Sorbonne Université
Master Mathématiques et Applications
M2 Statistique, M2 Apprentissage et Algorithmes

Année 2021/2022
Second Semestre

Méthodes Monte-Carlo

Arnaud Guyader

Table des matières

Chapitre 1

Génération de variables aléatoires

Introduction

Idéalement, une méthode Monte-Carlo repose sur la simulation d'une suite de variables aléatoires $(X_n)_{n\geq 1}$ indépendantes et identiquement distribuées (i.i.d.) selon une loi donnée. Ce chapitre expose quelques méthodes pour y parvenir, au moins de façon approchée, en commençant par la loi uniforme, sur laquelle toutes les autres sont basées.

1.1 Variables uniformes

Le but est de générer une suite de variables aléatoires $(U_n)_{n\geq 1}$ indépendantes et de loi uniforme sur $[0,1]$. Ceci est bien entendu impossible sur un ordinateur : d'abord parce que les nombres entre 0 et 1 sont en fait de la forme $k/2^p$ avec $k \in \{0,\dots,2^p-1\}$; ensuite parce que vérifier qu'une suite $(U_n)_{n\geq 1}$ est bien i.i.d. est une question très délicate. C'est essentiellement l'objet de la théorie de la complexité de Kolmogorov (voir par exemple [6], Chapitre 7, pour une introduction à ce sujet).

En pratique, les logiciels se basent sur une suite dite pseudo-aléatoire qui a la forme $X_{n+1} = f(X_n)$ avec X_n à valeurs dans un ensemble fini, typiquement de taille beaucoup plus grande que 2^p. L'objectif est alors d'en déduire une suite (U_n) qui ressemble autant que possible à une suite i.i.d. uniforme sur $\{0,\dots,2^p-1\}$. La fonction f étant déterministe, il est clair qu'un tel générateur est périodique. Il existe de nombreuses méthodes pour générer de telles suites et nous ne détaillerons pas ce sujet, nous contentant de mentionner celle utilisée par défaut sous Python, R et Matlab, à savoir le Mersenne Twister MT-19937 :

$$X_{n+1} = AX_n \qquad (\text{mod } 2)$$

où X_n est un vecteur de 19937 bits et A une matrice bien choisie. La variable pseudo-aléatoire U_n correspond alors aux 32 derniers bits de X_n. Proposé par Matsumoto et Nishimura en 1998 [23], ce générateur a une période de longueur $2^{19937}-1$ (i.e. le maximum possible vu la taille de X_n et le fait que 0 est un point fixe), lequel est un nombre premier de Mersenne.

Notons aussi que si, pour une raison ou une autre, on veut changer de générateur sous R, il suffit de le spécifier avant l'appel à `runif` via la fonction `RNGkind`. En voici un exemple pour utiliser un générateur proposé par L'Ecuyer :

```
> set.seed(123) # fixe la graine du générateur
> runif(1) # génération par Mersennne Twister
[1] 0.2875775
> RNGkind("L'Ecuyer-CMRG") # changement de générateur pseudo-aléatoire
> set.seed(123)
```

```
> runif(1)
[1] 0.1663742
> RNGkind("Mersenne-Twister") # retour au générateur par défaut
> set.seed(123)
> runif(1)
[1] 0.2875775
```

Remarque : Sur cet exemple, le fait de fixer la graine du générateur pseudo-aléatoire par la commande `set.seed` a simplement permis de vérifier les changements de générateurs. De façon plus générale, cette commande en début de programme peut s'avérer très utile puisqu'elle permet de refaire une simulation dans les mêmes conditions et, le cas échéant, de trouver une erreur dans un code.

Dans tout ce cours, nous supposerons donc disposer d'un générateur pseudo-aléatoire "satisfaisant" pour la loi uniforme. La suite de ce chapitre explique alors comment, partant de la loi uniforme, il est possible de générer d'autres lois. Pour une référence très complète sur le sujet, on pourra consulter le livre de Luc Devroye [8] en libre accès sur sa page.

1.2 Méthode d'inversion

Cette méthode part d'une propriété élémentaire, souvent vue en Licence dans un premier cours de probabilités. Dit rapidement, soit F une fonction de répartition bijective, alors si U suit une loi uniforme sur $[0,1]$, la variable aléatoire $X = F^{-1}(U)$ a pour fonction de répartition F. En effet, pour tout réel x,

$$\mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x),$$

où l'on a appliqué respectivement l'aspect bijectif de F, sa croissance et la forme spécifique de la fonction de répartition de la loi uniforme.

Pour ce qui nous concerne, l'intérêt de ce résultat est clair : pour peu que F soit facilement inversible, alors pour générer une variable aléatoire de loi[1] F, il suffit de générer une variable uniforme U et de lui appliquer F^{-1}. On peut en fait généraliser ceci à des fonctions de répartition non bijectives.

Définition 1 (Inverse généralisée)
Soit X une variable aléatoire de fonction de répartition F. On appelle inverse généralisée de F, ou fonction quantile de X, la fonction F^{-1} définie pour tout $u \in]0,1[$ par

$$F^{-1}(u) = \inf\{x \in \mathbb{R} : \ F(x) \geq u\}.$$

Rappel. Le réel $q_{1/2} := F^{-1}(1/2)$ est appelé la médiane de F. De façon générale, lorsque $0 < \alpha < 1$, $q_{1-\alpha} := F^{-1}(1-\alpha)$ est appelé quantile d'ordre $(1-\alpha)$ de F. On le rencontre constamment dans les tests d'hypothèses, le plus fameux d'entre eux étant celui associé aux intervalles de confiance à 95% de la loi gaussienne centrée réduite : $q_{0.975} := \Phi^{-1}(0.975) = 1.96\ldots \approx 2$, où Φ est la fonction de répartition de la loi $\mathcal{N}(0,1)$.

Si F est inversible, il est clair que cette fonction quantile correspond bien à l'inverse classique de F. A contrario, considérons une variable aléatoire X discrète à valeurs dans l'ensemble fini $\{x_1 < \cdots < x_m\}$ avec probabilités $(p_1, \ldots, p_m)$. Il est facile de vérifier que pour tout $u \in]0,1[$,

$$F^{-1}(u) = \begin{cases} x_1 & \text{si } 0 < u \leq p_1 \\ x_2 & \text{si } p_1 < u \leq p_1 + p_2 \\ \vdots & \\ x_m & \text{si } p_1 + \cdots + p_{m-1} < u \leq 1 \end{cases}$$

1. Rappelons que la fonction de répartition caractérise la loi.

c'est-à-dire

$$F^{-1}(u) = \sum_{k=1}^{m} x_k \mathbf{1}_{p_1+\cdots+p_{k-1}<u\leq p_1+\cdots+p_k}. \tag{1.1}$$

Si l'ensemble des valeurs prises par la variable discrète X n'est pas fini, il suffit de remplacer cette somme par une série. Quoi qu'il en soit, outre que, tout comme F, cette fonction quantile est croissante et en escalier, on notera que, contrairement à F, elle est continue à gauche. Ces propriétés sont en fait vraies de façon générale. Venons-en maintenant aux résultats utiles en simulation.

Proposition 1 (Méthode d'inversion)
Soit U une variable uniforme sur $[0,1]$, F une fonction de répartition et F^{-1} son inverse généralisée. Alors :

1. *la variable aléatoire $X = F^{-1}(U)$ a pour fonction de répartition F.*
2. *si X a pour fonction de répartition F et si F est continue, alors la variable aléatoire $F(X)$ est de loi uniforme sur $[0,1]$.*

Remarque : c'est le premier point qui porte le nom de méthode d'inversion. Le second point nous sera utile lorsque nous aborderons la théorie des copules en Section 1.4.4.

Preuve. Soit $X = F^{-1}(U)$ et x réel fixé. Montrons que pour tout $u \in]0,1]$, on a

$$F^{-1}(u) \leq x \iff u \leq F(x). \tag{1.2}$$

Par définition de $F^{-1}(u)$, si $u \leq F(x)$, alors $F^{-1}(u) \leq x$. Inversement, si $F^{-1}(u) \leq x$, alors pour tout $\varepsilon > 0$ on a $F^{-1}(u) < x + \varepsilon$, donc par définition de $F^{-1}(u)$ et par croissance de F, il vient $u \leq F(x+\varepsilon)$. Puisque F est continue à droite, on en déduit que $u \leq F(x)$ et l'équivalence (1.2) est établie. Dès lors, la fonction de répartition de X se calcule facilement :

$$\mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x),$$

la dernière égalité venant de ce que, pour tout $u \in [0,1]$, $\mathbb{P}(U \leq u) = u$. Le premier point est donc prouvé. On l'applique pour le second : la variable $Y = F^{-1}(U)$ a même loi que X, donc la variable $F(X)$ a même loi que $F(Y) = (F \circ F^{-1})(U)$. Or, d'après (1.2), pour tout $u \in]0,1]$, on a

$$F^{-1}(u) \leq F^{-1}(u) \implies u \leq (F \circ F^{-1})(u).$$

Réciproquement, pour tout $u \in]0,1]$ et pour tout $\varepsilon > 0$, on a, toujours par (1.2),

$$F^{-1}(u) - \varepsilon < F^{-1}(u) \implies F(F^{-1}(u) - \varepsilon) < u.$$

Etant donné que $u \in]0,1]$ et que F est supposée continue en $F^{-1}(u) > 0$, le passage à la limite lorsque $\varepsilon \to 0$ donne $(F \circ F^{-1})(u) \leq u$. Au total, on a donc prouvé que, pour tout $u \in]0,1]$, $(F \circ F^{-1})(u) = u$. En revenant aux variables aléatoires, ceci donne $F(Y) = (F \circ F^{-1})(U) = U$ presque sûrement.

■

Exemples :

1. Concernant la seconde propriété, si X suit une loi de Bernoulli de paramètre $1/2$, la variable $F(X)$ prend les valeurs $1/2$ et 1 de façon équiprobable, donc n'est pas uniforme. De façon générale, dès lors que la loi de X présente un atome, par exemple en x_0, sa fonction de répartition F ne prend aucune des valeurs dans l'intervalle $]F(x_0^-), F(x_0)[$, donc la variable aléatoire $F(X)$ non plus, donc elle ne peut correspondre à une variable uniforme.

2. Loi exponentielle : Soit $\lambda > 0$, alors
$$X = -\frac{1}{\lambda}\log(1-U) \overset{\mathcal{L}}{=} -\frac{1}{\lambda}\log U \sim \mathcal{E}(\lambda),$$
loi exponentielle de paramètre λ, c'est-à-dire de densité $f(x) = \lambda e^{-\lambda x}\mathbf{1}_{x\geq 0}$. Malgré sa simplicité, les logiciels n'utilisent pas tous cette méthode, l'appel au logarithme étant coûteux.
3. Loi de Cauchy : si $U \sim \mathcal{U}_{[0,1]}$, alors $X = \tan(\pi(U-1/2))$ suit une loi de Cauchy.
4. Loi discrète : on reprend l'exemple de la loi discrète ci-dessus. D'après la formule (1.1), il suffit dont de tirer une variable U uniforme et de regarder dans quel intervalle de la forme $(p_1 + \cdots + p_{k-1}; p_1 + \cdots + p_k]$ elle tombe. Noter que ceci se fait en testant successivement si $U > p_1$, puis si $U > p_1 + p_2$, etc. Dès lors, dans la mesure du possible, on aura tout intérêt à ranger les p_i de façon décroissante pour gagner du temps.
5. Supposons qu'on veuille simuler le résultat d'un dé équilibré : la méthode précédente s'applique, mais on aura plus vite fait de considérer $X = \lceil 6U \rceil$, où $\lceil\cdot\rceil$ désigne la partie entière par excès, ce qui en R donne `X <- ceiling(6*runif(1))`.

Bien entendu, la méthode d'inversion requiert la connaissance et le calcul rapide de $F^{-1}(u)$. Ceci n'est pas toujours envisageable : il suffit de penser à la loi normale centrée réduite, de fonction de répartition
$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$
et dont l'inverse n'admet pas d'expression analytique simple.

Dans ce cas, la méthode de rejet, détaillée en section suivante, peut représenter une alternative judicieuse. La méthode ziggourat est un exemple d'application de la méthode de rejet permettant de simuler la loi normale, ou encore la loi exponentielle sans faire appel à la fonction logarithme.

1.3 Méthode de rejet

Le contexte est le suivant : on veut simuler une variable aléatoire X de densité f, mais f est trop compliquée pour que ceci puisse se faire directement. On dispose par contre d'une densité auxiliaire g telle que :

(a) on sait simuler Y de densité g ;
(b) il existe une constante m telle que, pour tout y, $f(y) \leq mg(y)$;
(c) pour tout y, on sait calculer le rapport $f(y)/(mg(y))$.

Considérons alors deux suites indépendantes de variables aléatoires :

(a) $(Y_n)_{n\geq 1}$ i.i.d. de densité g ;
(b) $(U_n)_{n\geq 1}$ i.i.d. de loi uniforme sur $[0,1]$.

Concrètement, Y correspond à une proposition et U à un tirage pile ou face pour décider si on accepte ou non cette proposition. Nous noterons r la fonction rapport d'acceptation pour le pile ou face, à savoir : $r(y) = f(y)/(mg(y))$ si $g(y) > 0$, $r(y) = 0$ sinon. Par ailleurs, f et g étant des densités, la constante m intervenant dans la majoration est bien entendu supérieure à 1 :
$$1 = \int_{\mathbb{R}} f(x)dx \leq \int_{\mathbb{R}} mg(x)dx = m\int_{\mathbb{R}} g(x)dx = m.$$

Ceci étant acquis, le résultat suivant montre comment simuler suivant la densité f voulue.

Proposition 2 (Algorithme de rejet)
Soit $N = \inf\{n \geq 1,\ U_n \leq r(Y_n)\}$ le premier instant où le tirage est accepté, alors N suit une loi géométrique de paramètre $1/m$. En particulier N est presque sûrement fini donc la variable aléatoire $X = Y_N$ est bien définie presque sûrement. Alors X a pour densité f. De plus, les variables N et X sont indépendantes.

Remarque : si $p \in]0,1[$, nous disons que N suit une loi géométrique de paramètre p si N est à valeurs dans $\mathbb{N}^\star$, avec pour tout $n \in \mathbb{N}^\star$: $\mathbb{P}(N = n) = p(1-p)^{n-1}$, ce qui équivaut à dire que $\mathbb{P}(N > n) = (1-p)^n$. L'autre convention, qui est par exemple celle de `R` pour la fonction `rgeom`, consiste à considérer N à valeurs dans $\mathbb{N}$, avec pour tout $n \in \mathbb{N} : \mathbb{P}(N = n) = p(1-p)^n$. La seconde variable se déduit tout simplement de la première en enlevant 1.

Preuve. Commençons par montrer que N est p.s. finie. A priori, N est à valeurs dans $\mathbb{N}^\star \cup \{+\infty\}$. Pour tout $n \in \mathbb{N}^\star$, le fait que les couples (Y_i, U_i) sont i.i.d. permet d'écrire

$$\mathbb{P}(N > n) = \mathbb{P}(U_1 > r(Y_1), \ldots, U_n > r(Y_n)) = \mathbb{P}(U_1 > r(Y_1))^n.$$

Les variables Y_1 et U_1 étant indépendantes, leur loi jointe est le produit des marginales, ce qui donne par Fubini-Tonelli

$$\mathbb{P}(U_1 > r(Y_1)) = \mathbb{E}[\mathbf{1}_{U_1 > r(Y_1)}] = \int_{\mathbb{R}} \left(\int_0^1 \mathbf{1}_{u > r(y)} du \right) g(y) dy,$$

d'où, g et f étant des densités,

$$P(U_1 > r(Y_1)) = \int_{\mathbb{R}} (1 - r(y)) g(y) dy = 1 - \int_{\mathbb{R}} r(y) g(y) dy = 1 - \frac{1}{m},$$

donc

$$\mathbb{P}(N > n) = \left(1 - \frac{1}{m}\right)^n,$$

ce qui montre que N suit une loi géométrique de paramètre $1/m$ et en particulier que cette variable est p.s. finie. La variable $X = Y_N$ est donc p.s. bien définie. Notons maintenant F la fonction de répartition associée à la densité f. On a alors

$$\mathbb{P}(X \leq x, N = n) = \mathbb{P}(U_1 > r(Y_1), \ldots, U_{n-1} > r(Y_{n-1}), U_n \leq r(Y_n), Y_n \leq x).$$

Puisque les n tirages sont i.i.d., ceci s'écrit encore

$$\mathbb{P}(X \leq x, N = n) = (\mathbb{P}(U_1 > r(Y_1))^{n-1} \mathbb{P}(U_n \leq r(Y_n), Y_n \leq x).$$

Pour le premier terme, ce qui précède donne

$$(\mathbb{P}(U_1 > r(Y_1))^{n-1} = \left(1 - \frac{1}{m}\right)^{n-1}.$$

Pour le second, un raisonnement comparable donne

$$\mathbb{P}(U_n \leq r(Y_n), Y_n \leq x) = \mathbb{E}[\mathbf{1}_{U_n \leq r(Y_n)} \mathbf{1}_{Y_n \leq x}] = \int_{\mathbb{R}} \left(\int_0^1 \mathbf{1}_{u \leq r(y)} du \right) \mathbf{1}_{y \leq x}\ g(y) dy,$$

c'est-à-dire

$$\mathbb{P}(U_n \leq r(Y_n), Y_n \leq x) = \int_{\mathbb{R}} \mathbf{1}_{y \leq x}\ r(y) g(y) dy = \int_{-\infty}^x r(y) g(y) dy = \frac{1}{m} \int_{-\infty}^x f(y) dy = \frac{F(y)}{m}.$$

Au total, on a obtenu

$$\mathbb{P}(X \leq x, N = n) = \left(1 - \frac{1}{m}\right)^{n-1} \times \frac{F(x)}{m}.$$

Par sigma-additivité, il vient

$$\mathbb{P}(X \leq x) = \sum_{n=1}^{\infty} \mathbb{P}(X \leq x, N = n) = \frac{F(x)}{m} \sum_{n=1}^{\infty} \left(1 - \frac{1}{m}\right)^{n-1} = F(x),$$

ce qui prouve que X a bien la loi voulue. Enfin, le fait que $\mathbb{P}(X \leq x, N = n) = \mathbb{P}(X \leq x)\mathbb{P}(N = n)$ montre bien l'indépendance des variables X et N.

■

Remarque : la variable N étant géométrique de paramètre $1/m$, elle a pour moyenne m. Par conséquent, il faut faire en moyenne m essais pour obtenir une seule simulation selon la loi cible f. Dès lors, il s'agira de choisir le couple (g, m) de sorte que m soit aussi proche de 1 que possible. En d'autres termes, on a tout intérêt à choisir une densité g qui ressemble le plus possible à f.

Exemple jouet. On veut simuler X de densité $f(x) = 3x^2$ sur $[0, 1]$. Puisque $f(x) \leq 3$, on peut choisir pour g la loi uniforme sur $[0, 1]$ et prendre $m = 3$, ce qui donne pour le rapport d'acceptation

$$r(y) = \frac{f(y)}{mg(y)} = y^2.$$

L'algorithme est alors le suivant : on simule un couple (Y_1, U_1) de variables uniformes indépendantes : si $U_1 \leq Y_1^2$, on pose $X = Y_1$, sinon on recommence. Le nombre moyen d'essais pour obtenir une réalisation de X est donc égal à 3. Cet exemple est "jouet" en ce sens que, dans ce cas particulier, il vaut mieux appliquer la méthode d'inversion : un calcul immédiat montre en effet que si U est uniforme sur $[0, 1]$, la variable $X = U^{1/3}$ a pour densité f.

La preuve de la Proposition 2 est identique si au lieu de calculer la probabilité $\mathbb{P}(X \leq x, N = n)$, on étudie $\mathbb{P}(X \in A, N = n)$ pour A borélien quelconque de $\mathbb{R}$. L'intérêt est que cette version s'adapte directement en dimension supérieure. Autrement dit, la méthode de rejet reste valable si f et g sont des densités sur $\mathbb{R}^d$. De façon générale, pour pouvoir appliquer la méthode de rejet, il suffit que la loi de X soit absolument continue par rapport à celle de Y. Voyons un cas particulier d'application.

Exemple : simulation d'une loi conditionnelle. Soit B un ensemble (borélien) contenu dans le carré $[0, 1] \times [0, 1]$ de $\mathbb{R}^2$. Supposons qu'on veuille simuler des points uniformément dans B mais que, pour tout point x du plan, on ne dispose que d'une fonction boîte noire nous disant si, oui ou non, x est dans B (ce n'est rien d'autre que l'indicatrice de B). L'algorithme de simulation est alors naturel : il suffit de générer des points Y_n uniformément dans $[0, 1] \times [0, 1]$ jusqu'à l'instant aléatoire N où l'on tombe dans B. Par le résultat précédent, la variable $X = Y_N$ est alors uniforme sur B. En effet, dans ce contexte, en notant $\lambda(B)$ la mesure de Lebesgue de B (i.e. sa surface) et $C = [0, 1] \times [0, 1]$, on a, avec des notations évidentes,

$$f(x, y) = \frac{\mathbf{1}_B(x, y)}{\lambda(B)} \text{ et } g(x, y) = \mathbf{1}_C(x, y) \implies m = \frac{1}{\lambda(B)} \text{ et } r(x, y) = \mathbf{1}_B(x, y).$$

Le point remarquable de cet exemple est que la constante m est inconnue !

Généralisation. l'idée sous-jacente à l'exemple précédent est généralisable : supposons que $f(x) = c_1 f_u(x)$ où c_1 est une constante de normalisation inconnue et f_u est calculable en tout point (le "u" signifiant "unnormalized"). Cette situation est récurrente en statistique bayésienne et en physique statistique. On suppose qu'il existe une densité g facilement simulable, calculable en tout point

et vérifiant $f_u(x) \leq c_2 g(x)$ avec c_2 connue. Alors il suffit de prendre $m = c_1 c_2$: la méthode de rejet s'applique et ne nécessite nullement la connaissance de c_1 puisque $r(y) = f(y)/(mg(y)) = f_u(x)/(c_2 g(y))$, que l'on sait calculer. Ceci est un grand classique de certaines méthodes Monte-Carlo et on le recroisera par exemple dans l'algorithme de Metropolis-Hastings.

1.4 Vecteurs aléatoires

Puisqu'il n'existe pas de méthode universelle pour simuler une variable aléatoire de loi donnée, il en va de même pour les vecteurs aléatoires. Nous nous contentons donc de donner ici quelques pistes.

1.4.1 Conditionnement

Supposons que l'on veuille simuler un couple aléatoire (X, Y). Si les deux coordonnées sont indépendantes, on est ramené à ce qui précède puisqu'il suffit de simuler X et Y indépendamment.

Si tel n'est pas le cas, on peut éventuellement s'en sortir si, par exemple, X est facilement simulable et si la loi de Y sachant X l'est aussi : il suffit de simuler x selon $\mathcal{L}(X)$ puis y selon la loi conditionnelle $\mathcal{L}(Y|X = x)$. Dans le cas de variables à densité, ceci est tout simplement basé sur le fait que $f(x, y) = f(x) \times f(y|x)$.

Exemple. On considère la densité

$$f(x, y) = xe^{-xy}\mathbf{1}_{0<x<1}\mathbf{1}_{y>0}.$$

Il est facile de voir que X suit une loi uniforme sur $[0, 1]$ et que, sachant $X = x$, Y suit une loi exponentielle de paramètre x. La simulation du couple (X, Y) est donc triviale.

1.4.2 Changement de variables

La formule de changement de variables dans les intégrales doubles permet de déterminer la densité d'un couple aléatoire (U, V) à partir de celle d'un couple aléatoire (X, Y) lorsque ceux-ci sont en bijection.

Théorème 1 (Changement de variables)
Soit (X, Y) un couple aléatoire de densité $f_{X,Y}$ portée par l'ouvert A et soit $\varphi : A \to B$ un $\mathcal{C}^1$-difféomorphisme, alors le couple aléatoire $(U, V) = \varphi(X, Y)$ admet pour densité

$$f_{U,V}(u, v) = f_{X,Y}(\varphi^{-1}(u, v))|\det J_{\varphi^{-1}}(u, v)|\mathbf{1}_B(u, v).$$

Ce résultat se retrouve par la méthode de la fonction muette : on cherche $f_{U,V}(u, v)$ telle que pour toute fonction continue bornée $h : \mathbb{R}^2 \to \mathbb{R}$, on ait

$$\mathbb{E}[h(U, V)] = \iint_{\mathbb{R}^2} h(u, v) f_{U,V}(u, v) du dv.$$

Or, par le Théorème de Transfert,

$$\mathbb{E}[h(U, V)] = \mathbb{E}[h(\varphi(X, Y))] = \iint_{\mathbb{R}^2} h(\varphi(x, y)) f_{X,Y}(x, y) dx dy = \iint_A h(\varphi(x, y)) f_{X,Y}(x, y) dx dy.$$

Il ne reste plus qu'à appliquer la formule de changement de variable vue en cours d'analyse en considérant

$$(u, v) = \varphi(x, y) \iff (x, y) = \varphi^{-1}(u, v),$$

ce qui donne

$$\mathbb{E}[h(U,V)] = \iint_B h(u,v) f_{X,Y}(\varphi^{-1}(u,v)) |\det J_{\varphi^{-1}}(u,v)| du dv,$$

c'est-à-dire

$$\mathbb{E}[h(U,V)] = \iint_{\mathbb{R}^2} h(u,v) f_{X,Y}(\varphi^{-1}(u,v)) |\det J_{\varphi^{-1}}(u,v)| \mathbf{1}_B(u,v) du dv.$$

Rappelons que $J_{\varphi^{-1}}(u,v)$ désigne la matrice jacobienne de l'application φ^{-1} au point (u,v) avec formellement

$$J_{\varphi^{-1}}(u,v) = \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix} (u,v).$$

Exemple. Un changement de variables classiques est celui en coordonnées polaires $\varphi : \mathbb{R}^2 \setminus ([0,+\infty[\times\{0\}) \to]0,+\infty[\times]0,2\pi[$ avec

$$(x,y) = \varphi^{-1}(r,\theta) = (r\cos\theta, r\sin\theta),$$

donc $|J_{\varphi^{-1}}(r,\theta)| = r$. Par conséquent, si le couple (X,Y) a pour densité $f_{X,Y}$ alors le couple $(R,\Theta) = \varphi(X,Y)$ a pour densité

$$f_{R,\Theta}(r,\theta) = f_{X,Y}(r\cos\theta, r\sin\theta) r \mathbf{1}_{]0,+\infty[}(r) \mathbf{1}_{]0,2\pi[}(\theta).$$

Application : algorithme de Box-Muller. D'après ce qui vient d'être vu, si (X,Y) est un couple de variables gaussiennes centrées réduites indépendantes, alors les coordonnées polaires (R,Θ) ont pour densité

$$f_{R,\Theta}(r,\theta) = \left(re^{-r^2/2} \mathbf{1}_{]0,+\infty[}(r) \right) \left(\frac{\mathbf{1}_{]0,2\pi[}(\theta)}{2\pi} \right),$$

autrement dit elles sont indépendantes, l'angle Θ suit une loi uniforme sur $]0,2\pi[$ (isotropie) et la distance à l'origine a pour densité (dite loi de Rayleigh)

$$f_R(r) = re^{-r^2/2} \mathbf{1}_{]0,+\infty[}(r).$$

Puisque $R^2 = X^2 + Y^2$, cette variable R est la racine carrée d'un khi-deux à deux degrés de liberté, c'est-à-dire la racine carrée d'une loi exponentielle de paramètre $1/2$. Or on a vu précédemment que si U suit une loi uniforme sur $[0,1]$, alors $-2\log U$ suit une loi exponentielle de paramètre $1/2$. Réciproquement, soit (U,V) un couple de variables i.i.d. de loi uniforme sur $[0,1]$. Alors les variables X et Y définies par

$$\begin{cases} X = \sqrt{-2\log U} \times \cos(2\pi V) \\ Y = \sqrt{-2\log U} \times \sin(2\pi V) \end{cases}$$

sont i.i.d. gaussiennes centrées réduites. C'est l'algorithme de Box-Muller.

Remarque. La méthode de Box-Muller semble parfaite pour générer des variables gaussiennes. En fait, elle n'est pas utilisée en pratique en raison de l'appel à des fonctions coûteuses en temps de calcul (logarithme, cosinus et sinus). Comme mentionnée précédemment, la génération de gaussiennes est souvent plutôt basée sur une technique de rejet sophistiquée (méthode de ziggourat).

