Projet 1

Le modèle de Bak-Sneppen

sujet proposé par Olivier Hénard

olivier.henard@universite-paris-saclay.fr

On considère dans ce projet un modèle de *criticalité auto-organisée* introduit par deux physiciens Peter Bak et Kim Sneppen en 1993: par ce vocable on désigne un système sans paramètre qui évolue naturellement vers un état dit "critique" séparant un état "sous-critique" d'un état "sur-critique", tous trois étant caractérisés par des propriétés bien distinctes; l'exemple archétypal est le modèle du tas de sable (*sandpile* model).

Le modèle de Bak-Sneppen est inspiré par la biologie; dans un premier temps, on considère N espèces caractérisées par leur fitness (un nombre réel dans $[0,1]$ qui résume la capacité reproductive de l'espèce); à chaque instant, la plus faible fitness est remplacée par une fitness aléatoire uniforme sur $[0,1]$, ce qui simule la pression de la sélection naturelle; ce modèle donne lieu a un comportement dégénéré en temps long, c'est notre échauffement en partie 1.1.

Pour rendre le modèle intéressant, Bak et Sneppen ajoutent la donnée d'une géométrie au problème; la plus simple consiste à supposer que les espèces $1,\ldots,N$ vivent sur le cercle, de sorte que l'espèce i a deux voisins $i-1$ et $i+1$ (considérés modulo N); la fitness minimale mais aussi les fitness des deux espèces voisines sont alors remplacées par trois variables aléatoires uniformes indépendantes; si la simulation de ce modèle est aisée et la convergence en temps long de la distribution des fitness est bien visible (vous jugerez en partie 1.3), la preuve des résultats expérimentaux que vous observerez résiste encore à l'analyse !

Pour pouvoir faire un peu de mathématiques, nous étudierons aussi un modèle de difficulté intermédiaire, sans géométrie ("mean-field" en anglais), en partie 1.2.

1.1 Modèles avec une seule fitness updatée

On commence par un modèle jouet pour l'échauffement. On construit une suite de vecteurs aléatoires $X(t) = (X_1(t),\ldots,X_N(t)) \in [0,1]^n$ indicés par le temps $t \in \mathbb{N}$ comme suit : le vecteur initial $X(0) = (X_1(0),\ldots,X_N(0))$ est composé de N variables aléatoires uniformes sur $[0,1]$ indépendantes, puis, si $I(t)$ est l'indice tel que X_{I_t} est le minimum [1] de $X_1(t),\ldots,X_N(t)$, on pose de façon récursive:

$$\begin{cases} X_i(t+1) = X_i(t) & \text{pour tout } i \in \{1,\ldots,N\} \setminus \{I_t\}, \\ X_{I_t}(t+1) = U_{t+1}, \end{cases} \tag{1.1}$$

[1] ou le plus petit indice qui réalise le minimum en cas d'égalité

pour $(U_1, U_2, U_3, \ldots)$ une collection de variables aléatoires uniformes sur $[0,1]$ indépendantes, indépendante du vecteur $X(0)$.

En toutes lettres, on a à tout instant une collection de N variables aléatoires dont la plus petite est remplacée par une variable uniforme indépendante. Il sera utile de noter

$$X_{(1)}(t) \leq X_{(2)}(t) \leq \ldots \leq X_{(N)}(t)$$

le réordonnancement croissant de $X_1(t), \ldots, X_N(t)$, de sorte que $X_{I_t} = X_{(1)}(t)$.

S1. Pour $N = 100$, tracer sur un même graphe toutes les trajectoires $i \mapsto X_{(i)}(t)$ pour $i \in \{1, \ldots, N\}$ et $t = 1 \ldots 500$, et commenter.

T1. Discuter la monotonie de $t \mapsto X_{(i)}(t)$, pour $i \geq 2$. A-t-on convergence presque sûre de $t \mapsto X_{(i)}(t)$ quand $t \to \infty$? Observer que

$$\left\{ \sum_{r=1}^{t} \mathbf{1}_{U_r \geq x} \geq N-1 \right\} \subset \left\{ X_{(2)}(t) \geq x \right\}$$

et en déduire la valeur de la limite en probabilité, puis presque sûre, de $t \mapsto X_{(2)}(t)$ et $t \mapsto X_{(i)}(t)$ pour $i \geq 2$.

