)$$

T7. Que représente $q^A_{k,j}(m,n)$? Montrer que $q^A_{k,j}(m,n)$ ne dépend pas de m. On notera simplement $q_{k,j}(n)$ dans la suite.

On introduit maintenant $I^{A,B}_{k,j}$ le premier instant d'observation du mot A ou du mot B en partant de l'état (k,j).

T8. Montrer que

$$\mathbb{E}[I^{A,B}_{k,j}] = \sum_{n=0}^{+\infty} n(q^A_{k,j}(n) + q^B_{k,j}(n))$$

et

$$\mathbb{E}[I^{A,B}_{k,j}] = \sum_{i,i'=0}^{l-1} p^{A,B}_{(k,j),(i,i')} \mathbb{E}[I^{A,B}_{i,i'}]$$

T9. Quelle est la probabilité que le mot A apparaîsse avant (ou en même temps) que le mot B en partant de l'état (k,j) ? On notera $q^A_{k,j}$ cette probabilité. Montrer que

$$q^A_{k,j} = \sum_{i,i'=0}^{l-1} p^{A,B}_{(k,j),(i,i')} q^A_{i,i'} + \sum_{i'=0}^{l-1} p^{A,B}_{(k,j),(l,i')}$$

S8. Implémenter les formules qui permettent de calculer $q^A_{k,j}$. Calculer $q^A_{k,j}$ pour les mots binaires $A = 110$ et $B = 011$.

T10. Expliquer le paradoxe de Penney.

21 | Polynômes de Chaos

proposé par Clément Rey, clement.rey@polytechnique.edu

1 Objectif

Les polynôme de chaos fournissent un représentation des variables aléatoires de carré intégrable. L'objectif de ce projet est de calculer les coefficients qui apparaissent dans cette décomposition. Excepté dans certains cas spécifiques, ces coefficients ne peuvent pas être calculer de façon explicite. Ainsi, on étudie dans ce projet deux méthodes pour l'estimation de ces coefficients du chaos. On cherchera à étudier numériquement l'efficacité de ces deux méthodes.

2 Préliminaires

Soit λ une mesure de probabilité définie sur $\mathbb{R}$. Une famille de polynômes $(q_n)_{n\in\mathbb{N}}$ est dite orthogonale pour λ si q_n est de degré n et pour tout $m, n \in \mathbb{N}$, avec $n \neq m$,

$$\langle q_n, q_m\rangle_\lambda := \int_{\mathbb{R}} q_n(x)q_m(x)\lambda(dx) = \mathbb{E}[q_n(X)q_m(X)] = 0,$$

pour X une variable aléatoire de loi λ. De plus si pour tout $n \in \mathbb{N}$, $\|q_n\|_\lambda^2 := \langle q_n, q_n\rangle_\lambda = 1$, on dit que la famille de polynôme $(q_n)_{n\in\mathbb{N}}$ est orthonormale pour λ.

Des familles de polynômes de chaos classiques

T1. (facultatif) Montrer que la famille des polynômes de Legendre définie par

$$\mathrm{Le}_0(x) = 1, \quad \mathrm{Le}_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}[(x-1)^n], \quad n \in \mathbb{N}^*,$$

est orthogonale pour la mesure uniforme sur $[-1, 1]$, $\lambda_{\mathrm{Le}}(dx) = \mathbf{1}_{[-1,1]}(x)\frac{1}{2}dx$.

T2. (facultatif) Montrer que la famille des polynômes d'Hermite définie par

$$\mathrm{He}_0(x) = 1, \quad \mathrm{He}_n(x) = (-1)^n \exp(x^2/2)\frac{d^n}{dx^n}[\exp(-x^2/2)], \quad n \in \mathbb{N}^*,$$

est orthogonale pour la mesure gaussienne standard, $\lambda_{\mathrm{He}}(dx) = \mathbf{1}_{\mathbb{R}}(x)\frac{1}{\sqrt{2\pi}}\exp(-x^2/2)dx$. Pour tout $n \in \mathbb{N}$, calculer $\langle q_n, q_n\rangle_{\lambda_{\mathrm{He}}}$.