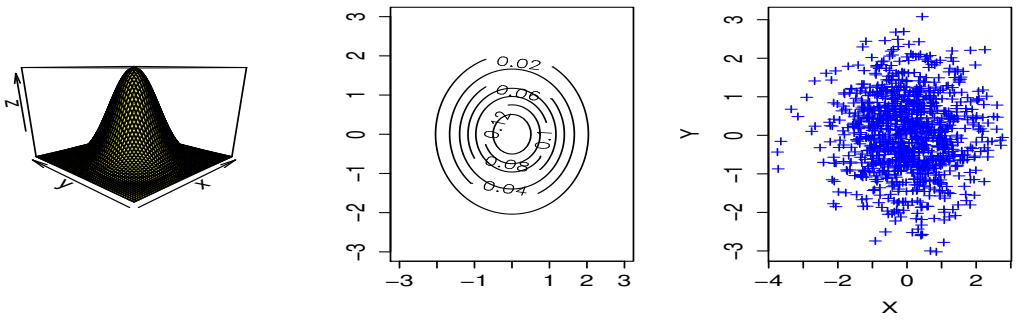

FIGURE 1.1 – Densité, lignes de niveaux et échantillon de la loi normale standard.

1.4.3 Vecteurs gaussiens

Rappelons que si $X = [X_1, \ldots, X_d]'$ est un vecteur aléatoire dont les composantes admettent toutes des moments d'ordre 2, l'espérance du vecteur X est définie par $\mathbb{E}[X] = [\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d]]'$ et sa matrice de covariance (ou de dispersion) Γ est la matrice $d \times d$ symétrique semi-définie positive de terme générique

$$\Gamma_{i,j} = \mathrm{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])] = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j].$$

Par ailleurs, si A est une matrice $d_0 \times d$ et b un vecteur de taille d_0, tous deux déterministes, le vecteur aléatoire $Y = AX + b$ admet pour espérance $\mathbb{E}[Y] = A\mathbb{E}[X] + b$ et pour matrice de covariance $\Gamma_Y = A\Gamma A'$.

D'autre part, un vecteur $X = [X_1, \cdots, X_d]'$ est dit gaussien si toute combinaison linéaire de ses variables est gaussienne, c'est-à-dire si pour tout d-uplet de coefficients $(\alpha_1, \ldots, \alpha_d)$, la variable $\alpha_1 X_1 + \cdots + \alpha_d X_d$ est gaussienne.

Du point de vue de la simulation, les vecteurs gaussiens ont un double intérêt : d'une part, ils sont complètement caractérisés par moyenne et dispersion, d'autre part ils sont stables par transformation affine. C'est ce que résume le résultat suivant.

Proposition 3 (Transformation affine d'un vecteur gaussien)
Soit $X \sim \mathcal{N}(m, \Gamma)$ un vecteur gaussien de dimension d, A une matrice $d_0 \times d$ et b un vecteur de taille d_0, tous deux déterministes, alors le vecteur aléatoire $Y = AX + b$ est gaussien, avec plus précisément :

$$Y \sim \mathcal{N}(Am + b, A\Gamma A').$$

Supposons maintenant qu'on veuille simuler un vecteur aléatoire gaussien $X \sim \mathcal{N}(m, \Gamma)$ dans $\mathbb{R}^d$, m et Γ étant donnés. Soit R une racine carrée de Γ, c'est-à-dire une matrice $d \times d$ vérifiant $\Gamma = RR'$. La matrice Γ étant semi-définie positive, elle se diagonalise en base orthonormée, c'est-à-dire sous la forme $\Gamma = P\Delta P'$ avec $P' = P^{-1}$ et $\Gamma = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$, les valeurs propres λ_i étant toutes positives ou nulles. Il suffit donc de prendre

$$R = P\sqrt{\Delta} \qquad \text{avec} \qquad \sqrt{\Delta} = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_d}). \tag{1.3}$$

Si Γ est définie positive, on peut aller plus vite via la méthode de Choleski, qui fournit une matrice R triangulaire inférieure vérifiant cette propriété. Bref, une racine carrée étant obtenue, il suffit de

partir d'un vecteur gaussien centré réduit, donc facile à simuler, pour arriver à nos fins. En effet, en vertu de la proposition précédente,

$$X_0 \sim \mathcal{N}(0, I_d) \implies X = RX_0 + m \sim \mathcal{N}(m, \Gamma),$$

et l'affaire est dans le sac.

Exemple. Sous R, le calcul d'une racine carrée est effectué comme expliqué en (1.3) et la génération de vecteurs gaussiens se fait très simplement par l'appel préalable à la librairie `MASS` :

```
> library(MASS)
> mvrnorm(1000,m,Gamma)
```

génère 1000 vecteurs gaussiens de moyenne m et de matrice de dispersion Γ. Ceci est illustré Figure 1.2 pour $m = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ et $\Gamma = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}$. Voir aussi l'exercice 1.12.

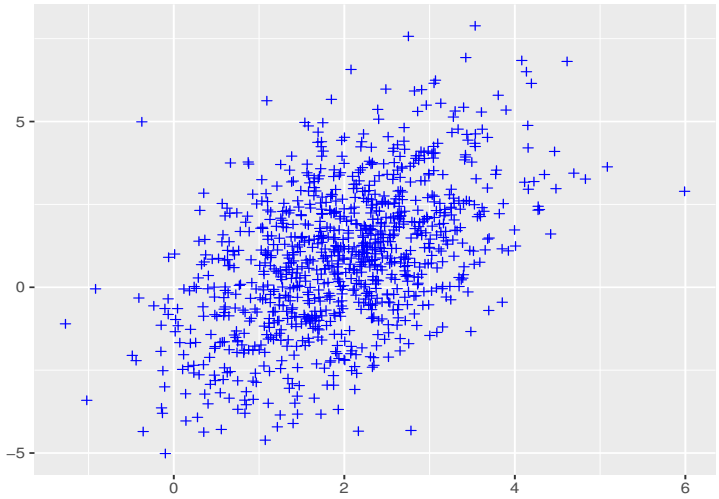

FIGURE 1.2 – Echantillon de 1000 points suivant la loi $\mathcal{N}(m, \Gamma)$.

1.4.4 Copules

Tout ce qui suit est exposé en dimension 2, mais la généralisation en dimension supérieure se fait sans problème. Pour des compléments sur toute cette section, on pourra se reporter à [26] ou au Chapitre 5 de [24]. Pour un couple aléatoire (X, Y), l'idée des copules est de séparer la dépendance entre les variables X et Y de leurs lois marginales.

On rappelle que la fonction de répartition d'un couple (X, Y) est définie par :

$$\forall (x, y) \in \mathbb{R} \times \mathbb{R} \qquad F(x, y) = \mathbb{P}(X \leq x, Y \leq y).$$

La définition d'une copule est liée à la fonction de répartition d'un couple dont les marginales sont uniformes.

Définition 2 (Copule)
On dit qu'une fonction $C : [0,1]^2 \to [0,1]$ est une copule si C est la fonction de répartition d'un couple (U, V) tel que U et V suivent chacune une loi uniforme sur $[0,1]$, autrement dit :

$$C(u, v) = \mathbb{P}(U \leq u, V \leq v),$$

avec $U \sim \mathcal{U}_{[0,1]}$ et $V \sim \mathcal{U}_{[0,1]}$.

En toute rigueur, C est plutôt la restriction à $[0,1]^2$ de la fonction de répartition du couple (U,V) puisque cette dernière est définie sur $\mathbb{R}^2$ tout entier.

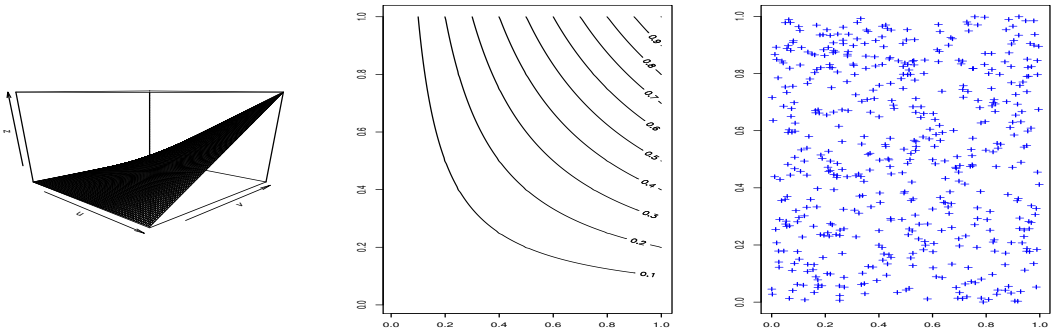

FIGURE 1.3 – Copule indépendante $c(u,v) = uv$, lignes de niveaux et nuage de points $(U_i,V_i)_{1\leq i\leq 500}$.

Exemples : dans tout ce qui suit, on considère $0 \leq u,v \leq 1$. Voir l'exercice 1.17 pour les représentations.

- copule comonotone : si $U \sim \mathcal{U}_{[0,1]}$ alors $V = U \sim \mathcal{U}_{[0,1]}$ et $C(u,v) = \min(u,v)$.
- copule anti-comonotone : si $U \sim \mathcal{U}_{[0,1]}$ alors $V = 1-U \sim \mathcal{U}_{[0,1]}$ et $C(u,v) = \max(u+v-1,0)$.
- copule indépendante : si $U \sim \mathcal{U}_{[0,1]}$ et $V \sim \mathcal{U}_{[0,1]}$, avec U et V indépendantes, alors $C(u,v) = uv$ (voir Figure 1.3).

Le fait qu'une copule permette de séparer la loi jointe des lois marginales est illustré par le résultat suivant.

Théorème 2 (Théorème de Sklar (1959))
Soit (X,Y) un couple aléatoire de fonction de répartition F, avec X de fonction de répartition F_X et Y de fonction de répartition F_Y. Alors il existe une copule C telle que

$$\forall (x,y) \in \mathbb{R}\times\mathbb{R} \qquad F(x,y) = C(F_X(x),F_Y(y)). \tag{1.4}$$

De plus, si F_X et F_Y sont continues, la copule C est unique et définie par :

$$\forall\ 0 \leq u,v \leq 1 \qquad C(u,v) = F(F_X^{-1}(u),F_Y^{-1}(v)),$$

où F_X^{-1} et F_Y^{-1} sont les inverses généralisées de X et Y.

Preuve. On se contente du cas où F_X et F_Y sont continues, qui est la seule situation qui nous intéressera dans les exemples. D'après le second point de la Proposition 1, les variables aléatoires $U = F_X(X)$ et $V = F_Y(Y)$ sont alors toutes deux uniformes sur $[0,1]$. Par définition, la fonction de répartition $C(u,v) = \mathbb{P}(U \leq u, V \leq v)$ du couple (U,V) est donc une copule. De plus, puisque U et V sont sans atome

$$C(u,v) = \mathbb{P}(U \leq u, V \leq v) = \mathbb{P}(U < u, V < v) = \mathbb{P}(F_X(X) < u, F_Y(Y) < v).$$

Or la relation (1.2) est équivalente à

$$u > F_X(x) \iff F_X^{-1}(u) > x,$$

ce qui donne ici

$$C(u,v) = \mathbb{P}(X < F_X^{-1}(u), Y < F_Y^{-1}(v)) = \mathbb{P}(X \leq F_X^{-1}(u), Y \leq F_Y^{-1}(v)) = F(F_X^{-1}(u), F_Y^{-1}(v)),$$

le passage des inégalités strictes aux larges étant dues au fait que X et Y sont sans atome. La seconde relation est donc établie. Pour montrer la première, on peut remarquer que si X est sans atome, alors on a presque sûrement l'égalité suivante :

$$\{X \leq x\} \overset{p.s.}{=} \{F_X(X) \leq F_X(x)\}.$$

C'est clair dans un sens par croissance de F_X :

$$\{X \leq x\} \implies \{F_X(X) \leq F_X(x)\}.$$

Dans l'autre sens, on raisonne par contraposition :

$$\{X > x\} \implies \{F_X(X) \geq F_X(x)\} \overset{p.s.}{=} \{F_X(X) > F_X(x)\},$$

la dernière égalité p.s. étant due au fait que que $F_X(X)$ est uniforme, donc sans atome. Ainsi, vu la définition de la copule C, il vient

$$C(F_X(x), F_Y(y)) = \mathbb{P}(U \leq F_X(x), V \leq F_Y(y)) = \mathbb{P}(F_X(X) \leq F_X(x), F_Y(Y) \leq F_Y(y)),$$

et il reste à utiliser la relation précédente pour conclure que

$$C(F_X(x), F_Y(y)) = \mathbb{P}(X \leq x, Y \leq y) = F(x,y).$$

L'unicité vient du fait que si X et Y sont continues, $F_X(\bar{\mathbb{R}}) = F_Y(\bar{\mathbb{R}}) = [0,1]$ et la relation précédente permet de conclure.

■

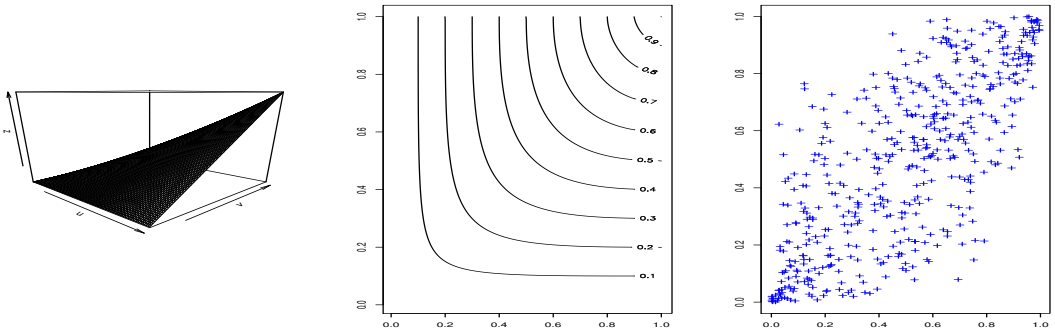

FIGURE 1.4 – Copule gaussienne de paramètre $\rho = 3/4$, lignes de niveaux et nuage de points $(U_i, V_i)_{1 \leq i \leq 500}$.

Remarque. Lorsqu'une copule C vérifie la relation (1.4), on dit que c'est une copule pour le couple (X,Y).

Venons-en à quelques exemples de copules classiques.

Exemple : copule gaussienne. Si (X,Y) est un couple gaussien centré tel que $\mathrm{Var}(X) = \mathrm{Var}(Y) = 1$ et $\mathrm{Cov}(X,Y) = \rho \in]-1,1[$, alors la copule gaussienne de paramètre ρ est définie par

$$C(u,v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right\} dxdy,$$

où Φ^{-1} est la réciproque de Φ, fonction de répartition de la loi normale centrée réduite (voir Figure 1.4).

Une autre famille classique est celle des copules archimédiennes. Si X est une variable aléatoire positive, la fonction

$$\varphi(\lambda) = \mathbb{E}[e^{-\lambda X}]$$

est appelée transformée de Laplace de X. Elle est dérivable sur $]0,\infty[$ avec $\varphi'(\lambda) = -\mathbb{E}[Xe^{-\lambda X}]$. Dès lors, si X est strictement positive, φ est strictement décroissante et établit une bijection de $[0,\infty[$ dans $]0,1]$. On note φ^{-1} sa fonction réciproque, avec les conventions $\varphi(\infty) = 0$ et $\varphi^{-1}(0) = \infty$.

Exemple : Si $X \sim \mathcal{E}(1)$, alors $\varphi(\lambda) = 1/(1+\lambda)$ donc $\varphi^{-1}(u) = (1-u)/u$ (voir Figure 1.5).

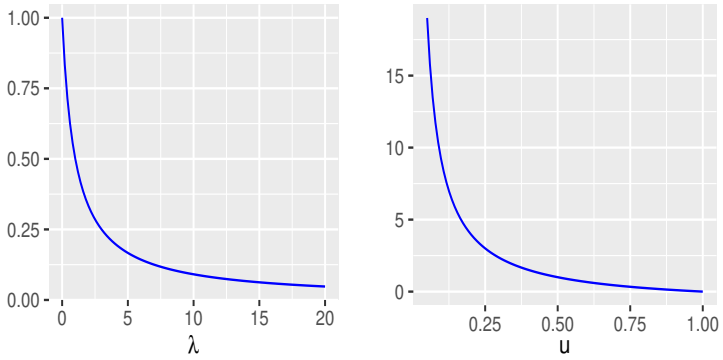

FIGURE 1.5 – Transformée de Laplace $\varphi(\lambda)$ de la loi exponentielle et réciproque $\varphi^{-1}(u)$.

Proposition 4 (Copule archimédienne)
Soit X une variable aléatoire strictement positive, de transformée de Laplace φ. La fonction $C : [0,1]^2 \to [0,1]$ définie par

$$\forall\ 0 \le u, v \le 1 \qquad C(u,v) = \varphi(\varphi^{-1}(u) + \varphi^{-1}(v))$$

est une copule, dite copule archimédienne associée à la loi de X. Si U_1 et U_2 sont uniformes sur $[0,1]$ et indépendantes, et si X a pour transformée de Laplace φ, alors le couple (U,V) défini par $U = \varphi(-\frac{1}{X}\log U_1)$ et $V = \varphi(-\frac{1}{X}\log U_2)$ a des marges uniformes et pour copule C.

Preuve. Le couple (U,V) est bien à valeurs dans $[0,1]\times[0,1]$. Pour $0 \le u, v \le 1$, on a

$$\mathbb{P}(U \le u, V \le v) = \mathbb{P}\left(\varphi(-\frac{1}{X}\log U_1) \le u, \varphi(-\frac{1}{X}\log U_2) \le v\right).$$

La stricte décroissance de φ^{-1} et la positivité de X donnent alors

$$\mathbb{P}(U \le u, V \le v) = \mathbb{P}\left(U_1 \le e^{-\varphi^{-1}(u)X}, U_2 \le e^{-\varphi^{-1}(v)X}\right).$$

Il suffit de conditionner par rapport à X :

$$\mathbb{P}(U \le u, V \le v) = \mathbb{E}\left[\mathbb{P}\left(U_1 \le e^{-\varphi^{-1}(u)X}\middle| X\right)\mathbb{P}\left(U_2 \le e^{-\varphi^{-1}(v)X}\middle| X\right)\right] = \mathbb{E}\left[e^{-(\varphi^{-1}(u)+\varphi^{-1}(v))X}\right],$$

et par définition de la transformée de Laplace :

$$\mathbb{P}(U \le u, V \le v) = \varphi(\varphi^{-1}(u) + \varphi^{-1}(v)) = C(u,v).$$

Dès lors, puisque $\varphi^{-1}(1) = 0$, on a pour $0 \leq u \leq 1$,

$$\mathbb{P}(U \leq u) = \mathbb{P}(U \leq u, V \leq 1) = \varphi(\varphi^{-1}(u) + \varphi^{-1}(1)) = \varphi(\varphi^{-1}(u)) = u.$$

Ceci assure que U suit une loi uniforme, tout comme V. Ainsi $C(u, v)$ définit bien une copule.

■

Exemple : copule de Clayton. Si $X \sim \mathcal{E}(1)$, alors $\varphi(\lambda) = 1/(1+\lambda)$ donc $\varphi^{-1}(u) = (1-u)/u$ et la copule associée à la loi exponentielle, dite copule de Clayton, est

$$C(u, v) = \frac{1}{\frac{1}{u} + \frac{1}{v} - 1}.$$

La Proposition 4 donne un façon très simple de simuler une telle copule.

Le résultat suivant assure qu'on ne change pas la copule d'un couple par transformations strictement croissantes des marginales.

Lemme 1 (Copule et stricte monotonie)
Soit (X, Y) un couple de variables aléatoires continues de copule C. Soit f et g deux fonctions strictement croissantes sur les supports respectifs de X et Y, alors le couple $(f(X), g(Y))$ a également pour copule C.

Remarque : Noter que $(f(X), g(Y))$ est encore un couple de variables continues, d'où l'unicité de la copule.

Simulation : D'après le Théorème de Sklar, un couple aléatoire (X, Y) est complètement caractérisé par la donnée de sa copule et de ses lois marginales. Considérons donc C, F_X et F_Y données et supposons qu'on veuille simuler une réalisation de ce couple. D'après ce qui précède, il "suffit" en général de :

1. simuler (U, V) de lois marginales uniformes et de copule C ;
2. calculer $X = F_X^{-1}(U)$ et $Y = F_Y^{-1}(V)$.

Lorsque F_X^{-1} et F_Y^{-1} sont strictement croissantes, le Lemme 1 assure en effet que (X, Y) a même copule que (U, V). De plus, le premier point de la Proposition 1 nous dit que X et Y ont bien les marginales souhaitées.

Enfin, par le Lemme 1 et le second point de la Proposition 1, pour simuler (U, V) de lois marginales uniformes et de copule C, il suffit de savoir simuler (S, T) de copule C et de fonctions de répartition F_S et F_T strictement croissantes et continues, puis de considérer $U = F_S(S)$ et $V = F_T(T)$. Ceci est illustré en exercice 1.17.

1.5 Exercices

Exercice 1.1 (Loi de Cauchy)
On rappelle que X suit une loi de Cauchy standard si elle a pour densité

$$f(x) = \frac{1}{\pi(1 + x^2)}$$

Grâce à la méthode d'inversion, donner un moyen simple de simuler une telle loi.

Exercice 1.2 (Loi de Pareto)
Soient x_m et α deux réels strictement positifs. Déterminer la densité de la variable

$$X = \frac{x_m}{U^{1/\alpha}},$$

où U suit une loi uniforme sur $[0,1]$. On dit que X suit une loi de Pareto de paramètres (x_m, α).

Exercice 1.3 (Conditionnement et inversion)
On considère le couple (X,Y) de densité

$$f(x,y) = yx^{y-1}e^{-y}\mathbf{1}_{y>0}\mathbf{1}_{0<x<1}.$$

1. Quelle est la loi de Y ?
2. En déduire la loi de X sachant $Y = y$, puis $\mathbb{P}(X \leq x|Y = y)$.
3. Proposer une méthode de simulation du couple (X,Y).

Exercice 1.4 (Conditionnement)
On considère le couple (X,Y) de densité

$$f(x,y) = \frac{1}{\sqrt{8\pi}}e^{-y^2x/2}e^{-\sqrt{x}}\mathbf{1}_{x>0}.$$

1. Quelle est la loi de Y sachant $X = x$?
2. Quelle est la loi de $\sqrt{X}$?
3. En déduire une méthode de simulation du couple (X,Y).

Exercice 1.5 (Maximum)
Soit X et Y deux variables indépendantes de fonctions de répartition F_X et F_Y.

1. Quelle est la fonction de répartition de la variable $T = \max(X,Y)$?
2. En déduire une façon de générer une variable aléatoire de fonction de répartition

$$F(t) = \min(t,1)\left(1 - e^{-t}\right)\mathbf{1}_{t>0}.$$

3. Tracer la densité de cette variable.

Exercice 1.6 (Rejet optimisé)
On veut simuler une loi normale $\mathcal{N}(0,1)$ en utilisant comme proposition une loi de Laplace de paramètre $\lambda > 0$, c'est-à-dire de densité

$$g(x) = \frac{\lambda}{2}e^{-\lambda|x|}.$$

Déterminer la valeur de λ qui permet de minimiser la probabilité de rejet. Vérifier que cette dernière vaut environ 0,24.

Exercice 1.7 (Rejet discret)
Soit $\lambda \in]0,1[$ fixé.

1. Utiliser la méthode de rejet pour simuler une variable X de Poisson de paramètre λ à partir d'une variable Y de loi

$$\forall n \in \mathbb{N} \qquad \mathbb{P}(Y = n) = (1-\lambda)\lambda^n.$$

2. Comment simuler simplement Y ?

3. Quelle est la probabilité de rejet ?

Exercice 1.8 (Loi uniforme sur le disque)

1. Proposer une méthode de rejet pour simuler une variable uniforme sur le disque unité sans utiliser de fonctions trigonométriques. Quelle est la probabilité de rejet ?
2. Donner la loi du couple $(X, Y) = (\sqrt{U}\cos(2\pi V), \sqrt{U}\sin(2\pi V))$ si U et V sont i.i.d. selon une loi uniforme sur $[0, 1]$.

Exercice 1.9 (Rejet et conditionnement)

Soit X une variable aléatoire de fonction de répartition F supposée inversible.

1. Comment simuler la loi de X conditionnellement à $X > a$ à l'aide d'une méthode de rejet ? Que se passe-t-il lorsque a devient grand ?
2. Soit U une variable de loi uniforme sur $[0, 1]$ et T définie par

$$T = F^{-1}(F(a) + (1 - F(a))U).$$

   Déterminer la fonction de répartition de T et en déduire une méthode de simulation de X conditionnellement à $X > a$. Comparer à la méthode de la question précédente.
3. Soit $a > 0$ fixé. On suppose que l'on cherche à simuler X de loi gaussienne centrée réduite conditionnellement à $X > a$. Proposer une méthode de rejet basée sur la loi exponentielle translatée de densité

$$g_\lambda(x) = \lambda e^{-\lambda(x-a)}\mathbf{1}_{x>a}.$$

   Comment choisir le paramètre λ ?

Exercice 1.10 (Durée d'une simulation)

1. On veut générer $n = 1000$ variables exponentielles de paramètre 1. Utiliser la méthode d'inversion vue en cours, tracer l'histogramme et superposer la densité théorique.
2. Même question avec les fonctions dédiées `rexp` et `dexp` sous R.
3. Interpréter les commandes suivantes :

```
> t0 <- proc.time()
> E <- runif(10^6)
> proc.time()-t0
```

   Comparer les durées respectives des méthodes des questions précédentes pour n grand.

Exercice 1.11 (Lien géométrique/exponentielle)

Soit $0 < p < 1$ et X une variable aléatoire suivant une loi géométrique $\mathcal{G}(p)$.

1. Rappeler la loi de X et son espérance.
2. Soit $(B_n)_{n\geq 1}$ une suite de variables i.i.d. selon une loi de Bernoulli $\mathcal{B}(p)$. Comment obtenir une loi géométrique à partir de celles-ci ?
3. Proposer un moyen de simuler B_1 à partir d'une variable uniforme U_1 sur $[0, 1]$. En déduire une simulation de X à partir de lois uniformes pour $p = 1/3$. Retrouver par simulation l'espérance de la loi géométrique. Que se passe-t-il lorsque p est proche de 0 ?
4. Soit $\lambda > 0$ et T une variable aléatoire suivant une loi exponentielle $\mathcal{E}(\lambda)$. Soit $X = \lceil T \rceil$ la partie entière par excès de T (i.e. $\lceil 0.4 \rceil = 1$ et $\lceil 2 \rceil = 2$). Quelles valeurs peut prendre X ? Avec quelles probabilités ? En déduire un nouveau moyen de générer une loi géométrique $\mathcal{G}(p)$. Comparer la vitesse de cette méthode à celle de la question 3 ainsi qu'à celle de `rgeom`.

Exercice 1.12 (Nul n'est censé ignorer la loi normale)

1. Utiliser la méthode vue en exercice 1.6 pour simuler une loi normale centrée réduite à partir d'une loi de Laplace. Estimer le temps nécessaire pour simuler un million de gaussiennes avec cette méthode.
2. Comparer la méthode précédente à celle de Box-Muller.
3. Comparer à celle implémentée sous R via `rnorm`.
4. Simuler $n = 1000$ réalisations d'une loi normale multivariée $\mathcal{N}(m, \Gamma)$ avec

$$m = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \qquad \text{et} \qquad \Gamma = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}.$$

   Retrouver approximativement m et Γ à partir de cet échantillon.

Exercice 1.13 (Méfiez-vous des mélanges !)
On considère une variable X de densité

$$f(x) = \frac{1}{6\sqrt{2\pi}} e^{-\frac{(x+3)^2}{8}} + \frac{2}{3\sqrt{2\pi}} e^{-\frac{(x-3)^2}{2}}.$$

1. Représenter la fonction f.
2. En voyant f comme un mélange de lois, proposer une méthode de simulation de X.
3. Représenter sur un même graphique un histogramme de réalisations de X et la densité f.

Exercice 1.14 (Comparaison de méthodes)
Soit X une variable de fonction de répartition F telle que $F(x) = 0$ si $x \leq 0$, $F(x) = 1$ si $x \geq 1$, et pour tout $x \in [0, 1]$:

$$F(x) = \frac{1}{2}(x + x^2). \tag{1.5}$$

1. Comment simuler X par la méthode d'inversion ? Implémenter cette méthode pour simuler $n = 10000$ réalisations de X. Sur un même graphe, représenter la vraie densité f sur $[0, 1]$ et un estimateur de la densité obtenu à partir de cet échantillon.
2. Majorer la densité de X et en déduire une façon de simuler X par la méthode de rejet. Implémenter cette méthode pour simuler $n = 10000$ réalisations de X.
3. Soit f une densité qui se décompose sous la forme $f(x) = pf_1(x) + (1-p)f_2(x)$, où $p \in]0, 1[$ est connu, et f_1 et f_2 sont deux densités selon lesquelles on sait simuler.
   (a) Expliquer comment simuler une variable aléatoire X ayant pour densité f.
   (b) En déduire une façon de simuler une variable X ayant la fonction de répartition donnée par (1.5). Implémenter cette méthode pour simuler $n = 10000$ réalisations de X.
4. Des trois méthodes précédentes, laquelle choisiriez-vous si vous devez simuler un très grand nombre de variables suivant la loi de X ?