T2. Commenter brièvement le cas $i = 1$: quelle est la loi limite de $X_{(1)}(t)$ quand $t \to \infty$? A-t-on convergence presque sûre ou en probabilité de $t \mapsto X_{(1)}(t)$?

On propose de préciser les résultats de la question 1 en étudiant les échelles de temps en jeu en fonction de N

T3. Montrer que $X_{(2)}(t)$ est la $(N-1)$-ième plus *grande* valeur de la collection $\{X_1(0), \ldots, X_N(0), U_1, \ldots, U_t\}$ de variables uniformes indépendantes et en déduire[2] la convergence en probabilité suivante quand $N \to \infty$, pour tout $s \geq 0$ réel, avec $\lfloor \cdot \rfloor$ la partie entière:

$$X_{(2)}(\lfloor Ns \rfloor) \to \frac{s}{s+1}.$$

On définit une généralisation du premier modèle comme suit: on se donne k entier entre 1 et N, puis I_t est désormais tel que X_{I_t} est le k-ième plus petit élément de $X_1(t), \ldots, X_N(t)$, soit $X_{I_t} = X_{(k)}(t)$, et enfin $X(t+1)$ est défini à partir de $X(t)$ comme en 1.1. En toutes lettres, on a à tout instant une collection de N variables aléatoires dont la k-ième plus petite est remplacée par une variable uniforme indépendante.

S2. Reprendre la question S1 pour $k = 50$.

T4. Discuter à nouveau la monotonie de $t \mapsto X_{(i)}(t)$, on pourra distinguer les indices $i < k$ et les indices $i > k$.

T5. Étudier la convergence presque sûre de ces trajectoires ainsi que les valeurs des limites, toujours pour les indices i tq $i < k$ et tq $i > k$. Mentionner enfin le cas $i = k$.

[2] un peu plus délicat, c'est facultatif !

1.2 Modèles avec fitness minimale mise à jour plus des fitness aléatoires.

1.2.1 Modèle avec 2 fitness mises à jour

On considère maintenant le modèle où I_t^1 est l'indice tel que $X_{I_t^1} = X_{(1)}(t)$ est le minimum de $X_1(t), \dots, X_N(t)$, et $I^2(t)$ est un indice choisi uniformément parmi $\{1, \dots, N\} \setminus \{I_t^1\}$. On se donne alors $(U_t^j)_{1 \leq j \leq 2, t \geq 1}$ un tableau de variables i.i.d. uniformes sur $[0,1]$, et on pose:

$$\begin{cases} X_i(t+1) = X_i(t) & \text{pour tout } i \in \{1, \dots, N\} \setminus \{I_t^1, I_t^2\}, \\ X_{I_t^j}(t+1) = U_{t+1}^j & \text{pour } j \in \{1, 2\}. \end{cases}$$

S3. Pour $N = 100$, superposer sur un même graphe toutes les trajectoires $i \mapsto X_{(i)}(t)$ pour $i \in \{1, \dots, N\}$ et $t = 1 \dots 500$, et commenter.

On définit F_t^N la fonction de répartition empirique de $(X_1(t), \dots, X_n(t))$ comme l'application (aléatoire) de $[0,1]$ dans $[0,1]$ suivante:

$$x \longmapsto F_t^N(x) = \frac{\sum_{i=1}^N \mathbf{1}_{\{X_i(t) \leq x\}}}{N} = \frac{\max\{i : X_{(i)}(t) \leq x\}}{N}.$$

S4. Superposer sur un même graphe les fonctions de répartition empirique $x \longmapsto F_t^N(x)$ pour $t = 0, 1000, 2000, 3000, 4000, 5000$ et $N = 2000$. De quelle fonction de répartition d'une loi usuelle la courbe de F_t^N semble-t-elle se rapprocher lorsque t devient grand ?

On cherche à préciser le résultat de simulation de S4 en précisant en fonction de N les échelles de temps en jeu ainsi que les fonctions de répartition limites. On pose, pour $x \in [0,1]$ et $s \geq 0$:

$$\phi_s(x) = \big((2x-1) + (1-x)e^{-s}\big)_+,$$

avec $x_+ = \max\{x, 0\}$ la partie positive de x.