3 Décomposition en polynôme de chaos

On dit qu'une suite de variables aléatoires $(Y_N)_{N\in\mathbb{N}}$ converge dans L_2 vers une variable aléatoire Y si

$$\lim_{N\to\infty} \mathbb{E}[|Y - Y_N|^2] = 0.$$

On note alors $Y_N \underset{N\to\infty}{\overset{\mathrm{L}_2}{\to}} Y$.
Le résultat suivant fournit l'existence d'une décomposition en chaos pour une variable de carré intégrable.

Théorème 1. *Soit λ une mesure de probabilité sur $\mathbb{R}$ de support infini telle qu'il existe $C_\lambda > 0$ tel que $\int \exp(C_\lambda x)\lambda(dx) < +\infty$ et soit X une variable aléatoire de loi λ. Alors pour toute famille de polynômes orthogonaux $(q_n)_{n\in\mathbb{N}}$ pour λ et toute fonction $f \in L_2(\mathbb{R},\lambda)$ ($\lambda(|f|^2) < +\infty$), alors il existe une suite $(y_n)_{n\in\mathbb{N}}$, telle que*

$$\sum_{n=0}^{N} y_n q_n(X) \underset{N\to\infty}{\overset{L_2}{\to}} f(X).$$

T3. Montrer que

$$y_n = \frac{\langle q_n, f\rangle_\lambda}{\|q_n\|_\lambda^2}.$$

4 Estimation des coefficients de chaos

On étudie la décomposition en chaos suivant les polynôme d'Hermite. Soit X une variable aléatoire de loi normale centrée réduite (*i.e.* de loi λ_{He}) et $f : \mathbb{R} \to \mathbb{R}$ telle que $f \in \mathrm{L}_2(\mathbb{R}, \lambda_{\mathrm{He}})$.

T4. Montrer qu'il existe une suite $(y_n)_{n\in\mathbb{N}}$ telle que

$$\sum_{n=0}^{N} y_n \mathrm{He}_n(X) \underset{N\to\infty}{\overset{\mathrm{L}_2}{\to}} f(X).$$

T5. Montrer que

$$\mathbb{E}[f(X)] = y_0, \qquad \text{et} \qquad Var(f(X)) = \sum_{n=1}^{\infty} n! y_n^2.$$

Cas d'un calcul explicite des coefficients du chaos

Soit $\gamma > 0$ et $f : \mathbb{R} \to \mathbb{R}_+, x \mapsto \exp(\gamma x)$.

T6. Montrer que

$$y_n = \frac{\exp(\gamma^2/2)}{n!}\mathbb{E}[\mathrm{He}_n(X+\gamma)]$$

T7. On admet que (à montrer en bonus)

$$\forall x \in \mathbb{R}, \qquad \mathrm{He}_n(x+\gamma) = \sum_{k=0}^{n} \binom{n}{k} \gamma^{n-k} \mathrm{He}_k(x)$$

Montrer que

$$y_n = \frac{\exp(\gamma^2/2)\gamma^n}{n!}$$

4.1 Méthode de Monte Carlo

S1. Soit $f \in \mathrm{L}_2(\mathbb{R}, \lambda_{\mathrm{He}})$. Proposer une méthode de calcul numérique de y_n, $n \in \mathbb{N}$, basée sur la simulation de $M \in \mathbb{N}^*$ réalisations aléatoires de la variable aléatoire X de loi normale centrée réduite. Justifier la convergence de cette méthode.

S2. Soit $f : \mathbb{R} \to \mathbb{R}_+, x \mapsto \exp(\gamma x)$, $\gamma > 0$. Retrouver le résultat obtenu précedemment (de façon explicite) numériquement pour y_n pour $\gamma = 0.1, 10$ et $n \in \{0, \dots, 5\}$. Estimer numériquement la vitesse de convergence en fonction de M.