Exercice 1.15 (Loi géométrique conditionnelle)
Soit $p \in]0, 1[$ fixé et X une variable aléatoire suivant une loi géométrique de paramètre p. On se fixe de plus un entier $m \in \mathbb{N}^*$ et on s'intéresse à la loi de X conditionnellement à $X > m$, notée $\mathcal{L}(X|X > m)$.

1. Proposer une méthode de rejet pour simuler une réalisation de $Y \sim \mathcal{L}(X|X > m)$. Quelle est la probabilité de rejet ? Quel est l'inconvénient de cette méthode ?
2. Donner la loi de $Z = X + m$, son espérance et sa variance. En déduire une nouvelle méthode de simulation suivant la loi $\mathcal{L}(X|X > m)$.

3. On veut retrouver $I = \mathbb{E}[X|X > m]$ par simulation. On fixe $p = 1/6$ et $m = 12$. Utiliser la question 2 pour tracer un estimateur $\hat{I}_n$ en fonction de n, pour n allant par exemple de 1 à 1000, les intervalles de confiance à 95% et la droite horizontale $y = I$ (en rouge).
4. Par rapport à la méthode de la question 1, toujours pour $p = 1/6$ et $m = 12$, par quel facteur en moyenne a-t-on divisé le temps de calcul ?

Exercice 1.16 (Loi Gamma)
On veut générer une variable X suivant une loi Gamma $\Gamma(3/2, 1)$, c'est-à-dire de densité

$$f(x) = \frac{2}{\sqrt{\pi}}\, x^{1/2} e^{-x} \mathbf{1}_{x>0}.$$

1. On utilise une technique de rejet avec comme loi de proposition une exponentielle $\mathcal{E}(2/3)$ de densité notée g. Déterminer $m = \sup_{x>0} f(x)/g(x)$. Quelle est le nombre moyen de simulations de la loi exponentielle pour aboutir à une réalisation de la loi Gamma ?
2. Représenter sur un même graphique un histogramme de réalisations de X (obtenues par cette méthode de rejet) et la densité f.
3. Intuitivement, qu'est-ce qui a guidé le choix du paramètre 2/3 comme paramètre de l'exponentielle ? On peut justifier cette intuition : considérer comme proposition la densité g_λ d'une exponentielle $\mathcal{E}(\lambda)$ et déterminer la valeur optimale de λ en terme de probabilité de rejet.

Exercice 1.17 (Représentation et simulation de copules)
Rappelons que pour visualiser une surface $z = f(x, y)$ sur $[0, 1] \times [0, 1]$, une méthode consiste à implémenter la fonction f, discrétiser x et y, appliquer `z=outer(x,y,f)` si f est vectorisable puis, via le package `rgl`, la commande `rgl.surface(x,y,z)`. On peut aussi utiliser `persp(x,y,z)` ou `contour(x,y,z)`.

1. Représenter les surfaces associées aux copules comonotone $C(u, v) = \min(u, v)$, anti-comonotone $C(u, v) = \max(u + v - 1, 0)$, indépendante $C(u, v) = uv$, de Clayton, et gaussienne de paramètre $\rho = 1/2$. Pour cette dernière, on pourra faire appel à la fonction `pmvnorm` du package `mvtnorm`.
2. On veut simuler deux nuages de 1000 points dont chaque marge suit une loi gaussienne centrée réduite. Sur la même fenêtre graphique, représenter à gauche un tel échantillon pour la copule gaussienne de paramètre $\rho = 1/2$, et à droite pour la copule de Clayton.
3. On veut simuler deux nuages de 1000 points dont chaque marge suit une loi de Laplace. Sur la même fenêtre graphique, représenter à gauche un tel échantillon pour la copule gaussienne de paramètre $\rho = 1/2$, et à droite pour la copule indépendante.

Exercice 1.18 (Mouvement brownien, pont brownien)
Un mouvement brownien ou processus de Wiener $(W_t)_{t\geq 0}$ peut être caractérisé par : $W_0 = 0$, les trajectoires de $t \mapsto W_t$ sont presque sûrement continues, W_t est à incréments indépendants avec, pour tout $0 \leq s \leq t$, $W_t - W_s \sim \mathcal{N}(0, t - s)$.

1. On veut simuler de façon approchée un mouvement brownien entre les dates $t = 0$ et $t = 1$. Simuler $n = 100$ variables gaussiennes et en déduire des réalisations de W_t pour $t \in \{1/100, \ldots, 99/100, 1\}$ (on pourra utiliser la fonction `cumsum`). Représenter la trajectoire avec en abscisse le temps et en ordonnée W_t, les points (t, W_t) étant reliés par des segments. Idem avec $n = 10^4$.
2. En dimension 2, un processus $(W_t)_{t\geq 0} = (W_t^x, W_t^y)_{t\geq 0}$ est un mouvement brownien si $(W_t^x)_{t\geq 0}$ et $(W_t^y)_{t\geq 0}$ sont deux mouvements browniens indépendants. Simuler et représenter une trajectoire $(W_t)_{0\leq t\leq 1}$ à partir de $n = 100$ points. Idem avec $n = 10^4$ points. Marquer l'origine d'un point rouge.

3. On revient en dimension 1 et on considère un mouvement brownien $(W_t)_{t\geq 0}$. Le processus $(B_t)_{t\geq 0}$ défini par $B_t = W_t - tW_1$ est appelé pont brownien. Utiliser la question 1 pour représenter des trajectoires d'un pont brownien.
4. Soit $(U_n)_{n\geq 1}$ une suite de variables i.i.d. uniformes sur $[0,1]$, $(F_n)_{n\geq 1}$ la suite des fonctions de répartition empirique et $\|F_n - F\|_\infty$ la distance du sup entre F_n et la fonction de répartition F de la loi uniforme. Le théorème de Kolmogorov-Smirnov dit que, en notant $(B_t)_{t\geq 0}$ un pont brownien,
$$\sqrt{n}\|F_n - F\|_\infty \xrightarrow[n\to\infty]{\mathcal{L}} \sup_{0\leq t\leq 1} |B_t|.$$
La loi de droite est connue sous le nom de loi de Kolmogorov-Smirnov.
   (a) Pour $n = 100$, à l'aide de l'argument `type=''s''`, représenter la fonction
$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{]-\infty,x]}(U_i) = \sum_{j=1}^{n} \frac{j}{n}\mathbf{1}_{[U_{(j,n)},U_{(j+1,n)}[}(x),$$
où $U_{(1,n)} < \cdots < U_{(n,n)}$ est le n-ème échantillon ordonné et $U_{(n+1,n)} = +\infty$. En déduire la représentation de $\sqrt{n}(F_n(x) - F(x))$ sur $[0,1]$ pour $n = 100$ et pour $n = 10^4$.
   (b) Fixons $n = 100$. En tenant compte du fait que
$$\|F_n - F\|_\infty = \max_{1\leq j\leq n}\left\{\max\left(\left|U_{(j,n)} - \frac{j-1}{n}\right|, \left|U_{(j,n)} - \frac{j}{n}\right|\right)\right\},$$
construire un échantillon de taille 1000 selon la loi $\sqrt{n}\|F_n - F\|_\infty$. Sur un même graphique, représenter un estimateur de la densité de cet échantillon et un estimateur de la densité d'un échantillon de taille 1000 de la loi de $\sup_{0\leq t\leq 1}|B_t|$.

Exercice 1.19 (Loi du demi-cercle et Théorème de Wigner)
On considère la densité
$$f(x) = \frac{1}{2\pi}\sqrt{4-x^2}\ \mathbf{1}_{[-2,2]}(x).$$
1. Représenter f sur $[-2,2]$.
2. Donner un majorant de f et en déduire une méthode de rejet pour simuler selon f. En moyenne, combien faut-il d'essais pour obtenir une réalisation X selon la densité f ?
3. Pour $n = 10^3$, implémenter la méthode précédente pour obtenir un échantillon $(X_1, \cdots, X_n)$ i.i.d. selon f. Superposer un estimateur de la densité de cet échantillon et la vraie densité f.
4. Grâce au code précédent, retrouver une estimation du nombre moyen d'essais nécessaires pour obtenir une réalisation X selon la densité f.
5. Pour $n = 100$, considérons $n(n+1)/2$ variables aléatoires $(X_{i,j})_{1\leq i\leq j\leq n}$ i.i.d. gaussiennes centrées réduites et X la matrice carrée symétrique de taille n construite à partir de ces variables. Obtenir les valeurs propres $\lambda_1, \cdots, \lambda_n$ de $X/\sqrt{n}$. Superposer un estimateur de la densité de cet échantillon de valeurs propres et la fonction f.
6. Même question si les variables aléatoires $(X_{i,j})_{1\leq i\leq j\leq n}$ sont i.i.d. uniformes sur $[-\sqrt{3};\sqrt{3}]$.

1.6 Corrigés

Voir la page du cours.

Chapitre 2

Intégration Monte-Carlo

Introduction

Une application classique des méthodes Monte-Carlo est le calcul de quantités du type

$$I = \mathbb{E}[\varphi(X)] = \int \varphi(x) f(x) dx, \tag{2.1}$$

où $\varphi : \mathbb{R}^d \to \mathbb{R}$ est une fonction donnée et X un vecteur aléatoire de densité f suivant laquelle on sait simuler (cf. Chapitre 1). Dans ce contexte, l'estimateur Monte-Carlo de base est défini par

$$\hat{I}_n = \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i), \tag{2.2}$$

où les X_i sont générées de façon i.i.d. selon f. Outre les propriétés de cet estimateur, ce chapitre explique comment on peut éventuellement améliorer sa précision grâce à des techniques de réduction de variance.

2.1 Généralités

2.1.1 Convergence et intervalles de confiance

Il est clair que l'estimateur $\hat{I}_n$ est sans biais, c'est-à-dire que $\mathbb{E}[\hat{I}_n] = I$. Mieux, la loi forte des grands nombres assure qu'il est convergent.

Proposition 5 (Loi des grands nombres)
Si $\mathbb{E}|\varphi(X)| < \infty$, *alors*

$$\hat{I}_n = \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i) \xrightarrow[n\to\infty]{p.s.} I.$$

Revenons sur un exemple élémentaire du chapitre précédent.

Exemple : estimation de π. Supposons que (X, Y) suive la loi uniforme sur le carré $C = [0,1] \times [0,1]$ et que $\varphi(x, y) = \mathbf{1}_{x^2+y^2 \leq 1}$. En notant $D = \{(x, y) \in \mathbb{R}_+^2,\ x^2 + y^2 \leq 1\}$ le quart de disque unité, on a donc

$$I = \iint_C \mathbf{1}_D(x, y) dx dy = \lambda(D) = \frac{\pi}{4}.$$

Simuler des points (X_i, Y_i) uniformément dans C est très facile et la propriété précédente assure donc que

$$\hat{I}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_D(X_i, Y_i) \xrightarrow[n\to\infty]{p.s.} \frac{\pi}{4} \iff 4 \times \hat{I}_n \xrightarrow[n\to\infty]{p.s.} \pi.$$

On dispose donc d'un estimateur Monte-Carlo[1] pour la constante π. Encore faut-il connaître sa précision : c'est tout l'intérêt du théorème central limite.

Proposition 6 (Théorème central limite)
Si $\mathbb{E}\left[\varphi(X)^2\right] < \infty$, *alors*

$$\sqrt{n}\ (\hat{I}_n - I) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2),$$

avec

$$\sigma^2 = \mathrm{Var}(\varphi(X)) = \mathbb{E}[\varphi(X)^2] - \mathbb{E}[\varphi(X)]^2 = \int \varphi(x)^2 f(x) dx - I^2.$$

Ainsi, lorsque n est grand, notre estimateur suit à peu près une loi normale : avec un abus de notation flagrant, on a $\hat{I}_n \approx \mathcal{N}(I, \sigma^2/n)$, c'est-à-dire que $\hat{I}_n$ tend vers I avec une vitesse en $\mathcal{O}(1/\sqrt{n})$. Plus précisément, le TCL doit permettre de construire des intervalles de confiance. Cependant, en général, l'écart-type σ est lui aussi inconnu, il va donc falloir l'estimer. Qu'à cela ne tienne, la méthode Monte-Carlo en fournit justement un estimateur à peu de frais puisque basé sur le même échantillon $(X_1, \ldots, X_n)$, à savoir

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} \varphi(X_i)^2 - \hat{I}_n^2 \xrightarrow[n\to\infty]{p.s.} \sigma^2 \tag{2.3}$$

par la loi des grands nombres pour le premier terme et la Proposition 5 pour le second. Le lemme de Slutsky implique donc que

$$\sqrt{n}\ \frac{\hat{I}_n - I}{\hat{\sigma}_n} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

Exemple : estimation de π. Dans ce cas, la variance vaut tout simplement

$$\sigma^2 = I - I^2 = \frac{\pi}{4}\left(1 - \frac{\pi}{4}\right),$$

laquelle est donc estimée par $\hat{\sigma}_n^2 = \hat{I}_n - \hat{I}_n^2$, et d'après ce qui vient d'être dit

$$\sqrt{n}\ \frac{\hat{I}_n - I}{\sqrt{\hat{I}_n - \hat{I}_n^2}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

On en déduit bien entendu des intervalles de confiance asymptotiques pour I.

Proposition 7 (Intervalles de confiance)
Soit $\alpha \in (0, 1)$ *fixé. Un intervalle de confiance de niveau asymptotique* $1 - \alpha$ *pour* I *est*

$$\left[\hat{I}_n - \Phi^{-1}(1 - \alpha/2)\frac{\hat{\sigma}_n}{\sqrt{n}}\ ;\ \hat{I}_n + \Phi^{-1}(1 - \alpha/2)\frac{\hat{\sigma}_n}{\sqrt{n}}\right],$$

où $\Phi^{-1}(1 - \alpha/2)$ *désigne le quantile d'ordre* $1 - \alpha/2$ *de la loi normale centrée réduite.*

Typiquement $\alpha = 0.05$ donne $\Phi^{-1}(1 - \alpha/2) = q_{0.975} = 1.96 \approx 2$, qui permet de construire un intervalle de confiance à 95% pour I. Un point de précision en passant : prendre 1.96 plutôt que 2 pour le quantile de la loi normale est tout à fait illusoire si le terme $\hat{\sigma}_n/\sqrt{n}$ n'est pas déjà lui-même de l'ordre de 0.01 (i.e., pour un écart-type unité, un échantillon de taille au moins 10 000).

1. tout à fait inutile en pratique, puisqu'il existe des formules analytiques bien plus efficaces pour approcher π.

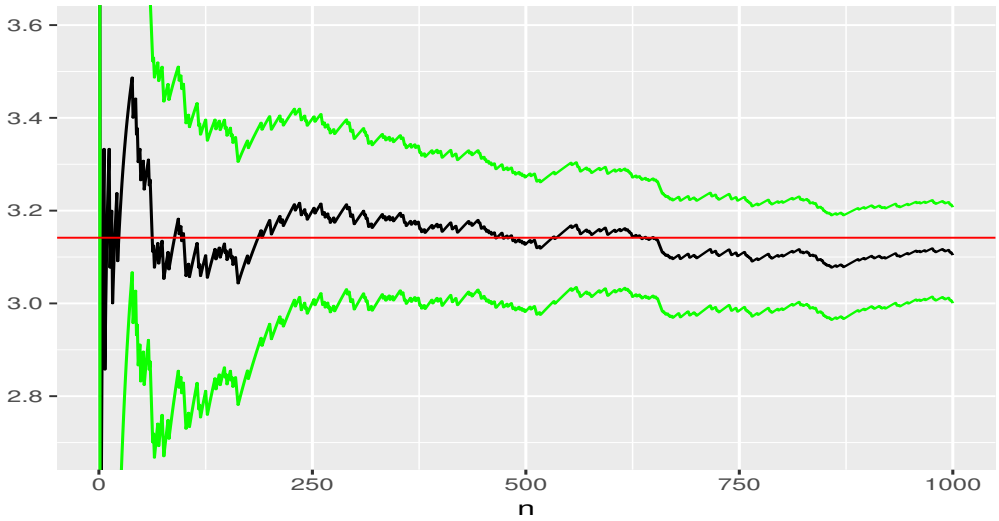

FIGURE 2.1 – Estimateur Monte-Carlo de π et intervalles de confiance asymptotiques.

2.1.2 Quelques mots sur l'intégration numérique

Supposons pour simplifier que, dans l'intégrale définie par (2.1), f soit la densité uniforme sur le pavé $C = [0,1]^d$, c'est-à-dire que

$$I = \int_{[0,1]^d} \varphi(x)dx.$$

D'après ce qui précède, un estimateur Monte-Carlo naturel de I est

$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} \varphi(U_i),$$

où les $U_i = (U_{i,1}, \ldots, U_{i,d})$ sont des vecteurs i.i.d. uniformes sur $[0,1]^d$. Par ailleurs, si la fonction φ est suffisamment régulière, des méthodes déterministes d'intégration numérique permettent également d'approcher cette intégrale.

Commençons par la dimension $d = 1$ et rappelons quelques résultats classiques :

— Dans ce cas, si φ est de classe $\mathcal{C}^1$, la méthode des rectangles sur la subdivision régulière $\{1/n, \ldots, (n-1)/n, 1\}$ a une précision en $\mathcal{O}(1/n)$. Plus précisément, si on convient de noter $M_1 = \sup_{x\in[0,1]} |\varphi'(x)|$ et R_n l'approximation obtenue, alors

$$|R_n - I| \leq \frac{M_1}{2n}.$$

— Si φ est de classe $\mathcal{C}^2$, la méthode des trapèzes sur la même subdivision a une précision en $\mathcal{O}(1/n^2)$. Plus précisément, si on note $M_2 = \sup_{x\in[0,1]} |\varphi''(x)|$ et T_n l'approximation obtenue, alors

$$|T_n - I| \leq \frac{M_2}{12n^2}.$$

— La méthode de Simpson consiste à approcher f sur chaque segment $[k/n, (k+1)/n]$ par un arc de parabole qui coïncide avec f aux deux extrémités et au milieu de ce segment. Si φ est de classe $\mathcal{C}^4$, cette méthode a une précision en $\mathcal{O}(1/n^4)$. Plus précisément, si on note $M_4 = \sup_{x\in[0,1]} |\varphi^{(4)}(x)|$ et S_n l'approximation obtenue, alors

$$|S_n - I| \leq \frac{M_4}{2880n^4}.$$

— De façon générale, si φ est de classe $\mathcal{C}^s$, une méthode adaptée à cette régularité (de type Newton-Cotes) permettra d'obtenir une erreur en $\mathcal{O}(1/n^s)$. Si l'on compare à l'erreur en $\mathcal{O}(1/\sqrt{n})$ de la méthode Monte-Carlo, il est donc clair que les méthodes déterministes sont préférables dès que f est $\mathcal{C}^1$.

Passons maintenant au cas $d = 2$ et focalisons-nous sur la plus simple des méthodes déterministes, à savoir celle des "rectangles" : on somme donc cette fois des volumes de pavés. Si l'on considère

$$I = \int_0^1 \int_0^1 \varphi(x,y)dxdy = \sum_{k,\ell=0}^{n-1} \int_{k/n}^{(k+1)/n} \int_{\ell/n}^{(\ell+1)/n} \varphi(x,y)dxdy$$

et son approximation numérique

$$R_n = \sum_{k,\ell=0}^{n-1} \int_{k/n}^{(k+1)/n} \int_{\ell/n}^{(\ell+1)/n} \varphi(k/n,\ell/n)dxdy = \frac{1}{n^2} \sum_{k,\ell=0}^{n-1} \varphi(k/n,\ell/n),$$

il vient

$$|I - R_n| \leq \sum_{k,\ell=0}^{n-1} \int_{k/n}^{(k+1)/n} \int_{\ell/n}^{(\ell+1)/n} |\varphi(x,y) - \varphi(k/n,\ell/n)|\, dxdy.$$

Par le théorème des accroissements finis (on rappelle que φ est à valeurs dans $\mathbb{R}$), pour tout couple (k,ℓ), il existe $c_{k,\ell} \in [k/n,(k+1)/n] \times [\ell/n,(\ell+1)/n]$ tel que

$$|\varphi(x,y) - \varphi(k/n,\ell/n)| = \left| \frac{\partial \varphi}{\partial x}(c_{k,\ell})(x - k/n) + \frac{\partial \varphi}{\partial y}(c_{k,\ell})(y - \ell/n) \right| \leq \frac{1}{n} \|\nabla \varphi(c_{k,\ell})\|_1.$$

En notant $M_1 = \sup_{0\leq x,y\leq 1} \|\nabla \varphi(x,y)\|_1$, on obtient donc

$$|I - R_n| \leq \frac{M_1}{n}.$$

Néanmoins, pour comparer ce qui est comparable, il faut considérer que les deux méthodes font le même nombre n d'appels [2] à la fonction φ. La subdivision $\{0, 1/n, \ldots, (n-1)/n\}$ de la dimension 1 est donc remplacée par le maillage de n points $A_{k,\ell} = (k/\sqrt{n}, \ell/\sqrt{n})$, avec $1 \leq k, \ell \leq \sqrt{n}$. La précision de la méthode s'en ressent puisqu'elle est alors logiquement en $\mathcal{O}(1/\sqrt{n})$, tout comme la méthode Monte-Carlo basée sur n appels à la fonction φ.

De façon générale, si φ est de classe $\mathcal{C}^s$ sur $[0,1]^d$, alors une méthode déterministe adaptée à cette régularité et faisant n appels à la fonction φ permettra d'atteindre une vitesse en $\mathcal{O}(n^{-s/d})$. Cette vitesse qui s'effondre avec d est symptomatique du fléau de la dimension. A contrario, la méthode Monte-Carlo, avec une vitesse en $1/\sqrt{n}$, est insensible à la dimension et peut donc s'avérer plus avantageuse dès que l'on travaille en dimension grande ou sur une fonction irrégulière. En fin de chapitre, nous dirons un mot des méthodes quasi Monte-Carlo, lesquelles représentent un compromis entre les méthodes déterministes et les méthodes Monte-Carlo.

2.1.3 Le cadre bayésien

De façon générale, les méthodes Monte-Carlo sont d'usage constant en statistique bayésienne. En particulier, celle-ci requiert souvent le calcul d'intégrales du type (2.1). Expliquons pourquoi en quelques mots.

Le cadre typique est celui où la loi de la variable X dépend d'un paramètre θ. Dans l'approche fréquentiste, θ est inconnu mais supposé avoir une valeur fixée et les observations $\mathbf{X} = (X_1, \ldots, X_N)$

2. Un appel pouvant être coûteux dès lors que φ est compliquée, ce critère fait sens.

permettent de l'estimer par une méthode donnée, par exemple au maximum de vraisemblance. L'approche bayésienne est différente : elle consiste à considérer que θ est lui-même une variable aléatoire $\boldsymbol{\theta}$ et suit une loi (dite a priori) donnée, les observations $\mathbf{X} = (X_1, \ldots, X_N)$ permettant d'affiner cette loi.

Plus formellement, notons π la densité de la loi a priori de $\boldsymbol{\theta}$ (ou prior) et $f(\mathbf{x}|\theta)$ la densité conditionnelle de $\mathbf{X}$ sachant $\boldsymbol{\theta} = \theta$ (ou vraisemblance). Nous supposons pour simplifier les écritures que toutes ces densités sont définies par rapport à la mesure de Lebesgue. Par la règle de Bayes, la densité a posteriori de $\boldsymbol{\theta}$ sachant l'observation $\mathbf{X}$ (ou posterior) s'écrit alors tout simplement

$$\pi(\theta|\mathbf{X}) = \frac{f(\mathbf{X}|\theta)\pi(\theta)}{f(\mathbf{X})} = \frac{f(\mathbf{X}|\theta)\pi(\theta)}{\displaystyle\int f(\mathbf{X}|t)\pi(t)dt}.$$

Puisqu'on cherche une densité par rapport à la variable θ, tout ce qui ne dépend pas de θ joue le rôle d'une constante de normalisation, d'où l'utilisation du symbole $\propto$ ("proportionnel à"). Ainsi l'écriture

$$\pi(\theta|\mathbf{X}) \propto f(\mathbf{X}|\theta)\pi(\theta)$$

signifie qu'il existe une constante de normalisation $c(\mathbf{X})$, ne faisant pas intervenir θ, telle que

$$\pi(\theta|\mathbf{X}) = c(\mathbf{X}) f(\mathbf{X}|\theta)\pi(\theta).$$

Exemple : Modèle gaussien. Supposons que $\boldsymbol{\theta} \sim \mathcal{N}(0,1)$ et que, sachant $\boldsymbol{\theta} = \theta$, $\mathbf{X} = (X_1, \ldots, X_N)$ soit un échantillon i.i.d. de loi $\mathcal{N}(\theta, 1)$. Alors pour tout $\mathbf{X} = (X_1, \ldots, X_N) \in \mathbb{R}^N$, la vraisemblance s'écrit

$$\pi(\mathbf{X}|\theta) = \prod_{i=1}^{N} \tfrac{1}{\sqrt{2\pi}} e^{-\frac{(X_i-\theta)^2}{2}} = \tfrac{1}{(2\pi)^{N/2}} \exp\left\{-\tfrac{1}{2}\sum_{i=1}^{N}(X_i-\theta)^2\right\},$$

d'où pour la densité a posteriori

$$\pi(\theta|\mathbf{X}) = \frac{\frac{1}{(2\pi)^{N/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}(X_i-\theta)^2\right\} \times \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{\theta^2}{2}\right\}}{f(\mathbf{X})} \propto \exp\left\{-\tfrac{1}{2}((N+1)\theta^2 - 2N\bar{X}_N\theta)\right\}.$$

Il suffit alors de faire apparaître une densité gaussienne pour en déduire la loi a posteriori :

$$\pi(\theta|\mathbf{X}) \propto \exp\left\{-\tfrac{N+1}{2}\left(\theta - \tfrac{N}{N+1}\bar{X}_N\right)^2\right\},$$

donc

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{X}) = \mathcal{N}\left(\tfrac{N}{N+1}\bar{X}_N, \tfrac{1}{N+1}\right).$$

La loi a posteriori est une loi normale de moyenne $N\bar{X}_N/(N+1)$ (qui dépend des observations) et de variance $1/(N+1)$ (qui ne dépend pas des observations). C'est la loi (aléatoire, car fonction de $\mathbf{X}$) ayant la densité suivante par rapport à la mesure de Lebesgue sur $\mathbb{R}$:

$$\pi(\theta|\mathbf{X}) = \sqrt{\frac{N+1}{2\pi}} \exp\left\{-\tfrac{N+1}{2}\left(\theta - \tfrac{N}{N+1}\bar{X}_N\right)^2\right\}.$$

Par rapport à la loi a priori (i.e. la gaussienne standard), la loi a posteriori est donc centrée grosso modo autour de la moyenne empirique $\bar{X}_N$ des données observées et est bien plus concentrée autour de cette moyenne que ne l'était la loi a priori autour de 0. Autrement dit, les observations $X_1, \ldots, X_N$ ont apporté de l'information sur le paramètre inconnu et aléatoire $\boldsymbol{\theta}$.

Ceci fait, on pourra s'intéresser à une valeur moyenne par rapport à cette loi a posteriori, laquelle s'écrira donc

$$\mathbb{E}[\varphi(\boldsymbol{\theta})|\mathbf{X}] = \int \varphi(\theta)\pi(\theta|\mathbf{X})d\theta = \frac{\int \varphi(\theta)f(\mathbf{X}|\theta)\pi(\theta)d\theta}{\int f(\mathbf{X}|\theta)\pi(\theta)d\theta}. \tag{2.4}$$

En particulier, on appelle estimateur de Bayes (sous-entendu : pour la perte L_2 ou perte quadratique) la moyenne a posteriori, qui vaut donc

$$\hat{\theta}_N(\mathbf{X}) = \mathbb{E}[\boldsymbol{\theta}|\mathbf{X}] = \int \theta\ \pi(\theta|\mathbf{X})d\theta = \frac{\int \theta f(\mathbf{X}|\theta)\pi(\theta)d\theta}{\int f(\mathbf{X}|\theta)\pi(\theta)d\theta}.$$

Exemple. Dans l'exemple précédent, puisque sachant $\mathbf{X}$, la variable $\boldsymbol{\theta}$ suit une loi gaussienne de moyenne $N\bar{X}_N/(N+1)$ et de variance $1/(N+1)$, l'estimateur de Bayes pour la perte quadratique est tout simplement

$$\hat{\theta}_N(\mathbf{X}) = \tfrac{N}{N+1}\bar{X}_N.$$

Remarque. Si l'on adopte une approche fréquentiste et que l'on considère une réalisation fixée θ_0 de la variable aléatoire $\boldsymbol{\theta}$, on voit que, lorsque N croît, l'estimateur de Bayes diffère de moins en moins de la moyenne empirique $\bar{X}_N$, qui est dans ce modèle l'estimateur du maximum de vraisemblance. Dans ce cas, cette moyenne empirique se concentrant elle-même autour de θ_0, la loi a posteriori se concentre autour de θ_0.