S5. Tracer pour $s = 1$ la fonction de répartition empirique $x \longmapsto F_{\lfloor sN \rfloor}^N(x)$ pour $N = 100, 1000, 2000$ ainsi que $x \longmapsto \phi_s(x)$. Reprendre pour quelques autres valeurs de s de votre choix. Qu'est-ce que cela suggère au sujet de la limite en probabilité de $x \longmapsto F_{\lfloor sN \rfloor}^N(x)$ pour $s \in \mathbb{R}$ fixé quand $N \to \infty$?

On admet que, pour tout $\varepsilon > 0$,

$$\lim_{N \to \infty} \mathbb{P}\Big(\sup_{x \in [0,1]} |\phi_s(x) - F_{\lfloor sN \rfloor}^N(x)| \geq \varepsilon \Big) \to 0$$

et on pose $\phi_\infty(x) := \lim_{s \to \infty} \phi_s(x) = (2x-1)_+$, pour $x \in [0,1]$; on reconnaît la fonction de répartition de la loi uniforme sur $[1/2, 1]$. On dit que $y_c = 1/2$ est la valeur critique du modèle; c'est l'infimum du support de la distribution limite. On propose d'expliquer la valeur critique observée dans la section suivante.

Lien avec une marche aléatoire On fixe un niveau z avec $0 < z < 1$ puis pour $k \geq 0$ et $t \geq 0$ entiers on définit:

$$P_k^N(t) = \mathbb{P}(X_{(k)}(t) \leq z < X_{(k+1)}(t)) = \mathbb{P}(\text{Card}\{i = 1 \dots N : X_i(t) \leq z\} = k).$$

T6. Montrer que $P_k^N(t)$satisfait :

$$P_k^N(t+1) = \sum_j P_j^N(t) M_{j,k}$$

pour des coefficients $M_{j,k}$ tels que, pour $k \geq 1$:

$$\begin{array}{llll}
M_{k,k-2} & = (1-z)^2 \dfrac{k-1}{N-1} & & \text{si } k \geq 2 \\
M_{k,k-1} & = (1-z)^2 \dfrac{N-k}{N-1} & +2z(1-z)\dfrac{k-1}{N-1} & \text{si } k \geq 1 \\
M_{k,k} & = 2z(1-z)\dfrac{N-k}{N-1} & +z^2 \dfrac{k-1}{N-1} & \\
M_{k,k+1} & = z^2 \dfrac{N-k}{N-1} & &
\end{array}$$

tandis que $M_{0,0} = (1-z)^2, M_{0,1} = 2z(1-z), M_{0,2} = z^2$, les coefficients non précisés étant nuls.

Le système obtenu en passant à la limite $N \to \infty$, k étant fixé, est:

$$P_k(t+1) = \sum_j P_j(t) L_{j,k}$$

pour des coefficients $L_{j,k}$ tels que, pour $k \geq 1$:

$$\begin{array}{ll}
L_{k,k-1} & = (1-z)^2 \\
L_{k,k} & = 2z(1-z) \\
L_{k,k+1} & = z^2
\end{array}$$

tandis que $L_{0,0} = (1-z)^2, L_{0,1} = 2z(1-z), L_{0,2} = z^2$, les coefficients non précisés étant nuls.

T7. Montrer que, pour $z < 1/2$, $(P_k)_{k \geq 0}$ défini par

$$\begin{aligned}
P_0 &= 1 - 2z \\
P_1 &= (1-2z)\left((1-z)^{-2} - 1\right) \\
P_n &= (1-2z) z^{2n-2} (1-z)^{-2n}, \qquad \text{si } n \geq 2.
\end{aligned}$$

est une solution au sens où:

$$P_k = \sum_j P_j L_{j,k} \text{ et } \sum_{k \geq 0} P_k = 1 \text{ et } P_k \geq 0.$$

On admet que c'est la seule solution, et qu'il n'existe pas de solution pour $z \geq 1/2$. Enfin, interpréter[3] ces résultats en terme du nombre de particules $X_i(t)$ de position $< 1/2$ dans le système en temps long.

On définit $(B_t, t \geq 1)$ des variables aléatoires indépendantes à valeurs dans $\{-1, 0, 1\}$ de même loi

$$\mathbb{P}(B_1 = +1) = z^2, \mathbb{P}(B_1 = 0) = 2z(1-z), \mathbb{P}(B_1 = -1) = (1-z)^2,$$

puis on pose

$$\begin{cases} S_0 = 0 \\ S_{t+1} = S_t + B_{t+1} + \mathbf{1}_{\{S_t = 0\}}. \end{cases}$$

Enfin on définit $Q_k(t) = \mathbb{P}(S_t = k)$ la loi de S_t.