S3. Soit $f : \mathbb{R} \to \mathbb{R}_+, x \mapsto \sqrt{|x|}$. Proposer une méthode pour estimer numériquement la vitesse de convergence en fonction de M et l'appliquer. On se concentrera sur les cas $n \in \{0, \dots, 5\}$.

4.2 Méthode de régression linéaire

Soit $f \in \mathrm{L}_2(\mathbb{R}, \lambda_{\mathrm{He}})$. On propose ici un estimateur basé sur la méthode des moindre carré appliquée à une version tronquée de $\sum_{n=0}^{\infty} y_n \mathrm{He}_n(X)$. En particulier, pour $N \in \mathbb{N}$ (moralement grand), on cherche les coefficients $\overline{y}_N := (y_0, \dots, y_N)$ qui minimisent,

$$\mathbb{E}\Big[\Big| \sum_{n=0}^{N} y_n \mathrm{He}_n(X) - f(X)\Big|^2\Big].$$

S4. Proposer et justifier une méthode d'approximation de cette espérance basée sur la simulation de $M \in \mathbb{N}^*$ réalisations indépendantes de la variable aléatoire X de loi normale centrée réduite.

On admet que pour $K, J \in \mathbb{N}$, $K \geqslant J + 1$, et $\overline{x} = (x_1, \dots, x_K)$,

$$\underset{\overline{a}=(a_0,\dots a_J)\in\mathbb{R}^{J+1}}{\operatorname{argmin}} \sum_{k=0}^{K} \Big| \sum_{j=0}^{J} a_n \mathrm{He}_j(x_k) - f(x_k)\Big|^2 = (A^T A)^{-1} A F$$

avec $A \in \mathbb{R}^{K\times(J+1)}$, $F \in \mathbb{R}^K$ tels que $A_{k,j} = \mathrm{He}_j(x_k)$ et $F_k = f(x_k)$. On admet notamment que $A^T A$ est inversible.

S5. Proposer une méthode d'approximation numérique $\overline{y}_N := (y_0, \dots, y_N)$ basée sur la méthode des moindres carrés et la simulation de $M \in \mathbb{N}^*$ réalisations aléatoires de la variable aléatoire X de loi normale centrée réduite.

S6. Soit $f : \mathbb{R} \to \mathbb{R}_+, x \mapsto \exp(\gamma x)$, $\gamma > 0$. Retrouver le résultat obtenu précedemment (de façon explicite et par méthode de Monte Carlo) numériquement pour y_n pour $\gamma = 0.1, 10$ et $n \in \{0, \dots, 5\}$. On pourra observer l'influence du choix de N. Estimer numériquement la vitesse de convergence en fonction de M lorsque N est fixé.

S7. Soit $f : \mathbb{R} \to \mathbb{R}_+, x \mapsto \sqrt{|x|}$. Retrouver numériquement les résultats obtenus avec la méthode de Monte Carlo. On se concentrera sur les cas $n \in \{0, \dots, 5\}$ et on pourra étudier la dépendance avec la valeur de N. Comparer les deux méthodes.

22 | Régressions linéaires

proposé par Clément Rey, clement.rey@polytechnique.edu

1 Objectif

Dans ce projet on souhaite mettre en oeuvre une méthode pour estimer les paramètres d'une fonction linéaire. Pour cela on on suppose avoir accès à des observations bruitées de cette fonction. En introduisant un modèle de bruit Gaussien, on étudie les propriétés de cet estimateur. Notamment, on montre la convergence (dans un sens que l'on précisera) de cet estimateur.