Contrairement à ce que pourrait laisser croire cet exemple jouet, le calcul de l'intégrale (2.4) est en général impossible analytiquement et difficile par intégration numérique déterministe, en particulier si le paramètre θ est multidimensionnel. On a donc naturellement recours à des méthodes Monte-Carlo.

Supposons en effet que l'on sache simuler suivant la loi a priori $\pi(\theta)$ et, pour tout θ, évaluer la quantité $\varphi(\theta)f(\mathbf{x}|\theta)$. On retombe alors exactement dans le cadre d'application des méthodes Monte-Carlo d'intégration, puisqu'il suffit de générer des réalisations i.i.d. $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$ selon $\pi(\theta)$ pour en déduire que, sachant $\mathbf{X}$,

$$\frac{1}{n}\sum_{i=1}^{n} \varphi(\boldsymbol{\theta}_i)f(\mathbf{X}|\boldsymbol{\theta}_i) \xrightarrow[n\to\infty]{p.s.} \int \varphi(\theta)f(\mathbf{X}|\theta)\pi(\theta)d\theta.$$

L'estimation du dénominateur de (2.4) correspond au cas particulier où $\varphi = 1$ et se traite donc de la même façon. Au total, l'estimateur Monte-Carlo de $I = \mathbb{E}[\varphi(\boldsymbol{\theta})|\mathbf{X}]$ s'écrit

$$\hat{I}_n = \frac{\sum_{i=1}^{n} \varphi(\boldsymbol{\theta}_i)f(\mathbf{X}|\boldsymbol{\theta}_i)}{\sum_{i=1}^{n} f(\mathbf{X}|\boldsymbol{\theta}_i)}.$$

En particulier, l'estimateur de Bayes $\hat{\theta}_N(\mathbf{X}) = \mathbb{E}[\boldsymbol{\theta}|\mathbf{X}]$ pour N observations $\mathbf{X} = (X_1, \ldots, X_N)$ admet lui-même pour estimateur Monte-Carlo basé sur n simulations $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$:

$$\hat{\theta}_N^n(\mathbf{X}) = \frac{\sum_{i=1}^{n} \boldsymbol{\theta}_i\ f(\mathbf{X}|\boldsymbol{\theta}_i)}{\sum_{i=1}^{n} f(\mathbf{X}|\boldsymbol{\theta}_i)}.$$

Voir l'exercice 2.15 pour un exemple d'application.

Remarque : d'un point de vue plus général, l'idéal serait en fait de disposer d'un échantillon i.i.d. selon la loi a posteriori, ce qui est bien plus difficile. On peut néanmoins évoquer une situation où

ceci peut se faire assez simplement. Supposons en effet que l'on adopte une méthode de rejet avec la loi a priori comme loi instrumentale : il faut alors trouver m tel que

$$\sup_\theta \frac{\pi(\theta|\mathbf{X})}{\pi(\theta)} \leq m \iff \sup_\theta \frac{f(\mathbf{X}|\theta)}{f(\mathbf{X})} \leq m \iff m = \frac{f(\mathbf{X}|\tilde{\theta})}{f(\mathbf{X})},$$

où, sous réserve d'existence et d'unicité, $\tilde{\theta} = \tilde{\theta}(\mathbf{X})$ est l'estimateur du maximum de vraisemblance. Notons qu'on a alors tout simplement

$$\frac{\pi(\theta|\mathbf{X})}{m\pi(\theta)} = \frac{f(\mathbf{X}|\theta)}{f(\mathbf{X}|\tilde{\theta})}.$$

Ainsi, dès lors que l'on sait simuler selon la loi a priori et, pour l'observation $\mathbf{X}$, calculer en tout θ la vraisemblance ainsi que l'estimateur du maximum de vraisemblance, il suffit de procéder comme suit :

— Pour l'observation $\mathbf{X}$, calculer l'estimateur du maximum de vraisemblance $\tilde{\theta}$;
— simuler des couples $(\boldsymbol{\theta}_k, U_k)$ i.i.d. selon la loi a priori $\pi(\theta)$ et la loi uniforme jusqu'à ce que

$$U_k \leq \frac{f(\mathbf{X}|\theta)}{f(\mathbf{X}|\tilde{\theta})}.$$

Le point remarquable de cette méthode est qu'elle ne nécessite pas le calcul de la densité marginale $f(\mathbf{X})$, qui est souvent très compliquée.

Exemple : si, sachant θ, les $(X_i)_{1\leq i\leq N}$ sont i.i.d. selon la loi $\mathcal{N}(\theta, 1)$, alors on sait que l'estimateur du maximum de vraisemblance est tout bonnement $\tilde{\theta} = \bar{X}_N$ et, après simplifications,

$$\frac{f(\mathbf{X}|\theta)}{f(\mathbf{X}|\tilde{\theta})} = \exp\left\{-\frac{N}{2}(\theta - \bar{X}_N)^2\right\}.$$

2.1.4 Tests d'hypothèses

Dans un cadre fréquentiste, on considère un test d'hypothèse $T(\mathbf{X}) \in \{0, 1\}$ consistant à accepter (respectivement rejeter) H_0 si $T(\mathbf{X}) = 0$ (respectivement $T(\mathbf{X}) = 1$). Très souvent, le test est obtenu par seuillage d'une statistique $S(\mathbf{X})$, i.e. on rejette H_0 au niveau α (erreur de première espèce) si et seulement si $S(\mathbf{X}) > c_\alpha$.

Exemple. Soit $\mathbf{X} = (X_1, \ldots, X_N)$ i.i.d. selon une loi $\mathcal{N}(\theta, 1)$, avec $\theta \in \mathbb{R}$ paramètre inconnu. On veut tester

$$H_0 \ : \ \theta = 0 \qquad \text{contre} \qquad H_1 \ : \ \theta \neq 0.$$

Puisque, sous H_0, la moyenne empirique $\bar{X}_n$ suit la loi $\mathcal{N}(0, 1/N)$, on a dans ce cas :

$$\mathbb{P}(|\sqrt{N}\bar{X}_N| > \Phi^{-1}(1 - \alpha/2)) = \alpha,$$

et on a bien construit un test de niveau α en seuillant la statistique $S(\mathbf{X}) = |\sqrt{N}\bar{X}_N|$ au niveau $c_\alpha = \Phi^{-1}(1 - \alpha/2)$, où Φ^{-1} correspond comme d'habitude à la fonction quantile de la loi normale standard.

Revenons au cadre général. Pour une réalisation $\mathbf{x}$, la p-value (ou probabilité critique, ou niveau de significativité) est alors définie par

$$\alpha_0(\mathbf{x}) = \inf\{\alpha \in [0, 1],\ H_0 \text{ est rejetée au niveau } \alpha\} = \inf\{\alpha \in [0, 1],\ S(\mathbf{x}) > c_\alpha\}.$$

Rappelons qu'une p-value très faible signifie que H_0 est très peu vraisemblable. Si H_0 est une hypothèse simple, i.e. $\theta = \theta_0$, alors sous certaines hypothèses (par exemple lorsque, sous H_0, la fonction de répartition de la variable aléatoire $S(\mathbf{X})$ est bijective), on peut montrer que, sous H_0,

$$\alpha_0(\mathbf{x}) = \mathbb{P}(S(\mathbf{X}) > S(\mathbf{x})),$$

où $\mathbf{X}$ est aléatoire suivant la loi prescrite par $H_0 : \theta = \theta_0$, et $S(\mathbf{x})$ fixée. On résume souvent ceci par la phrase : « La p-value est la probabilité, sous H_0, d'obtenir une statistique de test au moins aussi extrême que celle observée. »

Exemple. Dans l'exemple précédent, pour une réalisation $\mathbf{x}$, la p-value s'obtient via la première formulation comme suit :

$$\alpha_0(\mathbf{x}) = \inf\{\alpha \in [0,1],\ S(\mathbf{x}) > \Phi^{-1}(1-\alpha/2)\} = \inf\{\alpha \in [0,1],\ \alpha > 2(1-\Phi(S(\mathbf{x}))\} = 2(1-\Phi(S(\mathbf{x})).$$

On retrouve facilement ce résultat par la seconde formulation : sous H_0, puisque $S(\mathbf{X}) = |\sqrt{n}\bar{X}_n| \sim |\mathcal{N}(0,1)|$, on a

$$\mathbb{P}(S(\mathbf{X}) > S(\mathbf{x})) = \mathbb{P}(|\mathcal{N}(0,1)| > S(\mathbf{x})) = 2(1 - \Phi(S(\mathbf{x}))) = \alpha_0(\mathbf{x}).$$

Sur cet exemple élémentaire, toutes les lois sont explicites et connues, mais ce n'est pas toujours le cas. Supposons un test consistant à rejeter H_0 au niveau α si et seulement si $S(\mathbf{X}) > c_\alpha$ et que, pour une réalisation $\mathbf{x}$, on veuille calculer la p-value associée. Si on sait simuler $\mathbf{X}$ sous H_0, alors un estimateur Monte-Carlo $\hat{\alpha}_0^n(\mathbf{x})$ de $\alpha_0(\mathbf{x})$ s'obtient tout simplement en simulant des variables $\mathbf{X}_i$ i.i.d. sous H_0 et en considérant

$$\hat{\alpha}_0^n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{S(\mathbf{X}_i)>S(\mathbf{x})}.$$

Cet estimateur est non biaisé, consistant et asymptotiquement normal avec

$$\sqrt{n}(\hat{\alpha}_0^n(\mathbf{x}) - \alpha_0(\mathbf{x})) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \alpha_0(\mathbf{x})(1 - \alpha_0(\mathbf{x}))).$$

Un intervalle de confiance de niveau asymptotique 95% pour $\alpha_0(\mathbf{x})$ est ainsi donné par

$$\left[\hat{\alpha}_0^n(\mathbf{x}) - 2\frac{\sqrt{\hat{\alpha}_0^n(\mathbf{x})(1-\hat{\alpha}_0^n(\mathbf{x}))}}{\sqrt{n}}\ ;\ \hat{\alpha}_0^n(\mathbf{x}) + 2\frac{\sqrt{\hat{\alpha}_0^n(\mathbf{x})(1-\hat{\alpha}_0^n(\mathbf{x}))}}{\sqrt{n}}\right].$$

Remarque : Supposons que H_0 n'est pas vraie, alors la p-value peut être très petite, par exemple $\alpha_0(\mathbf{x}) = 10^{-9}$. Dans ce cas, il est clair qu'à moins de prendre n de l'ordre de 10^9, l'estimateur Monte-Carlo renverra typiquement la valeur 0. Si l'on souhaite une estimation précise de cette p-value, il faut alors faire appel à des méthodes Monte-Carlo pour événements rares.

2.2 Réduction de variance

Nous restons dans le cadre de la section précédente, à savoir l'estimation de l'intégrale $I = \mathbb{E}[\varphi(X)]$. Son estimation $\hat{I}_n$ par la méthode Monte-Carlo standard aboutit à une erreur en $\sigma/\sqrt{n}$. Dès que la dimension de X est grande ou φ irrégulière, les méthodes déterministes ou quasi-Monte-Carlo ne sont plus concurrentielles et la vitesse en $1/\sqrt{n}$ apparaît donc incompressible. L'idée est alors de gagner sur le facteur σ, ce qui est précisément l'objet des méthodes de réduction de variance.

Une remarque élémentaire au préalable : pour une précision ε voulue sur le résultat, donc en $\sigma/\sqrt{n}$ par Monte-Carlo classique, une méthode de réduction de variance permettant de diviser la variance

σ^2 par 2 permet de diviser le nombre n de simulations nécessaires par 2 pour atteindre la même précision. Si la nouvelle méthode n'est pas plus coûteuse en temps de calcul, c'est donc la durée de la simulation qui est ainsi divisée par deux. Si la nouvelle méthode requiert beaucoup plus de calculs, il faut en toute rigueur en tenir compte, par exemple en définissant l'efficacité d'une technique par

$$\text{Efficacité} = \frac{1}{\text{Complexité} \times \text{Variance}},$$

et en comparant les efficacités des différentes méthodes entre elles. Ceci n'est néanmoins pas toujours facile.

2.2.1 Echantillonnage préférentiel

On souhaite estimer

$$I = \mathbb{E}[\varphi(X)] = \int \varphi(x) f(x) dx.$$

Si la fonction φ prend ses plus grandes valeurs là où la densité f est très faible, i.e. là où X a très peu de chances de tomber, et réciproquement, l'estimateur Monte-Carlo classique

$$\hat{I}_n = \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i)$$

est très mauvais puisqu'à moins de prendre n très grand, il va estimer environ 0 même si I vaut 1.

Exemples :

1. Evénements rares : la variable X suivant une loi normale centrée réduite, on veut estimer la probabilité

$$\mathbb{P}(X > 6) = 1 - \Phi(6) = \mathbb{E}[\mathbf{1}_{X>6}] = \int_{\mathbb{R}} \mathbf{1}_{x>6} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

   La commande `1-pnorm(6)` en R montre qu'elle est de l'ordre de 10^{-9}. Ceci signifie qu'à moins de prendre n de l'ordre d'au moins un milliard, on a toutes les chances d'obtenir $\hat{I}_n = 0$. Un cas particulier est celui où, en section précédente, on voudrait estimer de façon précise (i.e. pas par zéro) une p-value alors qu'elle est très faible. Ceci est crucial dans certains domaines applicatifs où l'on doit assurer un risque de première espèce extrêmement faible (normes de sécurité drastiques, etc.).

2. Plus siouxe : soit m un réel, $X \sim \mathcal{N}(m, 1)$ et $\varphi(x) = \exp(-mx + m^2/2)$. Pour tout m, on a donc

$$I = \mathbb{E}[\varphi(X)] = \int_{\mathbb{R}} \varphi(x) f(x) dx = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 1.$$

   Or 95% des X_i tombent dans l'intervalle $[m-2, m+2]$ tandis que $\varphi(m) = \exp(-m^2/2)$ tend très vite vers 0 quand m augmente. Ainsi, dès lors que m est grand, on obtient $\hat{I}_n$ proche de 0, ce qui n'est clairement pas une approximation souhaitable de la valeur cherchée $I = 1$.

L'idée de l'échantillonnage préférentiel, ou échantillonnage pondéré, ou *importance sampling*, est de tirer des points non pas suivant la densité f de X mais selon une densité auxiliaire g réalisant un compromis entre les régions de l'espace où φ est grande et où la densité f est élevée, quitte à rétablir le tir ensuite en tenant compte du fait que la loi de simulation g n'est pas la loi initiale f.

Mathématiquement, il s'agit simplement d'une réécriture de I sous la forme

$$I = \mathbb{E}[\varphi(X)] = \int \varphi(x) f(x) dx = \int \frac{f(y)}{g(y)} \varphi(y) g(y) dy = \int w(y) \varphi(y) g(y) dy = \mathbb{E}[w(Y)\varphi(Y)],$$

où Y a pour densité g et $w(y) = f(y)/g(y)$ correspond à la pondération (ou rapport de vraisemblance) dû au changement de loi.

Attention ! Pour que celui-ci soit bien défini, il faut s'assurer qu'on ne divise pas par 0, c'est-à-dire que $g(y) = 0$ implique $f(y)\varphi(y) = 0$. Concrètement, il faut absolument que g mette du poids partout où le produit φf est non nul, sans quoi l'estimateur d'échantillonnage préférentiel est voué à l'échec.

Ceci supposé, si l'on sait :

(a) simuler suivant la densité g,

(b) calculer le rapport de vraisemblance $w(y) = f(y)/g(y)$ pour tout y,

l'estimateur par échantillonnage préférentiel prend la forme

$$\tilde{I}_n = \frac{1}{n}\sum_{i=1}^{n} w(Y_i)\varphi(Y_i),$$

où les Y_i sont i.i.d. de densité g. Mutatis mutandis, les résultats vus pour $\hat{I}_n$ s'appliquent à nouveau ici.

Proposition 8 (Echantillonnage préférentiel)
Si $\mathbb{E}|w(Y)\varphi(Y)| < \infty$, alors l'estimateur $\tilde{I}_n$ est sans biais et convergent, c'est-à-dire que $\mathbb{E}[\tilde{I}_n] = I$ et

$$\tilde{I}_n = \frac{1}{n}\sum_{i=1}^{n} w(Y_i)\varphi(Y_i) \xrightarrow[n\to\infty]{p.s.} \mathbb{E}[w(Y)\varphi(Y)] = I.$$

Si de plus $\mathbb{E}[w(Y)^2\varphi(Y)^2] < \infty$, alors

$$\sqrt{n}\ (\tilde{I}_n - I) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, s^2),$$

avec

$$s^2 = \mathrm{Var}(w(Y)\varphi(Y)) = \int w(y)^2\varphi(y)^2 g(y)dy - I^2 = \mathbb{E}[w(X)\varphi(X)^2] - I^2.$$

Remarque. Comme précédemment, la variance s^2 est naturellement estimée par

$$\tilde{s}_n^2 = \frac{1}{n}\sum_{i=1}^{n} w(Y_i)^2\varphi(Y_i)^2 - \tilde{I}_n^2,$$

d'où l'on peut déduire des intervalles de confiance asymptotiques.

La variance s^2 de $\tilde{I}_n$ est à comparer à la variance $\sigma^2 = \mathbb{E}[\varphi^2(X)] - I^2$ de $\hat{I}_n$. Il "suffit" donc de choisir la pondération w, c'est-à-dire la densité instrumentale g, de sorte que le terme $\mathbb{E}[w(X)\varphi^2(X)]$ soit le plus petit possible. La bonne nouvelle, c'est que ce problème a une solution explicite, la mauvaise c'est qu'elle est tout à fait hors d'atteinte...

Lemme 1 (Loi d'échantillonnage optimale)
Pour toute densité g telle que $\mathbb{E}[w(Y)^2\varphi(Y)^2] < \infty$, on a

$$s^2 = \mathrm{Var}(w(Y)\varphi(Y)) \geq \mathbb{E}[|\varphi(X)|]^2 - I^2 = \left(\int |\varphi(x)|f(x)dx\right)^2 - \left(\int \varphi(x)f(x)dx\right)^2,$$

la borne inférieure étant atteinte pour la densité $g^\star$ définie par

$$g^\star(y) = \frac{|\varphi(y)|f(y)}{\displaystyle\int |\varphi(y)|f(y)dy}.$$

Preuve. Toute variable aléatoire étant de variance positive, il vient

$$s^2 = \mathbb{E}[w(Y)^2\varphi(Y)^2] - I^2 = \mathbb{E}[(w(Y)|\varphi(Y)|)^2] - I^2 \geq \mathbb{E}[w(Y)|\varphi(Y)|]^2 - I^2 = \mathbb{E}[|\varphi(X)|]^2 - I^2.$$

L'égalité dans l'inégalité précédente n'est possible que si la variable aléatoire $w(Y)|\varphi(Y)|$ est p.s. constante. La variable Y étant de densité g, ceci signifie qu'il existe une constante c telle que pour tout y tel que $g(y) > 0$, on a

$$w(y))|\varphi(y)| = c \iff g(y) = \frac{|\varphi(y)|f(y)}{c} \iff g(y) = \frac{|\varphi(y)|f(y)}{\displaystyle\int |\varphi(y)|f(y)dy},$$

la dernière équivalence découlant du fait que g est une densité.

■

Si φ est de signe constant, la variance obtenue avec $g^\star$ est nulle, ce qui signifie qu'un seul tirage selon $g^\star$ suffit ! En effet, si $Y \sim g^\star$, alors

$$\tilde{I}_1 = w(Y)\varphi(Y) = \frac{f(Y)}{g^\star(Y)}\varphi(Y) = \int \varphi(x)f(x)dx = I.$$

Bien entendu, il y a deux obstacles : d'une part, rien n'assure que l'on sache simuler selon $g^\star$; d'autre part, même si on savait le faire, le calcul du rapport de vraisemblance $w(Y) = I/\varphi(Y)$ nécessite la connaissance de I, qui est précisément la quantité cherchée !

Même si la proposition précédente présente surtout un intérêt théorique, on retrouve l'idée selon laquelle la densité auxiliaire g doit réaliser un compromis entre la fonction à intégrer φ et la densité f : autant que possible, g doit mettre du poids là où le produit $|\varphi(x)|f(x)$ est le plus élevé.

2.2.2 Conditionnement

On cherche toujours à estimer $I = \mathbb{E}[\varphi(X)]$ en supposant $\mathbb{E}\left[\varphi^2(X)\right] < \infty$. L'idée force dans cette section a été vue en cours de probabilités, à savoir : le conditionnement ne change pas la moyenne, mais réduit l'incertitude. Appliqué à notre contexte, ceci signifie que pour toute autre variable aléatoire Y, on a d'une part

$$\mathbb{E}[\mathbb{E}[\varphi(X)|Y]] = \mathbb{E}[\varphi(X)] = I.$$

D'autre part, puisque $\varphi(X)$ est de carré intégrable, l'espérance conditionnelle est la projection orthogonale (définie p.s.) de $\varphi(X)$ sur le sous-espace[3] $\sigma(Y)$ des variables aléatoires de la forme $h(Y)$ où h est borélienne et telle que $\mathbb{E}[h(Y)^2] < \infty$. Par le Théorème de Pythagore, on a donc (voir Figure 2.2)

$$\mathbb{E}[\varphi(X)^2] = \mathbb{E}[\mathbb{E}[\varphi(X)|Y]^2] + \mathbb{E}[(\varphi(X) - \mathbb{E}[\varphi(X)|Y])^2] \geq \mathbb{E}[\mathbb{E}[\varphi(X)|Y]^2].$$

Puisque $\mathbb{E}[\mathbb{E}[\varphi(X)|Y]] = \mathbb{E}[\varphi(X)]$, on peut soustraire le carré de cette quantité des deux côtés pour aboutir à

$$\sigma^2 = \mathrm{Var}(\varphi(X)) \geq \mathrm{Var}(\mathbb{E}[\varphi(X)|Y]) = s^2.$$

Remarque. Une autre façon de le voir est d'appliquer la formule dite de la variance totale, à savoir

$$\mathrm{Var}(\varphi(X)) = \mathrm{Var}(\mathbb{E}[\varphi(X)|Y]) + \mathbb{E}[\mathrm{Var}(\varphi(X)|Y)] \geq \mathrm{Var}(\mathbb{E}[\varphi(X)|Y]),$$

où la variance conditionnelle de $\varphi(X)$ sachant Y est la variable aléatoire définie p.s. par

$$\mathrm{Var}(\varphi(X)|Y) = \mathbb{E}[(\varphi(X) - \mathbb{E}[\varphi(X)|Y])^2|Y] = \mathbb{E}[\varphi(X)^2|Y] - \mathbb{E}[\varphi(X)|Y]^2.$$

3. sous-entendu : de l'espace de Hilbert $L^2(\Omega, \mathcal{F}, \mathbb{P})$.

Soit donc Y une variable auxiliaire de densité g dont on sait simuler des réalisations et telle que, pour tout y, on puisse calculer $\psi(y) = \mathbb{E}[\varphi(X)|Y = y]$. On considère l'estimateur

$$\tilde{I}_n = \frac{1}{n}\sum_{i=1}^{n} \psi(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\varphi(X)|Y_i].$$

Ses propriétés découlent de ce qui vient d'être dit.

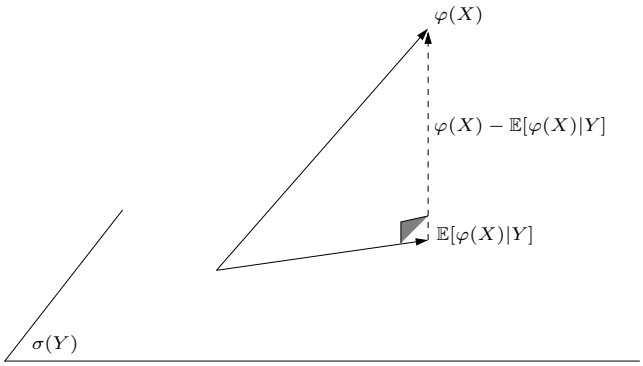

FIGURE 2.2 – Interprétation de $\mathbb{E}[\varphi(X)|Y]$ comme projeté orthogonal.

Proposition 9 (Estimation par conditionnement)
Si $\mathbb{E}|\varphi(X)| < \infty$, alors l'estimateur $\tilde{I}_n$ est sans biais et convergent, c'est-à-dire que $\mathbb{E}[\tilde{I}_n] = I$ et

$$\tilde{I}_n = \frac{1}{n}\sum_{i=1}^{n} \psi(Y_i) \xrightarrow[n\to\infty]{p.s.} \mathbb{E}[\psi(Y)] = \mathbb{E}[\mathbb{E}[\varphi(X)|Y]] = I.$$

Si de plus $\mathbb{E}[\varphi^2(X)] < \infty$, alors

$$\sqrt{n}\ (\tilde{I}_n - I) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, s^2),$$

avec

$$s^2 = \mathrm{Var}(\mathbb{E}[\varphi(X)|Y]) = \mathrm{Var}(\psi(Y)) = \mathbb{E}[\psi^2(Y)] - I^2.$$

A nouveau, la variance s^2 est estimée de façon immédiate par

$$\tilde{s}_n^2 = \frac{1}{n}\sum_{i=1}^{n} \psi^2(Y_i) - \tilde{I}_n^2,$$

et les intervalles de confiance asymptotiques en découlent.

Exemple : surface du quart de disque unité. On revient à l'exemple vu en Section 2.1.1 où $I = \pi/4$ est estimée par

$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_D(X_i, Y_i).$$

Les variables X et Y étant uniformes et indépendantes, on a

$$\mathbb{E}[\mathbf{1}_D(X,Y)|X = x] = \mathbb{P}(x^2 + Y^2 \leq 1) = \mathbb{P}(Y \leq \sqrt{1-x^2}) = \sqrt{1-x^2} = \psi(x),$$

ce qui conduit à l'estimateur

$$\tilde{I}_n = \frac{1}{n}\sum_{i=1}^{n}\sqrt{1-X_i^2}.$$

Sa variance se calcule facilement :

$$s^2 = \mathbb{E}[\psi^2(X)] - I^2 = \int_0^1 (1-x^2)dx - (\pi/4)^2 = 2/3 - (\pi/4)^2 \approx 0.05.$$

La variance de $\hat{I}_n$ étant de $\sigma^2 = \pi/4(1-\pi/4) \approx 0.17$, on gagne un facteur 2 en écart-type, c'est-à-dire que pour le même n, on a un estimateur deux fois plus précis avec deux fois moins de variables uniformes (puisqu'on ne simule plus les Y_i).

2.2.3 Stratification

Le principe est le même qu'en sondages : s'il existe une variable auxiliaire Y permettant de partitionner l'espace d'états $\mathcal{X} = \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_J$ en J strates sur chacune desquelles la variable X est "assez homogène", on a tout intérêt à tirer parti de cette information.

Formellement, supposons que l'on connaisse les probabilités $p_j = \mathbb{P}(X \in \mathcal{X}_j)$ et que l'on sache simuler selon les lois conditionnelles $\mathcal{L}(X|X \in \mathcal{X}_j)$. Par conditionnement, on peut écrire

$$I = \mathbb{E}[\varphi(X)] = \sum_{j=1}^{J}\mathbb{E}[\varphi(X)|X \in \mathcal{X}_j]\mathbb{P}(X \in \mathcal{X}_j) = \sum_{j=1}^{J} p_j \mathbb{E}[\varphi(X)|X \in \mathcal{X}_j].$$

L'idée est alors d'estimer séparément chacune des moyennes conditionnelles $\mathbb{E}[\varphi(X)|X \in \mathcal{X}_j]$ grâce à n_j simulations. Soit donc $(n_1, \ldots, n_J)$ un J-uplet tel que $n_1 + \cdots + n_J = n$ et, pour tout j, $(X_{1,j}, \ldots, X_{n_j,j})$ des réalisations i.i.d. de la loi $\mathcal{L}(X|X \in \mathcal{X}_j)$. L'estimateur par stratification a la forme

$$\tilde{I}_n = \sum_{j=1}^{J} p_j \left(\frac{1}{n_j}\sum_{i=1}^{n_j}\varphi(X_{i,j})\right)$$

et possède les propriétés suivantes.

