

3. Convergence presque sûre

Les précédents chapitres ont expliqué comment, à partir de la commande `random.random()`, on pouvait engendrer un échantillon $(X_1, X_2, \dots, X_N)$ de N variables aléatoires suivant une loi donnée (pas forcément la loi uniforme sur $[0, 1]$). On s'intéresse maintenant aux propriétés statistiques de ces échantillons, en particulier lorsque N est grand.

1. Loi des grands nombres.

La *loi des grands nombres* est le résultat suivant : si $(X_1, X_2, \dots, X_n)$ est un n -échantillon de variables réelles suivant toutes la même loi μ et si μ admet un moment d'ordre 1 :

$$\int_{\mathbb{R}} |x| \mu(dx) < +\infty,$$

alors la suite des *moyennes empiriques* $M_n = \frac{X_1 + X_2 + \dots + X_n}{n}$ tend presque sûrement lorsque n tend vers l'infini vers

$$m = \mathbb{E}[X_1] = \int_{\mathbb{R}} x \mu(dx).$$

Ainsi, $\mathbb{P}[\lim_{n \rightarrow \infty} M_n = m] = 1$.

- (1) Écrire un programme qui prend un échantillon $(X_1, X_2, \dots, X_N)$ et un paramètre m , et qui dessine le graphe de la suite des moyennes $(M_n)_{1 \leq n \leq N}$, et le compare à la valeur m (représentée par une droite horizontale d'ordonnée m).

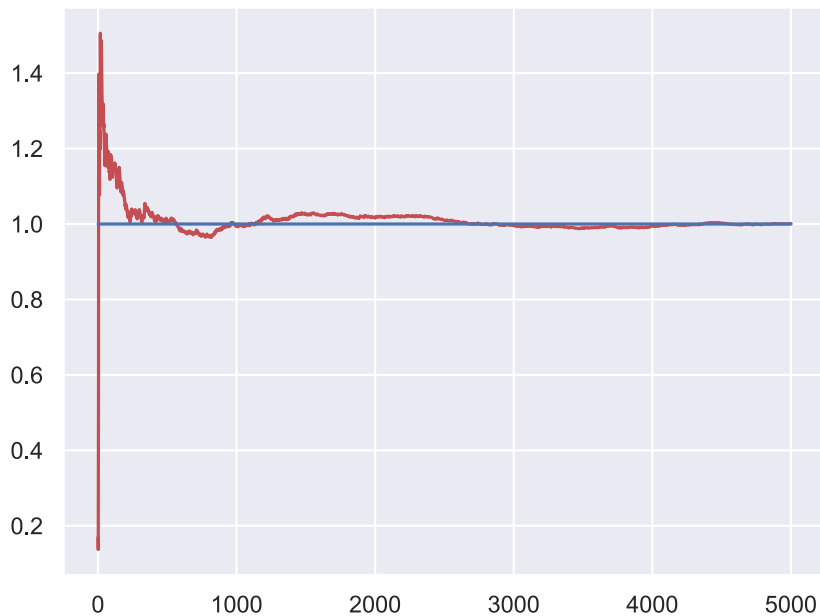


FIG. 3.1. Convergence presque sûre des moyennes empiriques d'une suite de variables i.i.d. vers la moyenne théorique.

- (2) Utiliser ce programme avec un échantillon de taille $N = 1000$ de variables indépendantes et exponentielles de loi $\text{Exp}(1)$. On prendra $m = \mathbb{E}[X_1]$. La loi des grands nombres est-elle bien mise en évidence ?

2. Urnes de Bernoulli–Laplace.

De façon générale, on dit qu’une suite de variables aléatoires $(M_n)_{n \in \mathbb{N}}$ converge presque sûrement vers une variable X si $\mathbb{P}[\lim_{n \rightarrow \infty} M_n = X] = 1$. La loi des grands nombres assure la convergence presque sûre lorsque M_n est une moyenne empirique de variables indépendantes et de même loi admettant une espérance m , avec une limite $X = m$. Il y a de nombreux autres résultats de convergence presque sûre en probabilité, avec une limite X qui peut être aléatoire. Cet exercice présente un tel exemple, à partir d’un modèle d’urne.