2 Préliminaires

Soit $\beta_0, \beta_1 \in \mathbb{R}$. On considère une fonction $Y : [0,1] \to \mathbb{R}$, telle que pour tout $x \in [0,1]$,

$$Y(x) = \beta_0 + \beta_1 x$$

On suppose que β_0, β_1 sont des paramètres inconnus que l'ont souhaite estimer à partir d'observation de la fonction Y en n points distincts $(x_1, \dots, x_n)$, avec $n \in \mathbb{N}$. On supposera que l'on observe la fonction Y en ces points mais que ces observation sont buitées. En particulier le practicien observe des réalisations de

$$Y_{obs}(x_i) = Y(x_i) + \epsilon_i, \qquad i \in \{1, \dots, n\},$$

où $(\epsilon_i)_{i\in\mathbb{N}}$ est une suite de variable aléatoire *i.i.d.* suivant la loi normale centrée de variance σ^2 avec $\sigma > 0$.

3 Résolution du problème des moindres carrés

S1. Proposer et implémenter un algorithme de simulation de $Y^n_{obs} := (Y_{obs}(x_1), \dots Y_{obs}(x_n))$ lorsque les $(x_1, \dots, x_n)$ sont choisis de façon uniforme sur $[0,1]$.

T1. Ecrire le problème sous la forme

$$Y^n = X^n \beta$$

où $Y^n = (Y(x_1), \dots, Y(x_n))^T \in \mathbb{R}^n$. On precisera X^n et β.
On admet maintenant que $(X^n)^T X^n$ est inversible (où $(X^n)^T$ est la transposé de la matrice X^n). On s'intéresse à l'estimateur des moindres carrés de β que l'on note $\hat{\beta}^n$ et qui est défini par

$$\hat{\beta}^n := \underset{\beta\in\mathbb{R}^2}{\operatorname{argmin}} \|Y^n_{obs} - X^n\beta\|^2$$

où $\|.\|$ est la norme euclidienne dans $\mathbb{R}^n$. On peut montrer que cette solution maximise la densité de Y^n_{obs}.

T2. Montrer que $\hat{\beta}^n = ((X^n)^T X^n)^{-1}(X^n)^T Y^n_{obs}$.

T3. Quelle est la loi de $\hat{\beta}^n$? En déduire que $\mathbb{E}[\hat{\beta}^n] = \beta$. On dit que l'estimateur $\hat{\beta}^n$ est non biaisé.

T4. Montrer que $\frac{1}{\sigma}(\hat{\beta}^n - \beta)^T((X^n)^T X^n)(\hat{\beta}^n - \beta)$ suit une loi χ^2_2 (chi2 à 2 degrés de libertés). On note $q^{\chi^2_2}(r)$ le quantile de la loi du χ^2_2 à l'ordre r et pour $\alpha \in [0,1]$, on définit

$$\mathcal{E}_{\alpha,n} := \left\{ \tilde{\beta} \in \mathbb{R}^2, \frac{1}{\sigma}(\hat{\beta}^n - \tilde{\beta})^T((X^n)^T X^n)(\hat{\beta}^n - \tilde{\beta}) \leqslant q^{\chi^2_2}(1-\alpha) \right\}.$$

Montrer que $\mathbb{P}(\beta \in \mathcal{E}_{\alpha,n}) = 1 - \alpha$. On appelle $\mathcal{E}_{\alpha,n}$ l'ellipsoïde de confiance à l'ordre α.

S2. Proposer et implémenter un algorithme pour représenter l'ellipsoïde de confiance et pour la vérification du résultat précédent pour $\alpha = 5\%$, $\sigma = 0.1$, et $\beta = (0,1), (1,0), (1,2)$. Réitérer l'expérience en faisant varier le nombre de points d'observation n. Que constatez vous ?