Proposition 10 (Estimation par conditionnement)
Si $\mathbb{E}|\varphi(X)| < \infty$, alors l'estimateur $\tilde{I}_n$ est sans biais et, si tous les n_j tendent vers l'infini, il est convergent, c'est-à-dire que

$$\tilde{I}_n = \sum_{j=1}^{J} p_j \left(\frac{1}{n_j}\sum_{i=1}^{n_j}\varphi(X_{i,j})\right) \xrightarrow[(n_1,\ldots,n_J)\to\infty]{p.s.} I.$$

Si de plus $\mathbb{E}[\varphi^2(X)] < \infty$, alors sa variance vaut

$$s_n^2 = \mathrm{Var}(\tilde{I}_n) = \sum_{j=1}^{J}\frac{p_j^2}{n_j}\,\mathrm{Var}(\varphi(X)|X \in \mathcal{X}_j) = \sum_{j=1}^{J}\frac{p_j^2}{n_j}\,\sigma_j^2,$$

avec

$$\sigma_j^2 = \mathrm{Var}(\varphi(X)|X \in \mathcal{X}_j) = \mathbb{E}\left[\varphi^2(X)|X \in \mathcal{X}_j\right] - \mathbb{E}\left[\varphi(X)|X \in \mathcal{X}_j\right]^2.$$

Les estimateurs naturels de σ_j^2 et s_n^2 sont donc respectivement

$$\tilde{\sigma}_j^2 = \frac{1}{n_j}\sum_{i=1}^{n_j}\varphi^2(X_{i,j}) - \left(\frac{1}{n_j}\sum_{i=1}^{n_j}\varphi(X_{i,j})\right)^2 \implies \tilde{s}_n^2 = \sum_{j=1}^{J}\frac{p_j^2}{n_j}\,\tilde{\sigma}_j^2.$$

Si l'on convient de noter $m_j = \mathbb{E}[\varphi(X)|X \in \mathcal{X}_j]$ les moyennes par classe, la formule de décomposition en variances intra et inter-classe nous assure que

$$\mathrm{Var}(\varphi(X)) = \sum_{j=1}^{J} p_j \sigma_j^2 + \sum_{j=1}^{J} p_j (I - m_j)^2.$$

Par ailleurs, si on effectue une allocation proportionnelle en prenant $n_j = p_j \times n$ (abstraction faite des arrondis), il vient pour l'estimateur par stratification

$$s_n^2 = \frac{1}{n} \sum_{j=1}^{J} p_j \sigma_j^2 \leq \frac{\mathrm{Var}(\varphi(X))}{n},$$

la variance inter-strate ayant disparu par rapport au Monte-Carlo standard. Théoriquement, on peut faire encore mieux grâce à l'allocation dite optimale : l'optimisation sous contrainte

$$\inf_{n_1,\ldots,n_J} \sum_{j=1}^{J} \frac{p_j^2}{n_j}\, \sigma_j^2 \qquad \text{avec} \qquad n_1 + \cdots + n_J = n$$

admet en effet la solution

$$(n_1^\star, \ldots, n_J^\star) = \left(\frac{p_1\sigma_1}{\sum_{j=1}^J p_j\sigma_j}\, n, \ldots, \frac{p_J\sigma_J}{\sum_{j=1}^J p_j\sigma_j}\, n \right) \implies \sum_{j=1}^{J} \frac{p_j^2}{n_j}\, \sigma_j^2 = \frac{1}{n} \left(\sum_{j=1}^{J} p_j \sigma_j \right)^2.$$

Cette allocation est impossible a priori puisqu'on ne connaît pas les σ_j. On sait cependant les estimer par les $\tilde{\sigma}_j$ déjà croisés, donc on peut envisager une méthode en deux temps : on commence par les estimer grâce à une première simulation, tandis qu'une seconde simulation effectue une allocation quasi-optimale avec

$$(\tilde{n}_1^\star, \ldots, \tilde{n}_J^\star) = \left(\frac{p_1\tilde{\sigma}_1}{\sum_{j=1}^J p_j\tilde{\sigma}_j}\, n \ldots, \frac{p_J\tilde{\sigma}_J}{\sum_{j=1}^J p_j\tilde{\sigma}_j}\, n \right).$$

Indépendamment de l'allocation choisie, les termes de variance intra-classe σ_j^2 rappellent qu'on a tout intérêt à choisir des strates $\mathcal{X}_j$ sur lesquelles la variable X est aussi homogène que possible.

Remarque. Notons Y la variable aléatoire discrète valant j si $X \in \mathcal{X}_j$. L'approche par stratification est liée à la méthode de conditionnement puisque dans les deux cas l'estimation est basée sur le calcul d'espérance par conditionnement :

$$\mathbb{E}[\varphi(X)] = \mathbb{E}[\mathbb{E}[\varphi(X)|Y]] = \left\{ \begin{array}{l} \displaystyle\int \mathbb{E}[\varphi(X)|Y = y] g(y) dy \\ \displaystyle\sum_{j=1}^{J} \mathbb{E}[\varphi(X)|Y = j] p_j \end{array} \right.$$

Elles sont en fait duales : dans un cas on simule la variable auxiliaire Y et on connaît l'espérance conditionnelle sachant celle-ci, dans l'autre on connaît la loi de Y et on estime l'espérance conditionnelle sachant celle-ci.

2.2.4 Variables antithétiques

Nous ne présentons pas cette méthode dans le cadre le plus général, mais plutôt par l'intermédiaire d'une application classique. Rappelons que l'estimateur Monte-Carlo standard de $I = \mathbb{E}[\varphi(X)]$ s'écrit

$$\hat{I}_n = \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i) = \frac{1}{n} \sum_{i=1}^{n} \varphi(F^{-1}(U_i)),$$

où F^{-1}, supposée connue, désigne comme au Chapitre 1 l'inverse généralisée de la fonction de répartition F de X. Les variables U_i sont quant à elles i.i.d. selon la loi uniforme sur $[0,1]$. Si les variables X_i sont simulées par la méthode d'inversion, c'est en fait sous cette dernière forme que l'estimateur est implémenté.

Puisque les variables $(1-U_i)$ sont elles-mêmes i.i.d. selon la loi uniforme, l'estimateur

$$\tilde{I}_n = \frac{1}{n}\sum_{i=1}^{n} \frac{\varphi(F^{-1}(U_i)) + \varphi(F^{-1}(1-U_i))}{2}$$

est lui aussi sans biais et convergent. Il est cependant deux fois plus coûteux si l'évaluation $\varphi(F^{-1}(u))$ n'est pas immédiate. Quid de sa variance ? Celle de $\hat{I}_n$ a déjà été vue et vaut $\mathrm{Var}(\varphi(X))/n$. Pour $\tilde{I}_n$, on a

$$\mathrm{Var}(\tilde{I}_n) = \frac{1}{n} \times \frac{\mathrm{Var}(\varphi(F^{-1}(U))) + \mathrm{Cov}(\varphi(F^{-1}(U), \varphi(F^{-1}(1-U)))}{2}.$$

Or $\mathrm{Var}(\varphi(F^{-1}(U))) = \mathrm{Var}(\varphi(X))$ et l'inégalité de Cauchy-Schwarz impose que

$$\mathrm{Cov}(\varphi(F^{-1}(U), \varphi(F^{-1}(1-U))) \leq \sqrt{\mathrm{Var}(\varphi(F^{-1}(U)))} \times \sqrt{\mathrm{Var}(\varphi(F^{-1}(1-U)))} = \mathrm{Var}(\varphi(X)),$$

si bien que l'on a toujours, en notant $\rho(X,Y) = \mathrm{Cov}(X,Y)/(\sigma(X)\sigma(Y))$ le coefficient de corrélation linéaire,

$$\frac{\mathrm{Var}(\tilde{I}_n)}{\mathrm{Var}(\hat{I}_n)} = \frac{1 + \rho(\varphi(F^{-1}(U), \varphi(F^{-1}(1-U)))}{2} \leq 1,$$

c'est-à-dire que $\tilde{I}_n$ est préférable à $\hat{I}_n$. Si de plus φ est monotone, alors l'inégalité de covariance de Tchebychev permet de montrer que

$$\mathrm{Cov}(\varphi(F^{-1}(U), \varphi(F^{-1}(1-U))) \leq 0 \implies \frac{\mathrm{Var}(\tilde{I}_n)}{\mathrm{Var}(\hat{I}_n)} \leq \frac{1}{2}.$$

Ainsi le surcoût algorithmique est au moins compensé par la réduction de variance.

Exemple jouet. Supposons que $X \sim \mathcal{U}_{[0,1]}$ et $\varphi(x) = x^2$. On obtient dans ce cas une division de la variance par 16.

2.3 Méthodes Quasi-Monte-Carlo

Revenons au problème d'intégration par rapport à la loi uniforme, autrement dit l'estimation de

$$I = \int_{[0,1]^d} \varphi(x)dx.$$

Si $d = 1$, pour une méthode Monte-Carlo classique, c'est-à-dire lorsque $(X_n)_{n\geq 1}$ est une suite de variables i.i.d. uniformes sur $[0,1]$, l'estimateur

$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} \varphi(X_i)$$

a une erreur moyenne en $\mathcal{O}(1/\sqrt{n})$, celle-ci étant mesurée par l'écart-type. Pour n fixé, puisque cette moyenne est mesurée par rapport à l'ensemble des séquences possibles $\{X_1, \ldots, X_n\}$ i.i.d. uniformes sur $[0,1]$, cela signifie qu'il en existe pour lesquelles l'erreur est plus petite que $1/\sqrt{n}$.

Par exemple, à n fixé et si φ est de classe $\mathcal{C}^1$, alors pour la séquence $\{1/n, \dots, (n-1)/n, 1\}$, la méthode des rectangles donne

$$|R_n - I| = \left| \frac{1}{n} \sum_{i=1}^{n} \varphi(i/n) - I \right| \leq \frac{\|\varphi'\|_\infty}{2n},$$

donc une vitesse bien meilleure. Néanmoins, avec cette méthode, l'estimateur à $(n+1)$ points requiert le calcul des $\varphi(i/(n+1))$, c'est-à-dire qu'on ne peut se servir ni de la valeur de R_n ni des valeurs $\varphi(i/n)$ déjà calculées. A contrario, pour une méthode Monte-Carlo, la relation entre $\hat{I}_n$ et $\hat{I}_{n+1}$ est élémentaire puisque

$$\hat{I}_{n+1} = \frac{n}{n+1}\hat{I}_n + \frac{1}{n+1}\varphi(X_{n+1}).$$

L'idée des méthodes Quasi-Monte-Carlo est de trouver un compromis entre ces deux techniques. Commençons par introduire la notion de discrépance, ou plutôt une notion de discrépance.

Définition 3 (Discrépance à l'origine)
Soit $(\xi_n)_{n\geq 1}$ une suite de $[0,1]^d$. La discrépance (à l'origine) de $(\xi_n)_{n\geq 1}$ est la suite $(D_n^\star(\xi))_{n\geq 1}$ définie par

$$D_n^\star(\xi) = \sup_{B\in\mathcal{R}^\star} |\lambda_n(B) - \lambda(B)| = \sup_{B\in\mathcal{R}^\star} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_B(\xi_i) - \lambda(B) \right|,$$

où $\mathcal{R}^\star = \{B = [0,u_1]\times\cdots\times[0,u_d], 0 \leq u_j \leq 1\ \forall j\}$ est l'ensemble des pavés contenus dans $[0,1]^d$ avec un sommet en l'origine, $\lambda(B) = u_1\times\cdots\times u_d$ est la mesure de Lebesgue d'un tel pavé tandis que $\lambda_n(B)$ est sa mesure empirique pour la suite $(\xi_n)_{n\geq 1}$, i.e. la proportion des n premiers points de cette suite appartenant à B.

En dimension $d = 1$, on a ainsi

$$D_n^\star(\xi) = \sup_{0\leq u\leq 1} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[0,u]}(\xi_i) - u \right|.$$

Autrement dit, la discrépance à l'origine mesure la distance en norme sup entre la fonction de répartition empirique et celle de la loi uniforme.

Exemples :

1. Suites de van der Corput : en dimension $d = 1$, on se fixe un nombre entier, par exemple 2, et on part de la suite des entiers naturels que l'on traduit en base 2 :

   $$1 = (1), 2 = (10), 3 = (11), 4 = (100), 5 = (101), 6 = (110), 7 = (111), 8 = (1000), \dots$$

   puis on applique la transformation "miroir" suivante :

   $$n = \sum_{b=0}^{B} x_b 2^b \implies \xi_n = \sum_{b=0}^{B} \frac{x_b}{2^{b+1}},$$

   ce qui donne, en décomposition binaire :

   $$\xi_1 = (0.1), \xi_2 = (0.01), \xi_3 = (0.11), \xi_4 = (0.001), \xi_5 = (0.101), \xi_6 = (0.011), \xi_7 = (0.111), \dots$$

   d'où, retraduit en décomposition décimale, la suite de van der Corput de base 2 :

   $$(\xi_n)_{n\geq 1} = \left(\frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{5}{8}, \frac{3}{8}, \frac{7}{8}, \frac{1}{16}, \frac{9}{16}, \frac{5}{16}, \frac{13}{16}, \frac{3}{16}, \frac{11}{16}, \frac{7}{16}, \frac{15}{16}, \frac{1}{32}, \dots \right).$$

   On peut montrer que $D_n^\star(\xi) = \mathcal{O}(\log n/n)$.

2. Suites de Halton : elles correspondent à la généralisation des suites de van der Corput en dimension d supérieure à 1. On considère d nombres premiers entre eux $b_1, \dots, b_d$ et on construit les d suites de van der Corput $\xi^{(1)}, \dots, \xi^{(d)}$ de bases respectives $b_1, \dots, b_d$. La suite de Halton associée est alors

$$\xi = (\xi_n)_{n\geq 1} = \left(\left(\xi_n^{(1)}, \dots, \xi_n^{(d)} \right) \right)_{n\geq 1}.$$

Par exemple, en dimension 2, en prenant $b_1 = 2$ et $b_2 = 3$, les premiers termes de la suite de Halton associée sont (voir aussi Figure 2.3)

$$\xi = \left(\left(\frac{1}{2}, \frac{1}{3}\right), \left(\frac{1}{4}, \frac{2}{3}\right), \left(\frac{3}{4}, \frac{1}{9}\right), \left(\frac{1}{8}, \frac{4}{9}\right), \left(\frac{5}{8}, \frac{7}{9}\right), \left(\frac{3}{8}, \frac{2}{9}\right), \left(\frac{7}{8}, \frac{5}{9}\right), \left(\frac{1}{16}, \frac{8}{9}\right), \dots \right)$$

Pour une suite de Halton en dimension d, on peut montrer que $D_n^\star(\xi) = \mathcal{O}((\log n)^d/n)$.

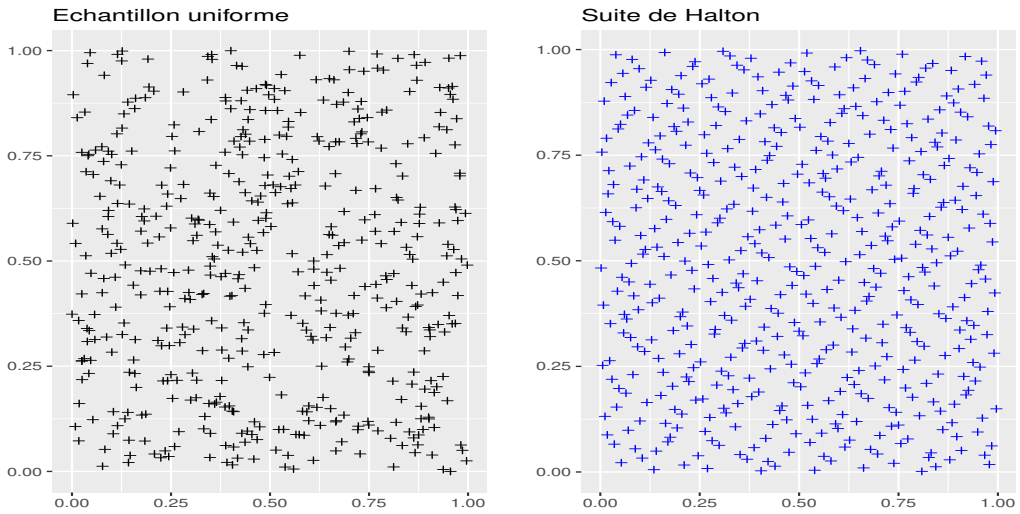

FIGURE 2.3 – Echantillons de $n = 500$ points d'une suite uniforme et d'une suite de Halton.

Avant de voir en quoi la notion de discrépance permet de contrôler l'erreur d'estimation, il faut introduire une notion de variation de la fonction à intégrer. Pour simplifier les choses, nous nous focaliserons sur les fonctions φ suffisamment régulières.

Définition 4 (Variation de Hardy-Krause)
La variation au sens de Hardy-Krause d'une fonction $\varphi : [0,1]^d \to \mathbb{R}$ de classe $\mathcal{C}^d$ est

$$V(\varphi) = \sum_{j=1}^{d} \sum_{i_1 < \dots < i_j} \int_{[0,1]^j} \left| \frac{\partial^j \varphi}{\partial x_{i_1} \dots \partial x_{i_j}} (x(i_1, \dots, i_j)) \right| dx_{i_1} \dots dx_{i_j},$$

avec $x(i_1, \dots, i_j)$ le point de $\mathbb{R}^d$ dont toutes les coordonnées valent 1 sauf celles de rangs $(i_1, \dots, i_j)$ qui valent respectivement $(x_{i_1}, \dots, x_{i_j})$.

Ainsi, lorsque $d = 1$, on a tout simplement

$$V(\varphi) = \int_0^1 |\varphi'(x)| dx = \|\varphi'\|_1.$$

Pour $d = 2$, ça se complique un peu :

$$V(\varphi) = \int_0^1 \left|\frac{\partial \varphi}{\partial x_1}(x_1, 1)\right| dx_1 + \int_0^1 \left|\frac{\partial \varphi}{\partial x_2}(1, x_2)\right| dx_2 + \iint_{[0,1]^2} \left|\frac{\partial^2 \varphi}{\partial x_1 \partial x_2}(x_1, x_2)\right| dx_1 dx_2.$$

Il est clair qu'en dimension supérieure, estimer la variation de Hardy-Krause devient vite inextricable : somme de $(2^d - 1)$ termes, dérivées partielles... Néanmoins, cette notion intervient de façon cruciale pour majorer la qualité d'un estimateur basé sur une séquence $(\xi_n)_{n\geq 1}$.

Théorème 3 (Inégalité de Koksma-Hlawka)
Pour toute fonction $\varphi : [0,1]^d \to \mathbb{R}$ et toute suite $(\xi_n)_{n\geq 1}$ de $[0,1]^d$, on a

$$\left|\frac{1}{n}\sum_{i=1}^{n} \varphi(\xi_i) - \int_{[0,1]^d} \varphi(x)dx\right| \leq V(\varphi) \times D_n^\star(\xi),$$

où $V(\varphi)$ est la variation de φ au sens de Hardy-Krause et $D_n^\star(\xi)$ la discrépance de ξ à l'ordre n.

L'intérêt de ce résultat est de séparer l'erreur d'estimation en deux termes : l'un faisant intervenir la régularité de φ, l'autre les propriétés de $(\xi_n)_{n\geq 1}$ en terme de discrépance. Puisqu'a priori on n'a aucune latitude sur φ, l'idée est de prendre $(\xi_n)_{n\geq 1}$ de discrépance minimale.

En dimension 1, il est prouvé qu'on ne peut faire mieux qu'une discrépance en $\mathcal{O}(\log n/n)$. En dimension supérieure, ce problème est encore ouvert à l'heure actuelle : on sait prouver que la discrépance ne peut être plus petite que $\mathcal{O}((\log n)^{d/2}/n)$, mais la conjecture est que la meilleure borne possible est en fait en $\mathcal{O}((\log n)^d/n)$. Ceci explique la définition suivante.

Définition 5 (Suites à discrépance faible et méthodes Quasi-Monte-Carlo)
Une suite $(\xi_n)_{n\geq 1}$ de $[0,1]^d$ vérifiant $D_n^\star(\xi) = \mathcal{O}((\log n)^d/n)$ est dite à discrépance faible et une méthode d'approximation basée sur ce type de suite est dite Quasi-Monte-Carlo.

Exemples. Outre les suites de Halton détaillées ci-dessus, on peut citer les suites de Faure, de Sobol et de Niederreiter comme exemples de suites à discrépance faible.

Si l'on revient au problème d'intégration

$$I = \int_{[0,1]^d} \varphi(x)dx,$$

et que l'on se donne une suite $(\xi_n)_{n\geq 1}$ à discrépance faible, par exemple une suite de Halton basée sur les d premiers nombres premiers, alors l'estimateur

$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} \varphi(\xi_i)$$

est un estimateur Quasi-Monte-Carlo de l'intégrale I, dont l'erreur (déterministe) est donc en $\mathcal{O}((\log n)^d/n)$. Rappelons qu'une méthode Monte-Carlo classique présente une erreur (moyenne) en $\mathcal{O}(1/\sqrt{n})$ et qu'une méthode de de type quadrature a une erreur (déterministe) en $\mathcal{O}(n^{-s/d})$. Dès lors, les méthodes Quasi-Monte-Carlo sont typiquement compétitives en dimension "intermédiaire" et pour des fonctions suffisamment régulières. C'est du reste leur cadre d'application en mathématiques financières.

Reférences. Sur toute cette section, on trouvera compléments et exemples en Chapitre 7 de [34] ainsi qu'en Chapitre 3 de [27].

2.4 Exercices

Exercice 2.1 (Estimation de π)
On revient sur la méthode vue en cours de simulation uniforme dans le quart de disque unité.

1. En déduire un estimateur $\hat{\pi}_n$ de π, ainsi qu'un intervalle de confiance asymptotique à 95%.
2. Sur un même graphique, représenter $\hat{\pi}_n$ et les intervalles de confiance en fonction de n pour n allant de 1 à 1000. Ajouter à ce graphique la droite horizontale $y = \pi$.
3. Comment choisir n pour obtenir une précision au centième sur π (avec 95% de chances) ?
4. Retrouver ce résultat par simulation.

Exercice 2.2 (Estimation de la constante d'Euler)
On considère sur $\mathbb{R}$ la fonction de répartition $F(x) = \exp(-\exp(-x))$. Si X a pour fonction de répartition F, on peut montrer que $\mathbb{E}[X] = \gamma \approx 0.577$, constante d'Euler.

1. Par la méthode d'inversion, proposer une méthode pour simuler une variable X de fonction de répartition F.
2. Implémenter la méthode précédente pour simuler un échantillon $X_1, \ldots, X_n$ de taille $n = 10^4$. En déduire un estimateur $\hat{\gamma}_n$ de γ ainsi qu'un intervalle de confiance asymptotique à 95%.
3. Sur un même graphique, pour n allant de 1 à 1000, représenter $\hat{\gamma}_n$ et les intervalles de confiance asymptotiques à 95% en fonction de n. Ajouter à ce graphique le droite horizontale $y = \gamma$ en rouge.
4. Si X a pour fonction de répartition F, quelle est sa densité f ? Soit $g(x) = \frac{1}{2}\exp(-|x|)$ la densité d'une variable de Laplace : représenter la fonction $x \mapsto f(x)/g(x)$ pour x variant de -10 à 10. Au vu de ce graphique, que vaut $m = \sup_{x\in\mathbb{R}} f(x)/g(x)$? Le démontrer.
5. Rappeler comment simuler Y selon une loi de Laplace. A partir de la question précédente, implémenter une méthode de rejet pour simuler X de densité f.
6. Des deux méthodes précédentes (inversion et rejet), laquelle choisiriez-vous pour simuler X ?

Exercice 2.3 (Bayes in memoriam (1761))
Une boule de billard est lancée au hasard uniforme sur une ligne de longueur 1, sa position étant notée θ. Ceci fait, une seconde boule est lancée de la même façon N fois de suite sur cette ligne et x est le nombre de fois où elle arrive à gauche de la première.

1. Avec les notations du cours, préciser la loi a priori $\pi(\theta)$, la vraisemblance $f(x|\theta)$ et la loi a posteriori $\pi(\theta|x)$. En déduire, en fonction de x et N, la moyenne a posteriori de θ (ou espérance de θ sachant x). Idem pour la variance a posteriori.
2. Pour $N = 10$ et $x = 7$, retrouver ces résultats par simulation.
3. Dans le cas général (i.e. on ne suppose plus $N = 10$ et $x = 7$), comparer la variance a posteriori de θ à sa variance a priori.
4. La moyenne a posteriori de θ correspond-elle à l'estimateur au maximum de vraisemblance ?
5. Que dire du mode a posteriori ?

Exercice 2.4 (Modèle bayésien gaussien)
Les réels τ et σ étant fixés non nuls, on suppose que θ suit une loi normale $\mathcal{N}(0, \tau^2)$ et que, sachant θ, X suit une loi normale $\mathcal{N}(\theta, \sigma^2)$.

1. Montrer que la densité a posteriori $\pi(\theta|x)$ est celle d'une loi normale dont on précisera moyenne et variance en fonction de x, τ et σ.

2. Comparer la variance a posteriori à la variance a priori.
3. La moyenne a posteriori de θ correspond-elle à l'estimateur au maximum de vraisemblance ?
4. Que dire du mode a posteriori ?

Exercice 2.5 (Estimation d'événement rare)
On veut retrouver par simulation la valeur de $p = \mathbb{P}(X \geq 6)$ avec $X \sim \mathcal{N}(0,1)$.

1. Déterminer p grâce à la fonction `pnorm` de R.
2. Rappeler l'estimateur Monte-Carlo standard basé sur n variables gaussiennes simulées via la fonction `rnorm`. Estimer p avec n le plus grand possible.
3. Si T suit une loi exponentielle de paramètre 1, donner la densité de l'exponentielle décalée $Y = 6 + T$. En déduire un estimateur d'échantillonnage préférentiel pour p, le représenter en fonction de n, ainsi que les intervalles de confiance à 95%.

Exercice 2.6 (Here comes trouble)
On veut retrouver par simulation la valeur de $p = \mathbb{P}(X \geq 10)$ avec X qui suit une loi de Pareto de paramètres $(1,3)$, c'est-à-dire de densité $f(x) = 3x^{-4}\mathbf{1}_{x\geq 1}$.

1. Déterminer la valeur de p.
2. Illustrer comme en exercice précédent la convergence de l'estimateur Monte-Carlo standard.
3. Utiliser une loi exponentielle translatée pour estimer p, représenter la convergence et expliquer ce qui se passe.

Exercice 2.7 (Somme d'exponentielles et conditionnement)
Soit X et Y deux variables indépendantes suivant respectivement des lois exponentielles de paramètres 1 et 2. On veut estimer $p = \mathbb{P}(X + Y > 5)$.

1. Pour un Monte-Carlo classique, représenter l'estimateur de la variance en fonction de n.
2. Soit $S = X + Y$ la somme des deux variables. Préciser $\mathbb{E}[\mathbf{1}_{S>5}|Y = y]$ et en déduire une méthode par conditionnement pour estimer p. Représenter le gain en variance par rapport à la question précédente.
3. Déterminer la densité de S et retrouver p.

Exercice 2.8 (Décembre 2016)
On suppose que $Y \sim \mathcal{N}(1,1)$ et que, sachant $Y = y$, X suit une loi normale de moyenne y et de variance 4. On veut estimer $I = \mathbb{P}(X > 1)$.

1. En écrivant X comme la somme de deux variables gaussiennes indépendantes, montrer que $I = 1/2$.
2. Proposer un estimateur Monte-Carlo $\hat{I}_n$ de I ainsi qu'un intervalle de confiance asymptotique à 95%. Sur un même graphique, pour n allant de 1 à 1000, représenter $\hat{I}_n$ et les intervalles de confiance asymptotiques à 95% en fonction de n. Ajouter à ce graphique la droite horizontale $y = 1/2$ en rouge.
3. Proposer un estimateur $\tilde{I}_n$ de I par la méthode de conditionnement (rappel : si $Y \sim \mathcal{N}(m, s^2)$, alors `pnorm(q,mean=m,sd=s)` correspond à $\mathbb{P}(Y \leq q)$). Sans faire de calculs, est-il meilleur que $\hat{I}_n$?

Exercice 2.9 (Vecteur gaussien et exponential tilting)
Dans cet exercice, on pourra utiliser le package `mvtnorm`. On veut estimer $p = \mathbb{P}((X,Y) \in \mathcal{R})$ où (X,Y) est un vecteur gaussien centré de matrice de covariance

$$\Gamma = \begin{bmatrix} 4 & -1 \\ -1 & 4 \end{bmatrix},$$

et $\mathcal{R}_a = \{(x,y),\ x \geq a,\ y \geq a\}$, avec successivement $a = 1$, $a = 3$, $a = 10$.

1. Pour chaque a, estimer p par Monte-Carlo classique.
2. Déterminer le point (x_0, y_0) de $\mathcal{R}_a$ où la densité de la loi normale $\mathcal{N}(0, \Gamma)$ est maximale.
3. Proposer une méthode d'Importance Sampling basée sur la densité $\mathcal{N}((x_0, y_0), \Gamma)$ et comparer aux résultats de la première question.
4. Reprendre la question précédente avec des densités instrumentales de la forme $\mathcal{N}((x_0, y_0), \delta\Gamma)$ pour différentes valeurs de δ.