- (1) On considère une urne qui contient initialement $B_0 = 1$ boule blanche et $R_0 = 1$ boule rouge. À chaque instant $n \geq 0$, on tire uniformément l’une des $n + 2$ boules de l’urne au temps n (indépendamment de tout ce qui s’est passé avant), et on la remplace par deux boules de la même couleur pour obtenir l’urne au temps $n + 1$. On a donc deux suites croissantes $(B_n)_{n \geq 0}$ et $(R_n)_{n \geq 0}$ d’entiers aléatoires, avec $n + 2 = B_n + R_n$ pour tout n . Écrire un programme `urn(N)` qui tire au hasard la suite $(B_n)_n$ jusqu’au rang N . On pourra calculer les probabilités conditionnelles $\mathbb{P}[B_{n+1} = k + 1 \mid B_n = k]$ et $\mathbb{P}[B_{n+1} = k \mid B_n = k]$.
- (2) On note $M_n = \frac{B_n}{n+2}$; c’est la proportion de boules blanches au rang n . Dessiner la fonction $n \mapsto M_n$ pour plusieurs tirages de la suite, et pour $n \in \llbracket 0, N \rrbracket$ avec $N = 1000$. La suite $(M_n)_{n \in \mathbb{N}}$ converge-t-elle presque sûrement ?
- (3) Dessiner la fonction de répartition empirique de 1000 tirages de la variable aléatoire M_{1000} . Que peut-on conjecturer ?

3. Moyennes de variables de Cauchy.

Cet exercice examine le cas où l’hypothèse de la loi des grands nombres d’existence d’un moment d’ordre 1 n’est plus vérifiée. Ainsi, on considérera des variables de Cauchy, dont on rappelle que la densité est la fonction $f(x) = \frac{1}{\pi(1+x^2)}$. Un échantillon de taille N de variables aléatoires de Cauchy peut être obtenu avec la commande `scs.cauchy.rvs(size=N)`.

- (1) Soit $(X_1, X_2, \dots, X_N)$ un N -échantillon de variables de Cauchy, et $(M_1, M_2, \dots, M_N)$ la suite des moyennes empiriques correspondantes. Dessiner plusieurs fois la fonction $n \mapsto M_n$ sur l’intervalle $\llbracket 1, N \rrbracket$, avec $N = 5000$. Qu’observe-t-on ? Y-a-t’il une convergence presque sûre ?
- (2) On note μ_n la loi de $\frac{X_1 + \dots + X_n}{n}$, où les X_i sont des variables aléatoires de Cauchy indépendantes. On souhaite déterminer cette loi μ_n , et sa fonction de répartition F_n . Écrire un programme `loi_empirique_moyenne_cauchy(N, n)` qui construit un échantillon $Z_1, \dots, Z_N$ de variables aléatoires indépendantes de loi μ_n , et qui trace la fonction de répartition empirique de cet échantillon.
- (3) Utiliser ce programme pour dessiner sur un même graphique des versions approchées des fonctions de répartition F_1 , F_2 et F_{10} . Que peut-on conjecturer sur la loi μ_n de $\frac{X_1 + \dots + X_n}{n}$ en fonction de n ?
- (4) Relier la conjecture de la question précédente aux observations de la première question.

4. Théorème de Glivenko–Cantelli.

Soit μ une loi de variables aléatoires réelles, $X_1, \dots, X_n$ des réalisations indépendantes de cette loi, et F_n^X la fonction de répartition empirique de cet échantillon. Le *théorème de Glivenko–Cantelli* assure que lorsque n tend vers l'infini, on a

$$\|F_n^X - F_\mu\|_\infty \rightarrow_{\text{presque sûrement}} 0.$$

On s'est déjà servi implicitement de ce résultat, en approximant une fonction de répartition ou un histogramme théorique par l'objet empirique correspondant obtenu à partir d'un échantillon de grande taille. Les premières questions de l'exercice donnent une preuve complète du théorème, et les questions suivantes s'intéressent à la vitesse de convergence.