4 Convergence de l'estimateur

On se propose maintenant de vérifier que $\hat{\beta}^n$ converge (dans un sens que l'on précisera) vers β. On commence par modifier la façon de choisir les $(x_1, \ldots, x_n)$ afin de pouvoir réutiliser ces points lorsqu'on modifie n. On va introduire la suite $(x_i)_{i \in \mathbb{N}^*}$ de variables aléatoires *i.i.d.* suivant la loi uniforme sur $[0,1]$ et on conservera la notation X^n pour la matrice construite à partir des n premiers points d'observation $(x_1, \ldots, x_n)$. On admet que pour tout $n \in \mathbb{N}$ $(X^n)^T X^n$ est presque sûrement inversible.

S3. Proposer et implémenter un algorithme de simulation des éléments de la suite $(Y^n_{obs})_{n\in\mathbb{N}} := (Y_{obs}(x_1), \ldots Y_{obs}(x_n))_{n\in\mathbb{N}}$.

T5. Montrer que presque sûrement

$$\lim_{n\to\infty} \frac{1}{n}(X^n)^T X^n = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1/3 \end{pmatrix}$$

En déduire que

$$\mathbb{P}(\lim_{n\to+\infty} \hat{\beta}^n = \beta) = 1.$$

S4. Vérifier numériquement cette convergence.

Calcul récursif des $\hat{\beta}^n$ La méthode décrite jusqu'à maintenant nécessite d'inverser la matrice $(X^n)^T X^n)$ pour chaque n. On se propose maintenant de construire l'inverse de cette matrice de façon récursive. On rappelle la *formule de Woodbury* : Si $A \in \mathbb{R}^{n\times n}, U \in \mathbb{R}^{n\times m}, C \in \mathbb{R}^{m\times m}, V \in \mathbb{R}^{m\times n}$ sont des matrices telles que A, C et $C^{-1} + VA^{-1}U$ sont inversibles, alors $A + UCV$ est inversible et

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}.$$

T6. En utilisant la formule de Woodbury, montrer qu'on a la relation de récurrence

$$((X^{n+1})^T X^{n+1})^{-1} = ((X^n)^T X^n)^{-1} - \frac{1}{\alpha} vv^T$$

où $\alpha > 0$ et $v \in \mathbb{R}^2$ sont à identifier.

S5. En déduire un algorithme récursif pour la construction de $\hat{\beta}^n$. Vérifier la convergence pour $\beta = (0,1), (1,0), (1,2)$ et $\sigma = 0.1$. Pour différentes valeurs de n comparer empiriquement la complexité (en temps) de cet algorithme par rapport au précédent.

5 Test des résidus

On veut tester si les erreurs de mesure sont bien de loi normale centrée de variance σ^2. On pose, pour $n \in \mathbb{N}^*$,

$$\hat{\epsilon}^n = Y^n - \hat{Y}^n$$

où $\hat{Y}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$. On admet que $\hat{\epsilon}^n = (Id(n) - X^n((X^n)^T X^n)^{-1}(X^n)^T) Y^n_{obs}$ où $Id(n)$ est la matrice identité de taille $n \times n$.

T7. Montrer que $\frac{1}{\sigma^2}\|\hat{\epsilon}^n\|^2$ suit la loi χ^2_{n-2} (chi2 à $n-2$ degrés de liberté).

S6. Construire un intervalle $Int(n)$ tel que $\mathbb{P}(\frac{1}{\sigma^2}\|\hat{\epsilon}^n\|^2 \in Int(n)) = 1 - \alpha$, pour $\alpha \in [0,1]$. Prendre $\alpha = 5\%$, $\beta = (0,1), (1,0), (1,2)$ et $\sigma = 0.1$ et $n = 10$. Simuler des entrée $(x_1, \ldots, x_n)$ *i.i.d.* de loi uniforme sur $[0,1]$. À partir de ces entrées, construire M (choisi par l'utilisateur, moralement grand) observations Y^n_{obs} pour chacune des trois valeurs possibles de β. Vérifier que pour ces trois valeurs, les résidus tombent bien dans $Int(n)$ dans (environ) 95% des cas.