Exercice 2.10 (Stratification et supercanonical rate)
On étudie ici une méthode d'accélération de la convergence se basant sur un principe de stratification. On commence par l'illustrer sur un exemple jouet, à savoir l'estimation de

$$I = \mathbb{E}[\cos X] = \int_0^1 \cos x dx. \tag{2.5}$$

1. Rappeler l'estimateur Monte-Carlo classique $\hat{I}_n$ de I.
2. Grâce à une représentation log-log, c'est-à-dire avec $\log n$ en abscisse et le log de l'écart-type empirique en ordonnée, retrouver le fait que la convergence est en $1/\sqrt{n}$.
3. L'entier n étant fixé, on considère les n strates $\mathcal{X}_j = [x_{j-1}, x_j] = [(j-1)/n, j/n]$ et on simule un seul point U_j par strate. Avec les notations du cours, préciser les probabilités p_j ainsi que, pour tout j, la loi $\mathcal{L}(X|X \in \mathcal{X}_j)$. En déduire l'estimateur stratifié $\tilde{I}_n$ de I et déterminer sa vitesse de convergence grâce à une représentation log-log.
4. En (2.5), on généralise cosinus en une fonction φ dérivable telle que $M_1 = \|\varphi'\|_\infty < \infty$.
   (a) Via les accroissements finis, montrer que si $U_j \sim \mathcal{U}_{[x_{j-1}, x_j]}$, alors $\mathrm{Var}(\varphi(U_j)) \leq M_1^2/n^2$.
   (b) En déduire que l'écart-type de $\tilde{I}_n$ est en $n^{-3/2}$. On parle de "supercanonical rate" par référence à la vitesse canonique en $n^{-1/2}$ d'une méthode Monte-Carlo classique.

Exercice 2.11 (Preuve par couplage de l'inégalité de covariance de Tchebychev)
Soit X une variable aléatoire, φ une fonction croissante et ψ une fonction décroissante telles que $\varphi(X)$ et $\psi(X)$ soient de carré intégrable.

1. Soit X' une variable de même loi que X et indépendante de celle-ci. Montrer que

$$\mathrm{Cov}(\varphi(X) - \varphi(X'), \psi(X) - \psi(X')) = \mathbb{E}[(\varphi(X) - \varphi(X'))(\psi(X) - \psi(X')] \leq 0.$$

2. Montrer par ailleurs que : $\mathrm{Cov}(\varphi(X) - \varphi(X'), \psi(X) - \psi(X')) = 2\,\mathrm{Cov}(\varphi(X), \psi(X))$, et en déduire l'inégalité de covariance de Tchebychev, à savoir que : $\mathrm{Cov}(\varphi(X), \psi(X)) \leq 0$.
3. En déduire le principe des variables antithétiques en Monte-Carlo, à savoir que si φ est monotone et h décroissante telle que $h(X)$ ait même loi que X alors, sous réserve d'existence des variances,

$$\mathrm{Var}\left(\frac{\varphi(X) + \varphi(h(X))}{2}\right) \leq \frac{1}{2}\mathrm{Var}(\varphi(X)).$$

4. Exemple jouet : proposer une méthode de variable antithétique pour estimer $\mathbb{E}[\exp X]$ lorsque $X \sim \mathcal{N}(0, 1)$ et illustrer la réduction de variance obtenue par rapport à une méthode Monte-Carlo classique.
5. On peut généraliser l'inégalité de Tchebychev comme suit : soit $\varphi(x, y)$ croissante en chaque variable, $\psi(x, y)$ décroissante en chaque variable, et (X, Y) deux variables indépendantes, alors sous réserve d'existence des variances, on a $\mathrm{Cov}(\varphi(X, Y), \psi(X, Y)) \leq 0$. En déduire une méthode de variables antithétiques pour estimer la valeur de p de l'Exercice 2.7 et la comparer à la méthode par conditionnement.

Exercice 2.12 (Variables antithétiques)
On s'intéresse à l'estimation de l'intégrale $I = \int_0^1 e^u du$.

1. Rappeler la formule de l'estimateur Monte-Carlo standard $\hat{I}_n$. Rappeler le Théorème Central Limite auquel il obéit et calculer la variance σ^2 qu'il fait intervenir. Donner un estimateur $\hat{\sigma}_n^2$ de σ^2.
2. Illustrer la convergence de $\hat{\sigma}_n^2$ vers σ^2.
3. Donner un estimateur $\tilde{I}_n$ de I à base de variables antithétiques. Quelle est sa variance théorique s^2 ? Par rapport au Monte-Carlo standard, par combien (environ) a-t-on divisé le temps de calcul pour atteindre la même précision ?
4. Soit c une constante et $X_c = \exp(U) + c(U - 1/2)$, où $U \sim \mathcal{U}_{[0,1]}$. Quelle est la moyenne de la variable X_c ? Exprimer la variance de X_c en fonction de c et des variances et covariance de U et $\exp(U)$. En déduire la valeur $c^\star$ de c rendant cette variance minimale et préciser $\mathrm{Var}(X_{c^\star})$. Comparer à s^2.

Exercice 2.13 (Marche aléatoire avec dérive et échantillonnage préférentiel)
Soit $(X_n)_{n\geq 1}$ une suite de variables gaussiennes indépendantes $\mathcal{N}(-m, 1)$ avec $m > 0$, et $S_n = X_1 + \cdots + X_n$. Les réels $a < 0$ et $b > 0$ étant fixés, on considère le temps d'arrêt $N = \min\{n : S_n < a \text{ ou } S_n > b\}$ et on s'intéresse à la probabilité $p = \mathbb{P}(S_N > b)$ que la marche aléatoire (S_n) sorte de l'intervalle $[a, b]$ par le haut.

1. Montrer que le temps d'arrêt N est presque sûrement fini.
2. Pour $m = 1$, $a = -50$ et $b = 5$, simuler et représenter une trajectoire $(S_n)_{1\leq n\leq N}$.
3. Proposer un estimateur $\hat{p}_k$ de p basé sur k simulations successives de trajectoires. L'implémenter pour $m = 1$, $a = -50$ et $b = 5$. Combien trouvez-vous pour $\hat{p}_k$ après $k = 1000$ simulations ?
4. Plutôt que de simuler X suivant la densité f d'une loi $\mathcal{N}(-m, 1)$, on simule X' suivant la loi $\mathcal{N}(m, 1)$. On note S'_n la marche aléatoire obtenue et N le temps de sortie de l'intervalle $[a, b]$. Calculer le rapport de vraisemblance $f(s)/g(s)$, où f et g sont les densités respectives des variables aléatoires S_N et S'_N.
5. Déduire de la question précédente un nouvel estimateur $\tilde{p}_k$ de p basé sur k simulations successives, et illustrer sa convergence en fonction de k pour $m = 1$, $a = -50$ et $b = 5$.
6. Déduire de la forme de $\tilde{p}_k$ que $p \leq \exp(-2mb)$. Ceci est-il en accord avec le résultat de la question 3 ?

Exercice 2.14 (Méthodes numériques, Monte-Carlo et Quasi-Monte-Carlo)
Soient k et d deux entiers naturels non nuls. On veut estimer par plusieurs méthodes l'intégrale

$$I(k, d) = \int_{[0,1]^d} f(x)dx = \int_{[0,1]^d} \left(\prod_{j=1}^{d} \tfrac{k\pi}{2} \sin(k\pi x_j) \right) dx_1 \dots dx_d.$$

1. Calculer à la main la valeur exacte de $I(k, d)$.
2. Importer le package `cubature` et retrouver cette valeur grâce à la fonction `adaptIntegrate` pour $d = 10$ (observer la différence entre $k = 1$ et $k = 2$).
3. Implémenter l'estimateur Monte-Carlo classique $\hat{I}_n$ de $I(k, d)$ ainsi qu'un estimateur $\widehat{\mathrm{Var}}(\hat{I}_n)$ de sa variance. Calculer à la main la valeur exacte de $\mathrm{Var}(\hat{I}_n)$. Pour $k = 1$ et $d = 10$, tracer sur un même graphique, en échelle log-log, $\mathrm{Var}(\hat{I}_n)$ et $\widehat{\mathrm{Var}}(\hat{I}_n)$ en fonction de n.

4. Sur un même graphique, représenter à gauche 500 points distribués selon une loi uniforme dans le carré $[0,1] \times [0,1]$, et à droite 500 points d'une suite de Halton.
5. Toujours à l'aide d'une suite de Halton, implémenter l'estimateur Quasi-Monte-Carlo $\tilde{I}_n$ de $I(k,d)$. Sur un même graphique, pour $k=1$, représenter I, $\hat{I}_n$ et $\tilde{I}_n$ en fonction de n. Faire varier d.

Exercice 2.15 (Bayes, Cauchy et Gauss)
Dans un cadre bayésien, on considère que la loi a priori sur θ est une loi de Cauchy standard et que, sachant θ, les variables X_i sont i.i.d. selon une loi normale de moyenne θ et de variance 1.

1. Donner la formule de la loi a posteriori $\pi(\theta|\mathbf{x}) = \pi(\theta|x_1,\ldots,x_N)$ et de la moyenne a posteriori $\hat{\theta}(\mathbf{x})$.
2. Pour $\theta_0 = 3$ et $N = 10$, générer $X_1,\ldots,X_N$ selon une loi normale de moyenne θ et de variance 1 pour obtenir une réalisation $\mathbf{x} = (x_1,\ldots,x_N)$.
3. En déduire un estimateur Monte-Carlo $\hat{\theta}_n(\mathbf{x})$ de l'estimateur de Bayes (moyenne a posteriori). Comment évolue cet estimateur avec N ?
4. Dans un cadre général, supposons qu'on veuille simuler selon la loi a posteriori $\pi(\theta|\mathbf{x})$. Si on adopte une méthode de rejet avec comme loi instrumentale la loi a priori $\pi(\theta)$, montrer que la constante optimale m est liée à l'estimateur du maximum de vraisemblance $\hat{\theta}_{mv}$.
5. Si on revient au cas particulier précédent avec $\theta_0 = 3$, $N = 10$ et $\mathbf{x} = (x_1,\ldots,x_N)$, en déduire une méthode de rejet pour simuler selon la loi a posteriori. Représenter un échantillon de taille $n = 100$ selon cette loi a posteriori via un histogramme ou via la fonction `density`. Idem avec $N = 100$.

Exercice 2.16 (Estimation d'une p-value)
On considère un vecteur gaussien en dimension 2, $X = [X_1, X_2]' \sim \mathcal{N}([\theta_1,\theta_2]', I_2)$, où I_2 est la matrice identité en dimension 2. On veut tester $H_0 : \theta_1 = \theta_2 = 0$ contre $H_1 : \exists j \in \{1,2\},\ \theta_j \neq 0$. Pour ce faire, on considère la statistique de test $T(X) = \max(|X_1|, |X_2|)$.

1. On notera G la fonction de répartition de la loi $\mathcal{N}(0,1)$ et F celle de la variable aléatoire $T(X)$ sous H_0. Pour tout $t \geq 0$, exprimer $F(t)$ en fonction de $G(t)$. Soit $\alpha \in]0,1[$ le risque de première espèce. En déduire $q = q_\alpha$ tel que, sous H_0, $\mathbb{P}(T(X) > q) = \alpha$, et un test de niveau α pour décider entre H_0 et H_1.
2. Soit $x = [x_1, x_2]'$ une réalisation de X. On rappelle que la p-value associée est la probabilité que, sous H_0, $\alpha_0(x) = \mathbb{P}(T(X) \geq T(x))$. Déduire de la question précédente cette p-value en fonction de $T(x)$. Que vaut-elle pour $x = [1,2]'$?
3. Toujours pour $x = [1,2]'$, retrouver approximativement cette p-value en simulant un grand nombre de vecteurs aléatoires gaussiens i.i.d. de loi $\mathcal{N}([0,0]', I_2)$.
4. On suppose maintenant que

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}, \begin{bmatrix} 1 & 2 \\ 2 & 5 \end{bmatrix} \right),$$

et on veut toujours tester $H_0 : \theta_1 = \theta_2 = 0$ contre $H_1 : \exists j \in \{1,2\},\ \theta_j \neq 0$, avec la même statistique de test $T(X) = \max(|X_1|, |X_2|)$. La formule théorique de la p-value n'est plus aussi simple et on ne cherchera pas à l'expliciter. Par contre, la méthode Monte-Carlo se généralise sans problème. Toujours pour $x = [1,2]'$, déterminer (approximativement) cette p-value en simulant un grand nombre de vecteurs aléatoires gaussiens.

Exercice 2.17 (Test du rapport de vraisemblance)
Soit $X \sim \mathcal{B}(N, \theta)$. On veut tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$. Le test du rapport de vraisemblance consiste à calculer la statistique de test

$$T(X) = -2 \log \frac{\theta_0^X (1-\theta_0)^{N-X}}{\hat{\theta}^X (1-\hat{\theta})^{N-X}},$$

où $\hat{\theta}$ est l'estimateur du maximum de vraisemblance et avec la convention usuelle $0^0 = 1$.

1. Exprimer $\hat{\theta}$ en fonction de X et N. Dans la suite, on considère $\theta_0 = 1/2$.
2. Soit x une observation selon la loi $\mathcal{B}(N, 1/2)$. La p-value du test est définie par

   $$\alpha_0(x) = \mathbb{P}(T(X) \geq T(x)),$$

   avec $X \sim \mathcal{B}(N, 1/2)$. Comment estimer $\alpha_0(x)$ par une méthode Monte-Carlo ? Implémenter cette méthode pour $N = 10$.
3. On peut montrer que, sous H_0, $T(X)$ tend en loi vers une χ^2_1 lorsque N tend vers l'infini. Pour $N = 10^4$, retrouver le résultat de la simulation de la question précédente à partir de la fonction `pchisq`.
4. Grâce à un développement limité en θ_0 autour de $\hat{\theta}$, justifier la convergence vers une loi χ^2_1 de la question précédente.

Exercice 2.18 (Décembre 2017)
On veut estimer de différentes façons la probabilité qu'une variable de Cauchy soit plus grande que 2, c'est-à-dire

$$p = \mathbb{P}(X > 2) = \int_2^\infty \frac{1}{\pi(1+x^2)} dx = \frac{1}{2} - \frac{\arctan 2}{\pi} \approx 0.148.$$

1. Donner l'estimateur Monte-Carlo standard $\hat{p}_n$ et la variance asymptotique σ^2 de $\sqrt{n}(\hat{p}_n - p)$.
2. Pour $n = 10^3$, implémenter l'estimateur $\hat{p}_n$ et illustrer graphiquement sa convergence vers p.
3. Donner un estimateur $\hat{\sigma}^2_n$ de σ^2 et illustrer graphiquement sa convergence vers σ^2.
4. Proposer un estimateur $\tilde{p}_n$ de p par variables antithétiques. Calculer la variance asymptotique s^2 de $\sqrt{n}(\tilde{p}_n - p)$. La comparer à σ^2.
5. Implémenter un estimateur $\tilde{s}^2_n$ de s^2 et illustrer graphiquement sa convergence vers s^2.
6. Montrer que, si les U_i sont des variables i.i.d. selon une loi uniforme sur $[0, 2]$, alors

   $$\bar{p}_n = \frac{1}{2} - \frac{1}{n} \sum_{i=1}^n \frac{2}{\pi(1+U_i^2)}$$

   est un estimateur sans biais et asymptotiquement normal de p. On ne demande pas de calculer la variance asymptotique v de cet estimateur.
7. Implémenter un estimateur $\hat{v}^2_n$ de v et illustrer graphiquement sa convergence.
8. Des trois estimateurs précédents, lequel choisissez-vous ?
9. Mêmes questions pour

   $$\breve{p}_n = \tfrac{1}{n} \sum_{i=1}^n \frac{1}{2\pi(1+V_i^2)},$$

   avec les V_i i.i.d. uniformes sur $[0, 1/2]$.

Exercice 2.19 (Sampling Importance Resampling)
Soit $N \in \mathbb{N}^\star$ fixé. Dans un cadre bayésien, on considère que la loi a priori sur θ est une loi normale centrée réduite et que, sachant θ, la variable X suit une loi binomiale $\mathcal{B}(N, e^\theta/(1+e^\theta))$.

1. Donner la formule de la loi a posteriori $\pi(\theta|x)$. Exprimer en fonction de $\pi(\theta|x)$ la probabilité $\mathbb{P}(\theta > 0|X = x)$. Proposer un estimateur Monte-Carlo $\hat{p}_{n,N}$ de cette probabilité.
2. Implémenter cet estimateur avec par exemple $N = 10$, $x = 8$ et $n = 1000$.
3. On considère la procédure suivante : générer un échantillon $(\tilde{\theta}_1, \cdots, \tilde{\theta}_n)$ i.i.d. suivant la loi a priori, calculer les poids
$$w_i = \frac{\mathbb{P}(X = x|\tilde{\theta}_i)}{\sum_{j=1}^n \mathbb{P}(X = x|\tilde{\theta}_j)},$$
puis effectuer R tirages i.i.d. multinomiaux $(\theta_1^*, \cdots, \theta_R^*)$ parmi $(\tilde{\theta}_1, \cdots, \tilde{\theta}_n)$ selon le vecteur de probabilités $w = [w_1, \cdots, w_n]$ (fonction **sample**). Pour $N = 10$, $x = 8$, $n = 1000$ et $R = 100$, représenter un estimateur de la densité de l'échantillon $(\theta_1^*, \cdots, \theta_R^*)$ ainsi obtenu.
4. Sur cet échantillon, calculer la proportion $\tilde{p}_{n,R}$ de θ_r^* qui sont positifs. Comparer à la probabilité obtenue en question 2.
5. Soit $\varphi : \mathbb{R} \to \mathbb{R}$ une fonction mesurable bornée, déterminer
$$\mathbb{E}[\frac{1}{R}\sum_{r=1}^R \varphi(\theta_r^*)|\tilde{\theta}_1, \cdots, \tilde{\theta}_n, x].$$
Montrer que
$$\mathbb{E}\left[\frac{1}{R}\sum_{r=1}^R \varphi(\theta_r^*)\middle|\tilde{\theta}_1, \cdots, \tilde{\theta}_n, x\right] \xrightarrow[n\to\infty]{p.s.} \mathbb{E}\left[\varphi(\theta)|x\right].$$

2.5 Corrigés

Voir la page du cours.

Chapitre 3

Monte-Carlo par Chaînes de Markov

Introduction

Le premier chapitre exposait quelques méthodes pour simuler des variables aléatoires $X_1, X_2, \ldots$ i.i.d. selon une loi f donnée. Ceci étant supposé possible, le deuxième chapitre en donnait une application classique pour l'estimation de

$$I = \mathbb{E}[\varphi(X)] = \int \varphi(x) f(x) dx, \tag{3.1}$$

en premier lieu par l'estimateur Monte-Carlo standard

$$\hat{I}_n = \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i). \tag{3.2}$$

Ceci étant, la convergence de $\hat{I}_n$ vers I est encore valable si les variables X_n ne sont plus i.i.d., mais constituent par exemple une chaîne de Markov de loi stationnaire f. On parle alors de méthodes Monte-Carlo par chaînes de Markov (MCMC), dont l'algorithme de Metropolis-Hastings est l'exemple le plus connu. Afin de simplifier les choses, ce chapitre en présente les idées dans le cadre d'un espace d'états fini, mais tout est transposable dans un cadre plus général.

3.1 Rappels sur les chaînes de Markov

Soit (X_n) une suite de variables aléatoires à valeurs dans un ensemble E supposé **fini** et qui sera parfois noté par commodité $E = \{1, 2, \ldots, d\}$. E est appelé l'espace d'états. On dit que (X_n) est une chaîne de Markov homogène si pour tout $n \geq 1$ et toute suite $(x_0, x_1, \ldots, x_{n-1}, x, y)$ de E telle que $\mathbb{P}(X_0 = x_0, \ldots, X_{n-1} = x_{n-1}, X_n = x) > 0$, on a l'égalité suivante :

$$\mathbb{P}(X_{n+1} = y | X_0 = x_0, \ldots, X_{n-1} = x_{n-1}, X_n = x) = \mathbb{P}(X_{n+1} = y | X_n = x) = \mathbb{P}(X_1 = y | X_0 = x).$$

Autrement dit, sachant le présent, le futur est indépendant du passé. Ou encore : l'état présent étant connu, toute information sur le passé est inutile pour prévoir l'état futur. On appelle alors probabilité de transition de l'état x vers l'état y la quantité

$$P(x, y) = \mathbb{P}(X_1 = y | X_0 = x),$$

et matrice de transition de la chaîne la matrice $P = [P(x, y)]_{1 \leq x, y \leq d}$ de taille $d \times d$. Cette matrice vérifie les propriétés suivantes :

— Encadrement des coefficients :

$$\forall (x, y) \in E^2, \ 0 \leq P(x, y) \leq 1.$$

— Somme par ligne : pour tout $x \in E$, on a

$$\sum_{y \in E} P(x, y) = 1.$$

Autrement dit, 1 est valeur propre de P, le vecteur $\mathbf{1} = [1, \ldots, 1]'$ étant un vecteur propre associé.

— Spectre : toute valeur propre de P est, en module, inférieure ou égale à 1. En effet, soit (λ, v) un couple propre. Il existe $x_0 \in E$ tel que $\|v\|_\infty = |v(x_0)| > 0$, or $Pv = \lambda v$ donne en particulier

$$|\lambda v(x_0)| = \left| \sum_{y \in E} P(x_0, y) v(y) \right| \leq \sum_{y \in E} P(x_0, y) |v(y)| \leq \left(\sum_{y \in E} P(x_0, y) \right) |v(x_0)| = |v(x_0)|,$$

donc $|\lambda| \leq 1$.

Le sous-espace propre associé à la valeur propre 1 n'est pas nécessairement de dimension égale à 1. Pour preuve l'exemple trivial de la matrice identité : noter que cet exemple correspondrait à une chaîne qui ne change jamais d'état, il ne présente donc pas un grand intérêt.

Par ailleurs, la spécification de la loi initiale, c'est-à-dire des $\mathbb{P}(X_0 = x)$ pour tout $x \in E$, et des probabilités de transition $P(x, y)$ permet d'écrire très simplement la loi jointe du vecteur aléatoire $(X_0, \ldots, X_n)$, puisque :

$$\begin{aligned} \mathbb{P}(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) &= \mathbb{P}(X_0 = x_0)\mathbb{P}(X_1 = x_1 | X_0 = x_0) ... \mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1}) \\ &= \mathbb{P}(X_0 = x_0) P(x_0, x_1) \ldots P(x_{n-1}, x_n). \end{aligned}$$

A toute chaîne de Markov peut être associé un graphe de transition de la façon suivante : les sommets du graphe sont les états de E et il existe un arc, étiqueté $P(x, y)$, de x vers y si $P(x, y) > 0$. Cette construction est commode lorsque E n'est pas trop grand ou lorsque la matrice P est très creuse, autrement dit lorsque d'un état on ne peut transiter que vers un petit nombre d'états.

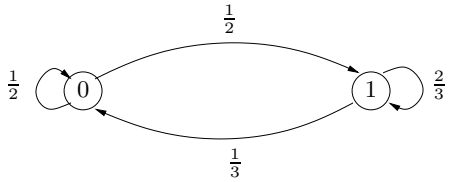

FIGURE 3.1 – Graphe de transition de la ligne téléphonique.

Exemple : la ligne téléphonique. On considère une ligne de téléphone. L'état X_n de cette ligne à l'étape n est 0 si elle est libre et 1 si elle occupée. Entre deux instants successifs, il y a une probabilité 1/2 pour qu'un appel arrive. Si la ligne est occupée et qu'un appel arrive, cet appel est perdu. La probabilité pour que la ligne se libère entre l'instant n et l'instant $(n + 1)$ est 1/3. Le graphe de transition de cette chaîne de Markov est donné figure 3.1. La matrice de transition est la suivante :

$$P = \begin{bmatrix} 1/2 & 1/2 \\ 1/3 & 2/3 \end{bmatrix}.$$

Les probabilités de transition en n étapes sont en fait complètement déterminées par les probabilités de transition en un coup, c'est-à-dire par la matrice de transition. Ceci est explicité par les équations de Chapman-Kolmogorov, que nous allons voir maintenant.

Notation. La probabilité d'aller de l'état x à l'état y en n coups est notée :

$$P(x,y)^{(n)} = \mathbb{P}(X_n = j | X_0 = i),$$

et la matrice de transition en n coups est notée :

$$P^{(n)} = \left[P(x,y)^{(n)} \right]_{(x,y)\in E^2}.$$

On adopte aussi la convention $P^{(0)} = I_d$, matrice identité de taille $|E|$.

Proposition 11 (Equations de Chapman-Kolmogorov)
Pour tout $n \geq 0$, la matrice de transition en n coups est la puissance n-ème de la matrice de transition de la chaîne, c'est-à-dire :

$$P^{(n)} = P^n.$$

Remarque. On en déduit que pour tout couple d'entiers naturels (n_1, n_2) :

$$P^{(n_1+n_2)} = P^{n_1+n_2} = P^{n_1} \times P^{n_2} = P^{(n_1)} \times P^{(n_2)}.$$

C'est plutôt cette équation qu'on appelle relation de Chapman-Kolmogorov. Ce qu'on traduit comme suit : aller de x à y en $(n_1 + n_2)$ pas, c'est d'abord aller de x à un certain x' en n_1 pas, puis de x' à y en n_2 pas.

Notation. Tout comme les transitions de la chaîne, la position initiale X_0 peut être aléatoire. On convient de noter la loi de X_0 comme un vecteur **ligne** de taille $|E| = d$:

$$\mu_0 = [\mu_0(1), \ldots, \mu_0(d)] = [\mathbb{P}(X_0 = 1), \ldots, \mathbb{P}(X_0 = d)].$$

De même, on notera en vecteur ligne la loi de X_n :

$$\mu_n = [\mathbb{P}(X_n = 1), \ldots, \mathbb{P}(X_n = d)].$$

Corollaire 1 (Loi marginale de la chaîne)
Soit (X_n) une chaîne de Markov de loi initiale μ_0 et de matrice de transition P, alors pour tout entier naturel n, la loi de X_n est :

$$\mu_n = \mu_0 P^n.$$

Pour une suite de variables aléatoires (X_n) à valeurs dans l'ensemble fini E, la convergence en loi correspond simplement à la convergence du vecteur ligne μ_n de taille d, c'est-à-dire à la convergence de chacune de ses d composantes. Puisque $\mu_n = \mu_0 P^n$, une condition suffisante pour la convergence en loi de (X_n) est donc la convergence de la suite (P^n) des puissances de la matrice P.

Tant qu'à faire, on aimerait avoir convergence de la loi de (X_n) vers une loi indépendante de la loi initiale μ_0, phénomène connu sous le nom d'oubli de la condition initiale. Mentionnons deux situations pathologiques :

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad \text{et} \qquad Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I_2. \tag{3.3}$$

Dans le premier cas, on a $P^{2n} = I_2$ et $P^{2n+1} = P$, donc (P^n) ne converge pas et, a fortiori, $\mu_n = \mu_0 P^n$ non plus sauf dans le cas très particulier $\mu_0 = [1/2, 1/2]$ d'équidistribution initiale. On a ici un problème de périodicité.

Dans le second cas, on a $\mu_n = \mu_0 Q^n = \mu_0$ trivialement convergente, mais on n'oublie pas la loi initiale pour autant. C'est un problème de communication entre états auquel on a cette fois affaire.

Afin d'éviter ces désagréments, nous nous focaliserons sur les chaînes irréductibles et apériodiques.

Une chaîne est dite **irréductible** si tous les états communiquent entre eux, autrement dit : pour tout couple de sommets du graphe de transition, il existe un chemin allant de l'un à l'autre en suivant le sens des flèches :

$$\forall (x,y),\ \exists n = n(x,y), \qquad P^n(x,y) > 0.$$

On peut alors montrer que 1 est valeur propre simple (ce qui n'était pas le cas pour $P = I_2$).

Remarque : la valeur propre 1 peut être simple sans que la chaîne soit irréductible, comme le montre la matrice

$$P = \begin{bmatrix} 1 & 0 \\ 1/2 & 1/2 \end{bmatrix}.$$

On parle dans ce cas de chaîne indécomposable (des états transitoires et une seule classe de récurrence).

Une chaîne est dite **apériodique** si

$$\forall x \in E, \qquad d(x) = \text{pgcd}\{n \geq 1,\ P^n(x,x) > 0\} = 1.$$

La quantité $d(x) \in \mathbb{N} \cup \{\infty\}$ est appelée période de l'état x. On peut montrer que, lorsqu'une chaîne est irréductible, tous les états ont même période. Par conséquent, pour une chaîne irréductible, une condition **suffisante** d'apériodicité est qu'il existe un état sur lequel elle puisse boucler, c'est-à-dire un indice x tel que $P(x,x) > 0$. Pour une chaîne irréductible, un critère d'apériodicité est le suivant :

$$\forall (x,y) \in E^2,\ \exists n_0 = n_0(x,y),\ \forall n \geq n_0,\ P^n(x,y) > 0. \tag{3.4}$$

Si une chaîne est apériodique, on peut montrer que 1 est la seule valeur propre de module 1.