- (1) On souhaite d'abord montrer que pour tout $x \in \mathbb{R}$, $F_n^X(x)$ converge presque sûrement vers $F_\mu(x)$. Relier ce résultat à la loi des grands nombres pour des variables de Bernoulli de paramètre $p \in (0, 1)$. Si $B_1, \dots, B_n$ sont des variables indépendantes de loi $\text{Ber}(p)$ et $S_n = B_1 + \dots + B_n$, montrer que la convergence presque sûre $\frac{S_n}{n} \rightarrow p$ est impliquée par la condition suivante :

$$\mathbb{E}[(S_n - np)^4] = O(n^2).$$

On pourra considérer la série $\sum_{n \geq 1} \mathbb{P}\left[\left|\frac{S_n}{n} - p\right| \geq n^{-\frac{1}{8}}\right]$, et utiliser une forme de l'inégalité de Markov.

- (2) Montrer que $\mathbb{E}[S_n] = np$ et $\mathbb{E}[(S_n)^2] = np + n(n-1)p^2$. Pour le second moment, on pourra développer $(\sum_{i=1}^n B_i)^2$ et regrouper les termes $B_i B_j$ selon que $i = j$ ou $i \neq j$. En exploitant cette idée, montrer qu'on a de même

$$\begin{aligned}\mathbb{E}[(S_n)^3] &= np + 3n^{\downarrow 2}p^2 + n^{\downarrow 3}p^3; \\ \mathbb{E}[(S_n)^4] &= np + 7n^{\downarrow 2}p^2 + 6n^{\downarrow 3}p^3 + n^{\downarrow 4}p^4\end{aligned}$$

avec $n^{\downarrow k} = n(n-1)\dots(n-k+1)$.

- (3) Dans **Python**, on peut manipuler des polynômes en n et p avec les commandes suivantes :

```
from sympy import var, expand
var("n,p")
M1 = n*p
M2 = n*p + n*(n-1)*(p**2)
```

Alors, la commande `expand(M2)` renvoie $n^2p^2 - np^2 + np$. En utilisant les formules pour les moments d'ordre 1 à 4, calculer $\mathbb{E}[(S_n - np)^4]$, et en déduire la convergence ponctuelle des fonctions de répartition empiriques.

- (4) Notons F la fonction de répartition théorique de la loi uniforme. On suppose établie la convergence presque sûre $\|F_n^U - F\|_\infty \rightarrow 0$ pour les fonctions de répartition empiriques de variables $U_1, U_2, \dots, U_n$ indépendantes et uniformes sur $[0, 1]$. En utilisant la méthode de simulation par inversion, en déduire la convergence presque sûre $\|F_n^X - F_\mu\|_\infty \rightarrow 0$ dans le cas général, c'est-à-dire pour les fonctions de répartition empiriques de variables $X_1, X_2, \dots, X_n$ indépendantes et de loi μ .
- (5) Soit $(f_n)_{n \in \mathbb{N}}$ une suite de fonctions de $\mathbb{R}$ vers $[0, 1]$ qui sont croissantes. On suppose que $(f_n)_{n \in \mathbb{N}}$ converge ponctuellement vers une fonction $f : \mathbb{R} \rightarrow [0, 1]$ qui est continue (et bien sûr croissante). Montrer alors que $\|f_n - f\|_\infty \rightarrow 0$ (c'est le *théorème de Dini*), et conclure la preuve du théorème de Glivenko–Cantelli.