[3] on n'attend pas de justification rigoureuse

T8. Montrer que

$$Q_k(t+1) = \sum_j Q_j(t) L_{j,k}.$$

S6. Tracer des trajectoires de $t \mapsto S_t$ pour $z = 0,4$, $z = 0,5$ et $z = 0,6$. Commenter.

T9. Montrer à l'aide de la loi des grands nombres pour la suite $(B_t)_{t \geq 1}$ que les trajectoires aléatoires $t \mapsto S_t$ divergent p.s. vers $+\infty$ quand $t \to +\infty$ si $z > 1/2$.

T10. Observer que

$$0 \leq S_t = \sum_{1 \leq r \leq t} B_r + \sum_{0 \leq r \leq t-1} \mathbf{1}_{\{S_r = 0\}},$$

et conclure, toujours à l'aide de la loi des grands nombres, que $\mathrm{Card}\{t \geq 0 : S_t = 0\} = \infty$ p.s. dans le cas $z < 1/2$. (Bonus : et dans le cas $z = 1/2$?)

S7. Tracer l'histogramme de S_t pour $t = 100$ pour $N = 100$ réalisations, pour les valeurs $z = 0,3$ et $z = 0,4$ et superposer sur le même graphe les valeurs de P_k. Commenter.

1.2.2 Modèle avec m fitness mises à jour

On suppose maintenant $N \geq m$. On définit une généralisation du modèle précédent comme suit : on choisit $I_t^2, I_t^3, \ldots, I_t^m$ des indices choisis uniformément sans remplacement (tous distincts donc) dans l'ensemble $\{1, \ldots, N\} \setminus \{I_t^1\}$, puis on pose, toujours pour $(U_t^j)_{1 \leq j \leq m, t \geq 1}$ un tableau de variables i.i.d. uniformes sur $[0, 1]$:

$$\begin{cases} X_i(t+1) = X_i(t) & \text{pour tout } i \in \{1, \ldots, N\} \setminus \{I_t^1, \ldots, I_t^m\} \\ X_{I_t^j}(t+1) = U_{t+1}^j & \text{pour } j \in \{1, \ldots, m\} \end{cases}$$

S8. Pour $m = 3$, puis $m = 4$, reprendre la question S3. Conjecturez quelle est la dépendance en m de la valeur critique $y_c^{(m)}$ de ce modèle à partir des simulations.

1.3 Modèle de Bak-Sneppen original.

On considère enfin le modèle de Bak-Sneppen original sur le tore; l'indice I_t est tel que $X_{I_t}(t) = X_{(1)}(t)$ puis on pose, pour $(U_t^j)_{-1 \leq j \leq 1, t \geq 1}$ un tableau de variables i.i.d. uniformes sur $[0, 1]$:

$$\begin{cases} X_i(t+1) = X_i(t) & \text{pour tout } i \in \{1, \ldots, N\} \setminus \{I_t - 1, I_t, I_t + 1\} \\ X_{I_t + j}(t+1) = U_{t+1}^j & \text{pour } j \in \{-1, 0, 1\} \end{cases}$$

avec $I_t - 1$ et $I_t + 1$ pris modulo N (c'est-à-dire qu'on travaille sur le tore). À la différence des modèles vus jusqu'ici, la structure spatiale compte véritablement.

S9. Reprendre la question S3 avec les mêmes valeurs que précédemment, et donner une valeur approchée de la valeur critique z_c de ce modèle à partir des simulations.

Remarque : on pourrait vérifier comme en S4 la convergence de la fonction de répartition empirique (dans la même échelle de temps) vers une certaine fonction $\phi_s(x)$, mais à la différence du modèle mean-field, une telle convergence n'est pas *prouvée* et il est fort peu probable que $\phi_s(x)$ admette une expression analytique simple.

T11. Comparer les deux valeurs critiques z_c et $y_c^{(3)}$, et proposer une intuition sur le classement de ces deux valeurs critiques.

S10. Pour $1 \leq i \leq N = 100$, et $0 \leq t \leq T = 300$, colorier le point (i, t) par l'âge

$$A_i(t) := t - \max\{0 \leq s \leq t : |i - I_s| \leq 1\} \in \{0, \ldots, t\}$$

de $X_i(t)$ en utilisant un gradient de couleur. Commenter le graphe obtenu.