Bilan : pour une chaîne apériodique et irréductible, 1 est valeur propre simple et c'est la seule de module 1.

Remarque : En (3.3), la chaîne de matrice de transition P est irréductible mais pas apériodique (période 2), tandis que celle correspondant à Q est apériodique, mais pas irréductible.

Notations :

— Tandis qu'une loi π sur E est un vecteur ligne, une fonction $\varphi : E \to \mathbb{R}$ est un vecteur colonne. La quantité $\pi\varphi$ n'est alors rien d'autre que la moyenne de la variable aléatoire $\varphi(X)$ lorsque X a pour loi π :

$$\pi\varphi = \sum_{x \in E} \pi(x)\varphi(x) = \mathbb{E}[\varphi(X)].$$

— Nous avons dit que si une chaîne est irréductible et apériodique, 1 est valeur propre simple de P et c'est la seule de module 1. Notons $1 > |\lambda_2| \geq |\lambda_3| \geq \cdots$ les valeurs propres **distinctes** rangées par ordre décroissant de leur module. Pour les valeurs propres distinctes de module $|\lambda_2|$, chacune a un degré, dit algébrique, dans le polynôme minimal de P : ce degré vaut 1 si le sous-espace propre associé à cette valeur propre est de dimension égale à la multiplicité de cette racine dans le polynôme caractéristique de P, sinon il correspond à la taille du bloc de Jordan associé. On note alors r le maximum des degrés algébriques associés aux valeurs propres distinctes de module $|\lambda_2|$. Par exemple, si P est diagonalisable, alors clairement $r = 1$.

— La distance en variation totale entre deux lois de probabilités μ et ν sur E est :

$$\|\nu - \mu\|_{vt} = \frac{1}{2}\sum_{x\in E} |\nu(x) - \mu(x)|.$$

Autrement dit, c'est la moitié de la distance L^1.

Théorème 4 (Convergence des chaînes irréductibles et apériodiques)
Si (X_n) est une chaîne de Markov irréductible de matrice de transition P sur E, il existe une ***unique*** *mesure de probabilité π invariante pour cette chaîne, c'est-à-dire telle que $\pi P = \pi$. Cette mesure est telle que $\pi(x) > 0$ pour tout x et, pour toute fonction $\varphi : E \to \mathbb{R}$, on a alors, quelle que soit la loi de X_0,*

$$\frac{1}{n}\sum_{k=1}^{n} \varphi(X_k) \xrightarrow[n\to\infty]{p.s.} \pi\varphi,$$

et

$$\sqrt{n}\left(\frac{1}{n}\sum_{k=1}^{n} \varphi(X_k) - \pi\varphi\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2(\varphi)).$$

Si, en outre, la chaîne est apériodique, alors il y a convergence à vitesse géométrique de la loi de X_n vers π : il existe une constante C telle que, pour toute loi initiale μ_0,

$$\|\mu_n - \pi\|_{vt} \leq Cn^{r-1}|\lambda_2|^n. \tag{3.5}$$

La propriété de convergence géométrique semble donc une excellente nouvelle, mais lorsque l'espace d'états est très grand, $|\lambda_2|$ peut être très proche de 1 et C peut être très grande : ce résultat n'a alors plus aucune implication pratique. La quantité $(1 - |\lambda_2|)$ est appelée le trou spectral.

Remarques :

1. On parle indifféremment de loi invariante, ou de loi stationnaire ou encore de loi d'équilibre, pour π vérifiant $\pi P = \pi$. Supposons la chaîne irréductible, donc l'unicité de π : en notant $T_x^+ = \inf\{n > 0,\ X_n = x\}$ le premier temps de retour de la chaîne en x, on a l'expression suivante pour cette loi stationnaire :

$$\pi(x) = \frac{1}{\mathbb{E}[T_x^+ | X_0 = x]}.$$

2. On parle de théorème ergodique pour la loi des grands nombres dans le cas des chaînes de Markov. Dans la même veine que la remarque précédente, la mesure d'équilibre π peut s'interpréter comme la proportion du temps passé par une trajectoire dans chaque état. Il suffit pour s'en convaincre de prendre par exemple $\varphi = \mathbf{1}_{x_0}$, pour x_0 un état fixé quelconque, et d'appliquer le théorème ergodique [1] :

$$\frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{x_0}(X_k) = \frac{|\{k \in \{1,\dots,n\},\ X_k = x_0\}|}{n} \xrightarrow[n\to\infty]{p.s.} \pi\mathbf{1}_{x_0} = \pi(x_0).$$

3. Contrairement au cas des suites de variables i.i.d., la variance $\sigma^2(\varphi)$ qui apparaît dans le TCL pour les chaînes de Markov n'est pas triviale du tout. En particulier, ça n'est pas tout bonnement

$$\mathrm{Var}_\pi(\varphi) = \sum_{j=1}^{M} \varphi^2(j)\pi_j - \left(\sum_{j=1}^{M} \varphi(j)\pi_j\right)^2,$$

1. $|A|$ désigne le cardinal de l'ensemble A.

qui correspondrait à la variance "limite", i.e. celle de $\varphi(X)$ lorsque X a pour loi π. Il faut en effet tenir compte des dépendances entre les variables X_n de la chaîne, ce qui complique tout... En considérant X_0 la condition initiale de la chaîne, supposée de loi π, on peut montrer que cette variance asymptotique vaut, :

$$\sigma^2(\varphi) = \mathbb{E}\left[(u(X_1) - (Pu)(X_0))^2\right],$$

où u est la solution de l'équation de Poisson $(I - P)u = \varphi$, i.e.

$$\forall x \in E \qquad u(x) = \sum_{n=0}^{\infty} (P^n \varphi)(x).$$

4. Si l'on revient à l'exemple de la matrice de transition P en (3.3), qui est irréductible, l'unique loi stationnaire est $\pi = [1/2, 1/2]$. D'après le théorème, cette chaîne vérifie donc une loi des grands nombres, à savoir

$$\frac{1}{n}\sum_{k=1}^{n} \varphi(X_k) \xrightarrow[n\to\infty]{p.s.} \frac{1}{2}\left(\varphi(1) + \varphi(2)\right),$$

ainsi qu'un théorème central limite, mais certainement pas de convergence en loi. La raison est intuitivement claire : la moyennisation sur une trajectoire de la chaîne permet de tuer le phénomène de périodicité, ce qui n'est pas le cas de la convergence en loi.

Définition 6 (Réversibilité)
Soit π une mesure de probabilité et (X_n) une chaîne de Markov de matrice de transition P. On dira que la chaîne est réversible pour π si elle vérifie les équations d'équilibre détaillé, à savoir

$$\pi(x)P(x,y) = \pi(y)P(y,x) \qquad \forall (x,y) \in E \times E..$$

Une interprétation est la suivante : considérons la mesure initiale μ comme une répartition de masses sur l'espace d'états, la masse totale valant donc 1. Entre l'instant 0 et l'instant 1, une proportion $P(x,y)$ de la masse $\mu(x)$ présente en x part en y, et ce pour tout couple de points (x,y). Il est facile de voir que la répartition des masses à l'instant 1 est μP, c'est-à-dire la loi de X_1. Ainsi, de façon générale, dire que π est une mesure d'équilibre pour P (c'est-à-dire que $\pi P = \pi$), signifie que si l'on part de cette mesure à un instant donné, tout ce qui quitte l'état x est égal à tout ce qui arrive dans l'état x : la répartition des masses ne change donc pas d'un instant au suivant (d'où la dénomination "loi d'équilibre"). Dire que la chaîne est réversible pour π est plus fort : sous la loi π, ce qui part de l'état x pour aller vers l'état y est égal à ce qui part de l'état y pour aller vers l'état x.

Remarque : réversibilité du temps. Le terme de réversibilité s'explique comme suit : si $X_0 \sim \pi$, alors les équations d'équilibre détaillé permettent d'écrire pour la loi jointe

$$\begin{aligned}
\mathbb{P}(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) &= \pi(x_0)P(x_0,x_1)\ldots P(x_{n-1},x_n) \\
&= P(x_1,x_0)\ldots P(x_n,x_{n-1})\pi(x_n) \\
&= \pi(x_n)P(x_n,x_{n-1})\ldots P(x_1,x_0) \\
&= \mathbb{P}(X_0 = x_n, X_1 = x_{n-1}, \ldots, X_n = x_0) \\
&= \mathbb{P}(X_n = x_0, X_{n-1} = x_1, \ldots, X_0 = x_n).
\end{aligned}$$

Autrement dit, la chaîne retournée dans le temps a la même loi :

$$\mathcal{L}(X_0, \ldots, X_n) = \mathcal{L}(X_n, \ldots, X_0).$$

Dit prosaïquement : pour une chaîne réversible à l'équilibre, face à une suite d'états, on ne sait pas dans quel sens le temps s'écoule. A contrario, considérons la chaîne irréductible de matrice de transition

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \tag{3.6}$$

d'unique loi stationnaire π la loi uniforme. Sans même faire de calculs, il est clair que la chaîne n'est pas réversible pour π puisqu'on voit très bien dans quel sens s'écoule le temps.

Exemple : marche aléatoire sur un graphe. Supposons donné un graphe non orienté de sommets numérotés de 1 à M et, pour chaque sommet $i \in \{1, \dots, M\}$, notons d_i le nombre d'arêtes connectées à celui-ci. Considérons maintenant la chaîne de Markov suivante : partant du sommet i, on va vers l'un des d_x sommets voisins de façon équiprobable. Il est facile de vérifier que cette chaîne est réversible pour la mesure $\pi = (d_x / \sum_{x' \in E} d'_x)_{x \in E}$. Cette marche aléatoire, mais sur un graphe orienté, est l'un des ingrédients du fameux algorithme PageRank de Google : les sommets correspondent aux pages web et il y a une arête de la page x vers la page y si, sur la page x, il y a un lien vers la page y.

Lemme 1 (Réversibilité ⇒ stationnarité)
Soit (X_n) une chaîne de Markov de matrice de transition P. Si la chaîne est réversible pour la mesure π, alors π est invariante, c'est-à-dire que $\pi P = \pi$.

Preuve. Si la chaîne est réversible pour la mesure π, alors pour tout y

$$(\pi P)(y) = \sum_{x \in E} \pi(x) P(x, y) = \pi(y) \sum_{x \in E} P(y, x) = \pi(y),$$

et π est bien stationnaire.

■

Remarque : Si on suppose de plus que la chaîne est irréductible, alors par le Théorème 4, les $\pi(x)$ sont tous strictement positifs et la matrice diagonale Δ_π telle que $\Delta_\pi(x, x) = \pi(x)$ est inversible. Cette remarque et la réversibilité assurent alors que

$$\Delta_\pi P = P' \Delta_\pi \implies \left(\Delta_\pi^{1/2} P \Delta_\pi^{-1/2} \right)' = \Delta_\pi^{1/2} P \Delta_\pi^{-1/2}.$$

La matrice symétrique réelle $\Delta_\pi^{1/2} P \Delta_\pi^{-1/2}$ est donc diagonalisable, par conséquent P aussi, ce qui implique $r = 1$ dans le Théorème 4. Ce raisonnement montre aussi que, si P est réversible, son spectre est réel.

Parmi les chaînes de Markov, les chaînes réversibles constituent donc un cadre d'étude privilégié. Ce sont elles qui vont nous intéresser dans la suite.

3.2 Algorithme de Metropolis-Hastings

Nous considérons toujours que l'espace d'états E est **fini**. En section précédente, considérant une chaîne de Markov définie d'une façon ou d'une autre, nous cherchions sa loi stationnaire π. Désormais, la démarche est **inverse** : on part d'une loi π donnée sur E et on cherche à construire une chaîne de Markov (X_n) dont π est la mesure stationnaire.

3.2.1 Algorithme de Metropolis

Nous présentons en premier lieu l'algorithme proposé par Metropolis et ses co-auteurs en 1953 [25] avant de passer à sa généralisation par Hastings en 1970 [20]. On veut simuler une chaîne de Markov de loi stationnaire π, celle-ci n'étant éventuellement connue qu'à une constante de normalisation près :

$$\forall x \in E \qquad \pi(x) = C\pi_u(x) = \frac{\pi_u(x)}{\sum_{x' \in E} \pi_u(x')}, \tag{3.7}$$

où π_u signifie "π unnormalized". Ce problème est tout sauf artificiel : il est par exemple récurrent en statistique bayésienne et en physique statistique. Nous supposerons pour simplifier que $\pi(x) > 0$ pour tout $x \in E$.

Pour construire cette chaîne, il nous faut spécifier une matrice de transition P ayant π comme loi stationnaire. A l'instar des méthodes de rejet et d'échantillonnage préférentiel, l'idée est d'utiliser un mécanisme de proposition auxiliaire (ou instrumental) et de corriger le tir ensuite.

Soit donc $Q = [Q(x,y)]_{x,y \in E}$ une matrice de transition telle que :
- symétrie : pour tout couple (x,y), on a $Q(x,y) = Q(y,x)$;
- partant de tout état x, on sait simuler facilement selon la loi $Q(x,.)$.

Comme dans les méthodes précédentes, l'algorithme fonctionne alors en plusieurs étapes :

1. partant de $X_n = x$, simuler $Y = y$ selon la loi $Q(x,.)$;
2. calculer le rapport d'acceptation $r(x,y) = \pi(y)/\pi(x)$;
3. tirer une loi uniforme $U \sim \mathcal{U}_{[0,1]}$, indépendant de toutes les autres variables aléatoires, et poser $X_{n+1} = Y = y$ si $U \leq r(x,y)$, $X_{n+1} = X_n = x$ sinon.

Interprétation : Si, partant de $X_n = x$, la proposition $Y = y$ est plus vraisemblable pour π (i.e. $\pi(y) \geq \pi(x)$), alors la transition est systématiquement acceptée ; dans le cas contraire, elle n'est acceptée qu'avec une probabilité égale à $r(x,y) = \pi(y)/\pi(x)$. Dans le premier cas, il est cohérent de favoriser les transitions vers les états plus vraisemblables pour π puisque celle-ci est notre loi cible ; dans le second cas, il ne faut pas refuser systématiquement une transition vers un état moins vraisemblable, sans quoi on resterait bloqué dans les maxima locaux de π vis-à-vis du système de voisinage induit par le graphe de transition de Q.

Achtung ! Dans la construction de la chaîne (X_n), il faut tout conserver, y compris les fois où l'on reste sur place (c'est-à-dire $X_{n+1} = X_n$), sans quoi tout est faussé.

Remarque. Le point a priori anecdotique mais en fait diabolique de cet algorithme est de ne faire intervenir que les rapports $\pi(y)/\pi(x) = \pi_u(y)/\pi_u(x)$. Il ne requiert donc nullement la connaissance de la constante de normalisation C de la formule (3.7). C'est ce qui explique sa popularité en physique statistique comme en statistique bayésienne.

Proposition 12 (Réversibilité de Metropolis)
Si la matrice de transition Q est irréductible, alors la chaîne (X_n) produite par l'algorithme de Metropolis est irréductible et réversible pour π. En particulier, π est son unique mesure d'équilibre.

Concernant le pénible phénomène de périodicité, la preuve va montrer qu'elle est en fait à peu près exclue ici, si ce n'est dans le cas très particulier où π est uniforme sur E et Q elle-même périodique : autant dire qu'il faut vraiment le faire exprès. Rappelons surtout que la dynamique instrumentale définie par Q est choisie par l'utilisateur : il suffit donc de prendre Q irréductible et apériodique pour que tout se passe bien, c'est-à-dire que le Théorème 4 s'applique.

Preuve. Notons $P = [P(x, y)]$ la matrice de transition de la chaîne (X_n). On veut montrer que $\pi(x)P(x, y) = \pi(y)P(y, x)$ pour tout couple (x, y). C'est clairement vrai si $x = y$. Supposons donc $x \neq y$, alors par définition de l'algorithme,

$$P(x, y) = \mathbb{P}(X_{n+1} = y|X_n = x) = \mathbb{P}(X_{n+1} = y, Y = y|X_n = x).$$

Or, par la formule de Bayes,

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

c'est-à-dire, à nouveau par définition de l'algorithme,

$$\mathbb{P}(X_{n+1} = y, Y = y|X_n = x) = \min\left(\frac{\pi(y)}{\pi(x)}, 1\right) Q(x, y),$$

d'où

$$\pi(x)P(x, y) = \pi(x) \min\left(\frac{\pi(y)}{\pi(x)}, 1\right) Q(x, y) = \min\left(\pi(x), \pi(y)\right) Q(x, y).$$

Par symétrie des rôles joués par x et y, on en déduit aussitôt que

$$\pi(y)P(y, x) = \min\left(\pi(x), \pi(y)\right) Q(y, x).$$

Puisque $Q(x, y) = Q(y, x)$, les équations d'équilibre détaillé sont bien vérifiées :

$$\forall (x, y) \in E \times E, \qquad \pi(x)P(x, y) = \pi(y)P(y, x).$$

L'irréductibilité de la chaîne (X_n) se déduit de celle induite par Q : s'il existe un chemin de x vers y dans le graphe de transition associé à Q, alors ce chemin reste de probabilité non nulle pour le graphe de transition associé à P. Passons à l'apériodicité : supposons la chaîne (X_n) non apériodique, alors en particulier $P(x, x) = 0$ pour tout x, ce qui implique d'une part que $Q(x, x) = 0$, d'autre part que $\pi(y) \geq \pi(x)$ pour tout Q-voisin y de x (i.e. tout y tel que $Q(x, y) > 0$). Puisque Q est irréductible, ceci implique de proche en proche que tous les $\pi(x)$ sont égaux à $1/|E|$. Dans cette situation, toutes les propositions sont acceptées, on a donc $P = Q$ et Q est elle-même périodique.

■

Exemple pathologique. Considérons

$$E = \{1, 2\}, \qquad \pi = [1/2, 1/2], \qquad Q = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

alors $P = Q$ et la chaîne de Metropolis est périodique de période 2.

Matrice de transition. Notons

$$\alpha(x, y) = \min\left(\frac{\pi(y)}{\pi(x)}, 1\right) = \min\left(r(x, y), 1\right).$$

La preuve ci-dessus montre que la matrice de la transition de la chaîne X_n est

$$\forall x \neq y, \qquad P(x, y) = \alpha(x, y)Q(x, y)$$

et

$$\forall x \in E, \qquad P(x, x) = 1 - \sum_{y \neq x} \alpha(x, y)Q(x, y).$$

3.2.2 Généralisation : méthode de Metropolis-Hastings

Dans l'algorithme précédent, la dynamique Q qui propose des transitions était supposée agir symétriquement sur l'espace d'états, à savoir que

$$\forall (x,y) \in E \times E, \qquad Q(x,y) = Q(y,x).$$

Autrement dit, sous Q, la probabilité d'aller de x vers y est la même que celle d'aller de y vers x. Cette condition n'est pas nécessaire. Pour s'en émanciper, il suffit de tenir compte de l'éventuelle dissymétrie de Q dans le rapport $r(x,y)$. C'est ainsi que Hastings a proposé, en 1970, de généraliser l'algorithme de Metropolis.

Par rapport à ce dernier, les ingrédients sont les mêmes, le prix à payer pour cette généralisation étant simplement de supposer que :

— partant de tout état x, on sait simuler facilement selon la loi $Q(x,\cdot)$.
— pour tout couple (x,y), on a $Q(x,y) > 0$ implique $Q(y,x) > 0$;
— pour tout couple (x,y) tel que $Q(x,y) > 0$, on sait calculer

$$r(x,y) = \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}.$$

L'algorithme de Metropolis-Hastings fonctionne alors comme suit :

1. partant de $X_n = x$, simuler $Y = y$ selon la loi $Q(x,\cdot)$;
2. calculer le rapport de Metropolis-Hastings

$$r(x,y) = \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)},$$

3. tirer une loi uniforme $U \sim \mathcal{U}_{[0,1]}$, indépendante de toutes les autres variables aléatoires, poser $X_{n+1} = Y = y$ si $U \leq r(x,y)$, et $X_{n+1} = X_n = x$ sinon.

Avec cette nouvelle définition de $r(x,y)$ et en notant encore $\alpha(x,y) = \min(r(x,y),1)$, la matrice de transition P de cette chaîne s'exprime comme précédemment, à savoir

$$P(x,y) = \alpha(x,y)Q(x,y)\mathbf{1}_{x\neq y} + \left(1 - \sum_{y\neq x} \alpha(x,y)Q(x,y)\right)\mathbf{1}_{x=y}. \tag{3.8}$$

En particulier, puisque $\alpha(x,y) \leq 1$, on remarque que, pour tous x et y de E, on a $P(x,y) \leq Q(x,y)$ et, pour tout x,

$$P(x,x) = 1 - \sum_{y\neq x} \alpha(x,y)Q(x,y) \geq 1 - \sum_{y\neq x} Q(x,y) = Q(x,x). \tag{3.9}$$

Proposition 13 (Réversibilité de Metropolis-Hastings)
Si la matrice de transition Q est irréductible et apériodique, alors la chaîne (X_n) produite par l'algorithme de Metropolis-Hastings est irréductible, apériodique et réversible pour π.

Preuve. La réversibilité et l'irréductibilité de la chaîne se prouvent de la même façon que pour la version originale de Metropolis. Concernant l'apériodicité, supposons P périodique, alors en particulier $P(x,x) = 0$ pour tout x, donc (3.9) implique $Q(x,x) = 0$ et

$$\sum_{y\neq x} \alpha(x,y)Q(x,y) = \sum_{y\neq x} Q(x,y),$$

d'où $\alpha(x,y) = 1$ pour tout couple (x,y) tel que $Q(x,y) \neq 0$. Ainsi $P = Q$ et Q serait elle-même périodique, ce qui est exclu.

■

Généralisation : Soit $h : \mathbb{R}_+ \to [0,1]$ vérifiant $h(0) = 0$ et $h(u) = uh(1/u)$ pour $u > 0$. En considérant, pour tout couple (x, y) tel que $Q(x, y) > 0$, le rapport

$$\alpha(x, y) = h\left(\frac{\pi(y)Q(y, x)}{\pi(x)Q(x, y)}\right),$$

on vérifie facilement que la chaîne de noyau de transition

$$P(x, y) = \alpha(x, y)Q(x, y)\mathbf{1}_{x\neq y} + \left(1 - \sum_{x'\neq x} \alpha(x, x')Q(x, x')\right)\mathbf{1}_{x=y}$$

est encore π-réversible. L'algorithme de Metropolis-Hastings classique correspond au choix $h(u) = \min(u, 1)$. Un autre exemple est donné par $h(u) = u/(1 + u)$, ce qui donne le noyau de Barker :

$$\alpha(x, y) = \frac{\pi(y)Q(y, x)}{\pi(y)Q(y, x) + \pi(x)Q(x, y)}.$$

Avec ce choix, comme $0 < \alpha(x, y) < 1$ pour tout couple (x, y) tel que $Q(x, y) > 0$, l'apériodicité de la chaîne (X_n) est claire puisqu'on peut boucler sur tout état :

$$P(x, x) = 1 - \sum_{x'\neq x} \alpha(x, x')Q(x, x') > 1 - \sum_{x'\neq x} Q(x, x') = Q(x, x) \geq 0.$$

Remarque. Tout comme pour la méthode de rejet ou pour l'échantillonnage préférentiel, il existe un noyau de proposition optimal, mais il est hors d'accès. En effet, il suffit de considérer $Q(x, y) = \pi(y)$ pour tout y. La condition $Q(x, y) > 0$ ssi $Q(y, x) > 0$ est bien vérifiée puisque π charge tous les points. De plus, le rapport de Metropolis-Hastings vaut $r(x, y) = 1$, donc toute proposition de transition est acceptée. En fait, quelle que soit la loi initiale, la loi de X_1 est exactement la loi cible π.

3.3 Le recuit simulé

3.3.1 Principe et convergence

Soit E un ensemble fini (très grand) et une fonction $V : E \to \mathbb{R}$ que l'on veut minimiser. On suppose savoir calculer $V(x)$ pour tout x mais l'espace E est trop grand pour qu'on puisse trouver ce minimum par recherche exhaustive : penser par exemple au voyageur de commerce où $|E| = (N - 1)!/2$ avec N le nombre de villes à parcourir.

Pour tout paramètre $T > 0$, la minimisation de V est équivalente à la maximisation de

$$\mu_T(x) = \frac{1}{Z_T} \exp\left(-\frac{1}{T}V(x)\right)$$

dite mesure de Boltzmann-Gibbs associée au potentiel V et à la température T, où

$$Z_T = \sum_{x'\in E} \exp\left(-\frac{1}{T}V(x')\right)$$

est la constante de normalisation, encore appelée fonction de partition en physique statistique. Selon sa valeur, le paramètre T a tendance à accentuer ou à lisser les extrema de V.

Exemple. Caricaturalement, prenons $E = \{1, 2, 3\}$, avec $V(1) = 0$, $V(2) = 1$ et $V(3) = -1$. On vérifie facilement les deux comportements extrêmes suivants

$$\mu_T \xrightarrow[T\to 0]{} [0, 0, 1] \qquad \text{et} \qquad \mu_T \xrightarrow[T\to+\infty]{} \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right].$$

En particulier, lorsque la température tend vers 0, la mesure de Gibbs se concentre sur le minimum de V. Ce résultat est général et l'idée sous-jacente est la même que pour la méthode de Laplace en analyse.

Lemme 2 (Principe de Laplace)
Soit $\mathcal{M}^\star$ l'ensemble des points où V atteint son minimum. Lorsque la température T tend vers zéro, la mesure de Gibbs μ_T se concentre sur ceux-ci, c'est-à-dire que

$$\lim_{T\to 0^+} \mu_T(x) = \begin{cases} 0 & \text{si } x \notin \mathcal{M}^\star \\ \frac{1}{|\mathcal{M}^\star|} & \text{si } x \in \mathcal{M}^\star \end{cases}$$

où $|\mathcal{M}^\star|$ désigne le cardinal de l'ensemble $\mathcal{M}^\star$. On en déduit en particulier que

$$\mu_T(\{x \in E,\ V(x) > V_\star\}) \xrightarrow[T\to 0^+]{} 0. \tag{3.10}$$

Ainsi, lorsque V atteint son minimum global en un unique point $x^\star$, ceci signifie que la mesure de Gibbs se concentre en $x^\star$ lorsque la température tend vers 0 (cf. exemple précédent avec $x^\star = 3$).

Preuve. Soit $V^\star$ le minimum de V, alors pour tout x de E

$$\mu_T(x) = \frac{\exp\left(-\frac{1}{T}(V(x) - V_\star)\right)}{\sum_{x'\in E} \exp\left(-\frac{1}{T}(V(x') - V_\star)\right)} = \frac{\exp\left(-\frac{1}{T}(V(x) - V_\star)\right)}{|\mathcal{M}^\star| + \sum_{x'\in E\setminus\mathcal{M}^\star} \exp\left(-\frac{1}{T}(V(x') - V_\star)\right)},$$

avec

$$\forall x' \in E \setminus \mathcal{M}^\star, \qquad \exp\left(-\frac{1}{T}(V(x') - V_\star)\right) \xrightarrow[T\to 0^+]{} 0.$$

La limite du numérateur est, quant à elle, 1 ou 0 selon que x appartient à $\mathcal{M}^\star$ ou non.

■

Remarque : A contrario, lorsque $T \to \infty$, on a tout simplement

$$\forall x \in E, \qquad \mu_T(x) \xrightarrow[T\to+\infty]{} \frac{1}{|E|},$$

c'est-à-dire que la mesure de Gibbs converge vers la loi uniforme sur E à haute température.

Puisqu'à basse température la mesure de Gibbs se concentre sur les minima globaux de V, une idée consiste à se donner une suite de températures (T_n) décroissant vers 0 et à simuler une chaîne (X_n) qui, à chaque étape n, transite selon un noyau de loi stationnaire $\mu_n = \mu_{T_n}$. Pour une matrice de transition $Q(x, y)$ donnée et vérifiant les propriétés requises par Metropolis-Hastings (irréductibilité et apériodicité), c'est ce que fait l'algorithme suivant, appelé **recuit simulé** [2] :

1. partant de $X_{n-1} = x$, simuler $Y = y$ selon la loi $Q(x, \cdot)$;
2. calculer le rapport de Metropolis-Hastings

$$r_n(x, y) = \frac{\mu_n(y)Q(y, x)}{\mu_n(x)Q(x, y)} = \frac{Q(y, x)}{Q(x, y)} \exp\left(-\frac{1}{T_n}(V(y) - V(x))\right);$$

2. *simulated annealing*, expression venant de la metallurgie.

3. tirer une loi uniforme $U \sim \mathcal{U}_{[0,1]}$, indépendante de toutes les autres variables aléatoires, poser $X_n = Y = y$ si $U \leq r_n(x,y)$, et $X_n = X_{n-1} = x$ sinon.