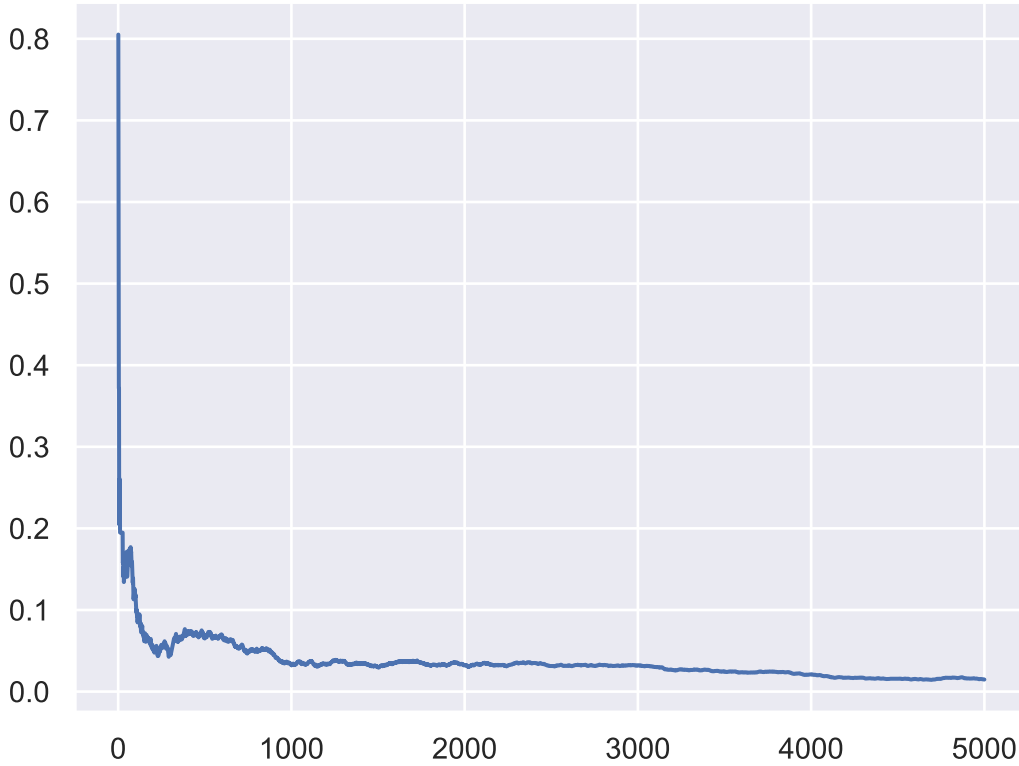


FIG. 3.2. Glivenko–Cantelli : décroissance presque sûre vers 0 de $\|F_n - F\|_\infty$.

Dans toute la suite, F_n désigne la fonction de répartition empirique d'un échantillon $U_1, \dots, U_n$ de variables indépendantes uniformes dans $[0, 1]$, et $(U_{(1)} \leq U_{(2)} \leq \dots \leq U_{(n)})$ désigne le réordonnement croissant de ces variables.

(6) Montrer que

$$\|F_n - F\|_\infty = \max \left(\max_{k=1, \dots, n} \left\{ \frac{k}{n} - U_{(k)} \right\}, \max_{k=1, \dots, n} \left\{ U_{(k)} - \frac{k-1}{n} \right\} \right).$$

Écrire un programme `distance_loi_uniforme(N)` qui tire au hasard un échantillon $U_1, \dots, U_N$, et qui renvoie le vecteur des distances $\|F_n - F\|_\infty$ avec $n \in \llbracket 1, N \rrbracket$. On prendra garde au fait qu'au temps n , on doit utiliser le réordonnement croissant de $U_1, \dots, U_n$, et pas les n premiers termes du réordonnement croissant de $U_1, \dots, U_N$.

(7) Tracer le graphe de

$$n \mapsto \|F_n - F\|_\infty$$

pour $n \in \llbracket 1, 5000 \rrbracket$, et vérifier qu'il semble y avoir une décroissance en $n^{-\alpha}$ pour un certain exposant α . Déterminer la valeur de cet exposant en traçant le graphe de

$$n \mapsto \frac{\log \|F_n - F\|_\infty}{\log(n+1)}.$$

(8) Tracer la différence renormalisée $n^\alpha(F_n(x) - F(x))$ sur l'intervalle $[0, 1]$, pour $n = 5000$. Commenter.