Remarque : Comme pour Metropolis-Hastings, (X_n) est bien une chaîne de Markov, mais elle n'est plus homogène en raison des changements constants de températures T_n.

Si les transitions sont symétriques, i.e. $Q(x,y) = Q(y,x)$, on obtient un algorithme de Metropolis inhomogène avec le rapport

$$r_n(x,y) = \exp\left(-\frac{1}{T_n}(V(y) - V(x))\right),$$

lequel est supérieur ou égal à 1 dès lors que $V(y) \leq V(x)$. On accepte donc systématiquement une transition vers une valeur de V plus basse, mais on ne refuse pas systématiquement une transition vers une valeur plus élevée, ceci afin d'éviter de rester piégé dans un minimum local de V. On voit aussi que, à x et y fixés tels que $V(y) > V(x)$, plus la température T_n sera basse, moins on acceptera de transiter vers une valeur plus élevée. Le schéma de température (T_n) joue clairement un rôle critique dans cette histoire. Le résultat suivant fournit une réponse partielle à la question du réglage de la suite (T_n).

Théorème 5 (Convergence du Recuit Simulé)
Si Q est irréductible et apériodique, il existe une constante c_0 dépendant de la fonction V et de la matrice de transition Q telle que, pour tout $c > c_0$, l'algorithme précédent appliqué avec le schéma de température $T_n = c/\log n$ et pour toute loi initiale ν_0, on a

$$\lim_{n\to\infty} \mathbb{P}(V(X_n) > V_\star) = 0.$$

Hypothèses de la preuve. Cette preuve est tirée du livre de Jean-François Delmas et Benjamin Jourdain [11]. Comme eux, nous nous contenterons de démontrer ce résultat dans le cas où $Q(x,y) = Q(y,x)$ (cadre Metropolis) et en supposant Q irréductible et apériodique. Par ailleurs, la constante c_0 apparaissant dans la preuve n'est pas la constante optimale : nous reviendrons sur ce point ultérieurement.

La preuve fait appel à plusieurs résultats intermédiaires. De façon générale, on dit que Q vérifie une **condition de Doeblin** s'il existe $a > 0$, une loi de probabilité π et $\ell \in \mathbb{N}^\star$ tels que

$$\forall (x,y) \in E^2,\ Q^\ell(x,y) \geq a\pi(y). \tag{3.11}$$

Dans ce cas, il est clair que $a \geq 1$. On commence par noter que si Q est irréductible et apériodique, le critère (3.4) assure qu'elle vérifie une condition de Doeblin : il suffit en effet de considérer $\ell = \max_{(x,y)} n_0(x,y)$,

$$\pi(y) = \frac{\min_x Q^\ell(x,y)}{\sum_y \min_{x'} Q^\ell(x',y)} \text{ et } a = \sum_y \min_{x'} Q^\ell(x',y).$$

On a alors, pour toutes mesures de probabilité μ et ν,

$$\|\nu Q^\ell - \mu Q^\ell\|_{vt} = \frac{1}{2}\sum_{y\in E} |(\nu Q^\ell)(y) - (\mu Q^\ell)(y)| = \frac{1}{2}\sum_{y\in E}\left|\sum_{x\in E}(\nu(x) - \mu(x))Q^\ell(x,y)\right|.$$

Puisque μ et ν somment toutes deux à 1, on peut encore écrire

$$\|\nu Q^\ell - \mu Q^\ell\|_{vt} = \frac{1}{2}\sum_{y\in E}\left|\sum_{x\in E}(\nu(x) - \mu(x))(Q^\ell(x,y) - a\pi(y))\right|$$

et, par la condition de Doeblin,

$$\|\nu Q^\ell - \mu Q^\ell\|_{vt} \leq \frac{1}{2}\sum_{x\in E}|\nu(x)-\mu(x)|\sum_{y\in E}(Q^\ell(x,y)-a\pi(y)),$$

or

$$\sum_{y\in E}(Q^\ell(x,y)-a\pi(y)) = 1-a,$$

donc

$$\|\nu Q^\ell - \mu Q^\ell\|_{vt} \leq (1-a)\|\nu-\mu\|_{vt}. \tag{3.12}$$

Ainsi Q^ℓ agit comme une application contractante de rapport $1-a<1$ sur l'ensemble $\mathcal{P}(E)$ des mesures de probabilité sur E. Notons que, de façon générale, si l'on sait juste que Q est une matrice de transition, alors on peut seulement dire que

$$\|\nu Q - \mu Q\|_{vt} \leq \frac{1}{2}\sum_{x\in E}|\nu(x)-\mu(x)|\sum_{y\in E}Q(x,y) = \|\nu-\mu\|_{vt}. \tag{3.13}$$

Nous convenons de noter

$$\delta = \max_{(x,y)\in E^2}(V(y)-V(x))\mathbf{1}_{Q(x,y)>0}.$$

Comme Q est supposée symétrique, il est clair que $\delta \geq 0$. C'est la barrière maximale d'énergie que Q est susceptible de franchir en un coup. Elle intervient dans le résultat suivant, qui quantifie l'effet contractant du recuit simulé.

Lemme 3 (Effet contractant du schéma de température)
Pour toutes mesures de probabilités μ et ν, on a

$$\|(\nu-\mu)P_{n+1}\cdots P_{n+\ell}\|_{vt} \leq \left(1 - a\ e^{-\frac{\delta\ell}{c}\log(n+\ell)}\right)\|\nu-\mu\|_{vt}.$$

Preuve. Pour tout $n\in\mathbb{N}^\star$, si $x\neq y$, le noyau de transition P_n à température T_n s'écrit, en notant $x_+ = \max(x,0)$,

$$P_n(x,y) = e^{-\frac{1}{T_n}(V(y)-V(x))_+}Q(x,y) \geq e^{-\frac{\delta}{T_n}}Q(x,y).$$

Par ailleurs, si $x\neq y$, il est clair que $P_n(x,y) = r_n(x,y)Q(x,y)) \leq Q(x,y)$, donc

$$P_n(x,x) = 1-\sum_{y\neq x}P_n(x,y) \geq 1-\sum_{y\neq x}Q(x,y) = Q(x,x) \geq e^{-\frac{\delta}{T_n}}Q(x,x),$$

si bien que

$$\forall (x,y)\in E\times E, \qquad P_n(x,y) \geq e^{-\frac{\delta}{T_n}}Q(x,y).$$

Il ne reste plus qu'à itérer cette inégalité et appliquer la condition de Doeblin (3.11) :

$$(P_{n+1}\cdots P_{n+\ell})(x,y) \geq e^{-\delta\left(\frac{1}{T_{n+1}}+\cdots+\frac{1}{T_{n+\ell}}\right)}Q^\ell(x,y) \geq e^{-\frac{\delta\ell}{c}\log(n+\ell)}a\pi(y).$$

Or ceci n'est rien d'autre qu'une condition de Doeblin pour la matrice de transition

$$P_{n+1}\cdots P_{n+\ell}$$

et l'on peut donc appliquer (3.12) pour arriver au résultat annoncé.

■

Nous aurons également besoin du résultat suivant pour contrôler la distance entre deux mesures de Gibbs en fonction de l'écart de température.

Lemme 4 (Distance entre mesures de Gibbs)
Notons $\Delta(V) = \max_{x\in E} V(x) - V^\star$, *alors*

$$\forall (T,T') \in]0,\infty[^2 \qquad \|\mu_T - \mu_{T'}\|_{vt} \leq \left|\frac{1}{T} - \frac{1}{T'}\right| \Delta(V).$$

Preuve. La mesure de Gibbs est inchangée si on remplace V par $V - V_\star$ (cf. par exemple la première égalité de la preuve du Lemme 2), donc on peut supposer sans perte de généralité que $V \geq 0$ et $V_\star = 0$, de sorte que $\Delta(V) = \max_{x\in E} V(x)$. Toujours sans perte de généralité, on va supposer $T \geq T' > 0$. En remarquant que $0 \leq 1 - e^{-x} \leq x$ pour tout $x \geq 0$, il vient

$$\left|e^{-\frac{V(x)}{T}} - e^{-\frac{V(x)}{T'}}\right| = e^{-\frac{V(x)}{T}} \left|1 - e^{-\left(\frac{1}{T'}-\frac{1}{T}\right)V(x)}\right| \leq \left(\frac{1}{T'} - \frac{1}{T}\right) V(x) e^{-\frac{V(x)}{T}} \leq \left(\frac{1}{T'} - \frac{1}{T}\right) \Delta(V) e^{-\frac{V(x)}{T}}$$

et la sommation sur x donne

$$|Z_T - Z_{T'}| \leq \sum_{x\in E} \left|e^{-\frac{V(x)}{T}} - e^{-\frac{V(x)}{T'}}\right| \leq \left(\frac{1}{T'} - \frac{1}{T}\right) \Delta(V) Z_T,$$

d'où, en divisant par $Z_T Z_{T'}$,

$$\left|\frac{1}{Z_T} - \frac{1}{Z_{T'}}\right| \leq \left(\frac{1}{T'} - \frac{1}{T}\right) \Delta(V) \frac{1}{Z_{T'}}.$$

On a alors

$$2\|\mu_T - \mu_{T'}\|_{vt} = \sum_{x\in E} \left|\frac{1}{Z_T} e^{-\frac{V(x)}{T}} - \frac{1}{Z_{T'}} e^{-\frac{V(x)}{T'}}\right| \leq \sum_{x\in E} \frac{1}{Z_T} \left|e^{-\frac{V(x)}{T}} - e^{-\frac{V(x)}{T'}}\right| + \left|\frac{1}{Z_T} - \frac{1}{Z_{T'}}\right| \sum_{x\in E} e^{-\frac{V(x)}{T'}}$$

c'est-à-dire, d'après ce qui précède,

$$2\|\mu_T - \mu_{T'}\|_{vt} \leq 2\left(\frac{1}{T'} - \frac{1}{T}\right) \Delta(V).$$

■

Nous avons besoin d'un dernier résultat élémentaire avant de passer à la preuve proprement dite.

Lemme 5 (Convergence de suite)
Soit (a_n), (b_n) *deux suites positives telles que* $0 < a_n < 1$ *pour tout* n, $\sum_n a_n = \infty$, $\lim b_n/a_n = 0$, *et une suite* (u_n) *telle que* $u_0 \geq 0$ *et, pour tout* $n \geq 0$,

$$u_{n+1} \leq (1 - a_n) u_n + b_n,$$

alors (u_n) *tend vers 0.*

Preuve. L'inégalité $\log(1-x) \leq -x$ pour tout $x < 1$ implique, pour tout n_0,

$$\prod_{i=n_0}^{n_0+p-1} (1 - a_i) = e^{\sum_{i=n_0}^{n_0+p-1} \log(1-a_i)} \leq e^{-\sum_{i=n_0}^{n_0+p-1} a_i} \xrightarrow[p\to\infty]{} 0.$$

Par ailleurs, pour tout $\varepsilon > 0$, il existe $n_0 = n_0(\varepsilon)$ tel que pour tout $n \geq n_0$ on ait $b_n \leq \varepsilon a_n$, donc

$$u_{n+1} \leq (1 - a_n) u_n + b_n \leq (1 - a_n) u_n + \varepsilon a_n,$$

ce qui implique, pour tout $n \geq n_0$,

$$u_{n+1} - \varepsilon \leq (1 - a_n)(u_n - \varepsilon),$$

d'où il découle que pour tout $p > 0$

$$u_{n_0+p} - \varepsilon \leq \left\{ \prod_{i=n_0}^{n_0+p-1} (1 - a_i) \right\} (u_{n_0} - \varepsilon),$$

ou encore

$$u_{n_0+p} \leq \varepsilon + \left\{ \prod_{i=n_0}^{n_0+p-1} (1 - a_i) \right\} (u_{n_0} - \varepsilon).$$

Par conséquent

$$\limsup_{n\to\infty} u_n = \limsup_{p\to\infty} u_{n_0+p} \leq \varepsilon,$$

et puisque ε est arbitraire, le résultat est établi.

■

Preuve du Théorème 5. En notant $T_n = c/\log n$ la température à l'étape n, la mesure de Gibbs $\mu_n \propto \exp(-V/T_n)$ associée vérifie $\mu_n P_n = \mu_n$. Par ailleurs, partant d'une loi initiale ν_0, la loi de X_n est donc

$$\nu_n = \nu_0 P_1 \cdots P_n.$$

Plus généralement, si n et ℓ sont deux entiers naturels,

$$\nu_{n+\ell} = \nu_n P_{n+1} \cdots P_{n+\ell},$$

et, puisque pour tout $1 \leq k \leq \ell - 1$,

$$\mu_{n+k} P_{n+k} \cdots P_{n+\ell} = \mu_{n+k} P_{n+k+1} \cdots P_{n+\ell},$$

on a la décomposition télescopique

$$\nu_{n+\ell} - \mu_{n+\ell} = (\nu_n - \mu_n) P_{n+1} \cdots P_{n+\ell} + \sum_{k=1}^{\ell} (\mu_{n+k-1} - \mu_{n+k}) P_{n+k} \cdots P_{n+\ell}.$$

Pour tout k, la relation (3.13) donne

$$\|(\mu_{n+k-1} - \mu_{n+k}) P_{n+k} \cdots P_{n+\ell}\|_{vt} \leq \|\mu_{n+k-1} - \mu_{n+k}\|_{vt}$$

donc

$$\|\nu_{n+\ell} - \mu_{n+\ell}\|_{vt} \leq \|(\nu_n - \mu_n) P_{n+1} \cdots P_{n+\ell}\|_{vt} + \sum_{k=1}^{\ell} \|\mu_{n+k-1} - \mu_{n+k}\|_{vt}. \tag{3.14}$$

Prenons $j \in \mathbb{N}^\star$ et $n = j\ell$, alors les Lemmes 3 et 4 impliquent

$$\|\nu_{(j+1)\ell} - \mu_{(j+1)\ell}\|_{vt} = \|\nu_{n+\ell} - \mu_{n+\ell}\|_{vt} \leq (1 - a_j)\|\nu_n - \mu_n\|_{vt} + b_j = (1 - a_j)\|\nu_{j\ell} - \mu_{j\ell}\|_{vt} + b_j,$$

où l'on a noté

$$a_j = a\, e^{-\frac{\delta\ell}{c}\log(\ell(j+1))} = \frac{a'}{j^{\frac{\delta\ell}{c}}},$$

et

$$b_j = \frac{\Delta(V)}{c}\sum_{k=1}^{\ell}(\log(j\ell+k)-\log(j\ell+k-1)) = \frac{\Delta(V)}{c}\log\left(1+\frac{1}{j}\right) \sim \frac{\Delta(V)}{cj},$$

les équivalents étant considérés lorsque j tend vers l'infini. En posant $u_j = \|\nu_{j\ell} - \mu_{j\ell}\|_{vt}$, on a bien $0 < a_j < 1$ pour tout j et le Lemme 5 s'applique dès lors que $c > c_0 := \delta\ell$, i.e.

$$\|\nu_{j\ell} - \mu_{j\ell}\|_{vt} \xrightarrow[j\to\infty]{} 0.$$

Ensuite, pour tout $p \in \{1, \cdots, \ell-1\}$, on déduit de (3.13), (3.14) et du Lemme 4 que

$$\|\nu_{j\ell+p} - \mu_{j\ell+p}\|_{vt} \le \|\nu_{j\ell} - \mu_{j\ell}\|_{vt} + \frac{\Delta(V)}{c}\log\left(1+\frac{p}{j\ell}\right) \xrightarrow[j\to\infty]{} 0.$$

Au total, on a prouvé que, si $c > c_0 = \delta\ell$,

$$\|\nu_n - \mu_n\|_{vt} \xrightarrow[n\to\infty]{} 0.$$

Ainsi, puisque $\|\nu_n - \mu_n\|_{vt} = \max_{A\subset E}|\nu_n(A) - \mu_n(A)|$, il vient

$$|\mathbb{P}(V(X_n) > V_\star) - \mu_n(\{x,\ V(X_n) > V_\star\})| \le \|\nu_n - \mu_n\|_{vt} \xrightarrow[n\to\infty]{} 0,$$

et la relation (3.10) permet de conclure que

$$\mathbb{P}(V(X_n) > V_\star) \xrightarrow[n\to\infty]{} 0.$$

■

Si ce résultat est satisfaisant théoriquement, il l'est moins en pratique car la convergence est très lente. C'est cependant une question très délicate en toute généralité. Quoi qu'il en soit, il est indispensable de stocker toutes les valeurs intermédiaires $V(X_n)$ et, en fin d'algorithme (disons après N étapes), de considérer non pas la dernière valeur $V(X_N)$ mais la valeur minimale $\min_{1\le n\le N} V(X_n)$ parmi toutes celles rencontrées. Cette règle de bon sens est en fait très difficile à prendre en compte d'un point de vue théorique.

Remarque : interprétation de $c_0^\star$. Comme indiqué précédemment, la constante c_0 apparaissant dans la preuve ci-dessus n'est pas la constante optimale. Supposons pour simplifier que le minimum $V^\star$ est atteint en un unique point $x^\star$ et que Q est symétrique (cas Metropolis). Pour tout point x, notons $\mathcal{C}(x, x^\star)$ l'ensemble des chemins menant de x à $x^\star$ via la relation de voisinage imposée par la matrice de transition Q, et

$$c_0^\star = c_0^\star(V, Q) = \max_{x\in E}\min_{\mathcal{C}\in\mathcal{C}(x,x^\star)}\max_{y\in\mathcal{C}}(V(y) - V(x))_+.$$

Alors un résultat dû à Hajek [19] assure que

$$\mathbb{P}(V(X_n) > V_\star) \xrightarrow[n\to\infty]{} 0$$

ssi (T_n) tend vers 0 et

$$\sum_{n=1}^{\infty} e^{-\frac{c_0^\star}{T_n}} = +\infty.$$

Ainsi, pour un schéma de température à décroissance logarithmique de la forme $T_n = c/\log n$, on a $\lim_{n\to\infty}\mathbb{P}(V(X_n) > V_\star) = 0$ ssi $c \ge c_0^\star$. De même, pour un schéma de température constant par paliers de la forme $T_n = 1/k$ si $e^{c(k-1)} \le n < e^{ck}$, le résultat est assuré ssi $c \ge c_0^\star$. On trouvera plus d'informations ainsi que des références dans le livre de Delmas et Jourdain [11].

3.3.2 Un mot sur les algorithmes génétiques

Cette section est tirée du livre de Bernard Ycart, *Modèles et algorithmes markoviens* [35]. Proposés dès les années 60, donc bien avant avant le recuit simulé (qui date du début des années 80), les algorithmes génétiques s'inspirent des mécanismes de la sélection naturelle. Certaines versions peuvent cependant être vues comme des algorithmes de recuit simulé en interaction.

Dans la version originale, le but est de maximiser une fonction f, dite fonction d'adaptation, définie sur $E = \{0,1\}^d$ et à valeurs dans $\mathbb{R}_+$. Pour ce faire, on considère une chaîne de Markov (X_n), où $X_n = (X_n^1, \ldots, X_n^N) \in E^N$ peut être vu comme une population de chromosomes, c'est-à-dire un N-uplet de mots binaires de longueur d (les chromosomes). Le passage de X_n à X_{n+1} s'effectue en 3 étapes :

$$X_n \xrightarrow{\text{mutation}} Y_n \xrightarrow{\text{croisement}} Z_n \xrightarrow{\text{sélection}} X_{n+1}$$

que l'on peut décrire grossièrement comme suit :

- Mutation : soit $p \in]0,1[$, alors pour chaque lettre de chaque chromosome, on décide de la modifier avec probabilité p et de la laisser inchangée avec probabilité $(1-p)$. Ceci donne une nouvelle population $Y_n = (Y_n^1, \ldots, Y_n^N)$.
- Croisement : soit $q \in]0,1[$, on forme $N/2$ couples de façon arbitraire, par exemple en considérant $(Y_n^1, Y_n^2), (Y_n^3, Y_n^4)$, etc. Pour chaque couple on décide avec probabilité q si le croisement a lieu. Dans ce cas, on tire un entier au hasard (uniforme) entre 1 et $(d-1)$ et les segments finaux des deux chromosomes sont échangés. Ceci donne une nouvelle population $Z_n = (Z_n^1, \ldots, Z_n^N)$.
- Sélection : Les N chromosomes de la population X_{n+1} sont choisis par tirage multinomial dans la population Z_n avec les poids

$$W_n^j := \frac{f(Z_n^j)}{\sum_{\ell=1}^N f(Z_n^\ell)}.$$

Pour pouvoir analyser la convergence de cet algorithme vers le ou les maximaux globaux de f, il convient de faire dépendre p, q et les poids W de l'itération n. Comme pour le recuit simulé, l'idée est, au fur et à mesure des itérations, de favoriser de moins en moins les mutations (donc $p = p_n$ et $q = q_n$ décroissants avec n) et de se concentrer de plus en plus sur les endroits où f est maximale, par exemple en considérant des poids de la forme

$$W_n^j := \frac{e^{\frac{f(Z_n^j)}{T_n}}}{\sum_{\ell=1}^N e^{\frac{f(Z_n^\ell)}{T_n}}},$$

en considérant une suite (T_n) décroissante vers 0.

Les premiers résultats théoriques généraux sur ce type d'algorithmes sont dus à Raphaël Cerf [3, 4, 5] . Ils sont néanmoins très techniques et nous ne les détaillerons pas ici. Nous nous contenterons de décrire une variante due à Olivier François, appelée MOSES pour Mutation-or-Selection Evolutionary Strategy [17]. Si tant l'algorithme que le résultat de convergence sont simples dans leurs formulations, la preuve l'est nettement moins, c'est pourquoi nous l'admettrons.

Algorithme MOSES

Par analogie avec le recuit simulé, nous revenons à une fonction V que l'on cherche à minimiser, supposée bijective par souci de simplicité, et notons $V_\star$ son minimum global. On associe à E un graphe non orienté $(E, \mathcal{G})$ connexe. Son diamètre est noté D, c'est le maximum sur l'ensemble de tous les couples (x, x') de la distance (i.e. le plus court trajet) entre x et x'. On note (T_n) le schéma de température.

A chaque étape, partant de $X_n = (X_n^1, \ldots, X_n^N) \in E^N$, en notant $X_n^\star = \text{argmin}_{1\leq j\leq N} V(X_n^j)$, on procède comme suit : $\forall 1 \leq j \leq N$, $X_{n+1}^j = X_n^\star$ avec probabilité $1 - e^{-1/T_n}$; sinon, c'est-à-dire avec probabilité e^{-1/T_n}, X_{n+1}^j est choisi au hasard (uniforme) parmi les voisins [3] de X_n^j pour la relation de voisinage induite par le graphe $(E, \mathcal{G})$.

Théorème 6
Si $N > D$ et si le schéma de température (T_n) tend vers 0 et vérifie

$$\sum_{n=1}^{\infty} e^{-\frac{D}{T_n}} = \infty,$$

alors, pour toute condition initiale,

$$\lim_{n\to\infty} \mathbb{P}(V(X_n^\star) > V_\star) = 0.$$

Le théorème énoncé en [17] est à la fois plus général et plus précis. De plus, à l'instar du résultat de Hajek mentionné précédemment, il existe en fait une constante optimale $D^\star$, dépendant du graphe $\mathcal{G}$ et de V, telle que pour toute suite (T_n) décroissant vers 0, le résultat précédent est vérifié si et seulement si

$$\sum_{n=1}^{\infty} e^{-\frac{D^\star}{T_n}} = \infty.$$

3.4 Espace d'états général

Tout ce qui a été dit dans le cas d'un espace d'états fini se généralise à un espace d'états infini, qu'il soit dénombrable ou continu. Par exemple, dans le cas d'un espace d'états continu, on peut avoir en tête $E = \mathbb{R}^d$ muni de la tribu borélienne $\mathcal{E} = \mathcal{B}(\mathbb{R}^d)$ et de la mesure de Lebesgue $\lambda(dx) = dx$. Rappelons que le principe des méthodes MCMC est de construire une chaîne de Markov convergeant vers une mesure de probabilité connue $\pi(dx)$ sur $(E, \mathcal{E})$. Le but de cette section est uniquement d'introduire de façon informelle le cadre des chaînes de Markov à espaces d'états généraux, de préciser les notations et de montrer que l'algorithme de Metropolis-Hastings est aussi facile à implémenter que dans le cas d'un espace d'états fini. En revanche, les résultats théoriques, nettement plus techniques, ne seront pas détaillés.

3.4.1 Noyaux de transition et chaînes de Markov

La matrice de transition $P = [P(x,y)]_{(x,y)\in E^2}$ est remplacée par un **noyau de transition** P, c'est-à-dire :

1. pour tout x de E, $P(x,\cdot)$ est une mesure de probabilité sur $(E, \mathcal{E})$;
2. pour tout ensemble B de $\mathcal{E}$, l'application $x \mapsto P(x,B)$ est mesurable.

A l'instar de son homologue discret, le noyau P peut se voir comme une façon de se déplacer dans l'espace d'états E. Ainsi, pour $E = \mathbb{R}^d$, la variable $x \in \mathbb{R}^d$ étant fixée, si $P(x,dy) = p(x,y)dy$, la quantité $p(x,y)$ s'interprète comme la densité de probabilité d'aller en y sachant que l'on part du point x.

La suite de variables aléatoires (X_n) est une chaîne de Markov homogène de noyau de transition P si, pour tout B de $\mathcal{E}$,

$$\mathbb{P}(X_{n+1} \in B | X_0, \cdots, X_n) = \mathbb{P}(X_{n+1} \in B | X_n) = P(X_n, B),$$

3. en excluant $X_n^\star$ le cas échéant.

ou de façon équivalente si, pour toute fonction mesurable bornée (dite fonction test) $\varphi : E \to \mathbb{R}$, on a

$$\mathbb{E}[\varphi(X_{n+1})|X_0, \cdots, X_n] = \mathbb{E}[\varphi(X_{n+1})|X_n] = (P\varphi)(X_n),$$

avec, pour tout $x \in E$,

$$(P\varphi)(x) = \int_E P(x, dy)\varphi(y).$$

Comme dans le cas discret, on peut montrer que ceci équivaut à l'existence d'une suite de variables aléatoires $(W_n)_{n\geq 1}$ i.i.d. et indépendantes de X_0 et d'une fonction h telles que

$$\forall n \geq 1, \qquad X_{n+1} = h(X_n, W_{n+1}).$$

Exemple : noyau gaussien. Soit $E = \mathbb{R}$ et $\sigma^2 > 0$ fixé. On appelle noyau de transition gaussien le noyau défini pour tout couple $(x, y) \in \mathbb{R}^2$ par

$$P(x, dy) = p(x, y)dy = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-x)^2}{2\sigma^2}} dy.$$

Autrement dit, une chaîne de Markov de noyau de transition P est définie par sa condition initiale $X_0 = x$ (ou plus généralement sa loi initiale μ_0, mesure de probabilité sur $\mathbb{R}$) et, si $X_n = x$, alors $X_{n+1} \sim \mathcal{N}(x, \sigma^2)$, c'est-à-dire de façon générale

$$X_{n+1} = X_n + \sigma W_{n+1},$$

où $(W_n)_{n\geq 1}$ est une suite de variables i.i.d. gaussiennes standards indépendantes de X_0. Le paramètre de réglage σ donne la taille moyenne du pas effectuée à chaque étape : un calcul élémentaire montre en effet que

$$\mathbb{E}[|X_{n+1} - X_n|] = \sigma\mathbb{E}[|W_{n+1}|] = \sigma\mathbb{E}[|\mathcal{N}(0, 1)|] = \sigma\sqrt{\frac{2}{\pi}}.$$

Notations, définitions et propriétés. Mutatis mutandis, toutes les propriétés classiques vues dans le cas d'un espace d'états fini passent dans un cadre plus général. En particulier, comme dans le cas discret, nous conviendrons qu'un noyau de transition agit **à gauche** sur les mesures et **à droite** sur les fonctions. Ainsi, si (X_n) est une chaîne de Markov de noyau de transition P et si $X_0 \sim \mu_0$, alors la loi μ_1 de X_1 est donnée par

$$\mu_1(dy) = (\mu_0 P)(dy) = \int_E \mu_0(dx)P(x, dy).$$

Comme indiqué ci-dessus, si $\varphi : E \to \mathbb{R}$ est une fonction mesurable bornée, on écrit

$$(P\varphi)(x) = \int_E P(x, dy)\varphi(y) = \mathbb{E}[\varphi(X_1)|X_0 = x],$$

de sorte que, si $X_0 \sim \mu_0$,

$$\mathbb{E}[\varphi(X_1] = \mu_0 P\varphi = \iint_{E^2} \mu_0(dx)P(x, dy)\varphi(y).$$

En notant P_1P_2 le noyau de transition défini par

$$(P_1P_2)(x, dy) = \int_E P_1(x, dx')P_2(x', dy),$$

on peut définir en particulier de proche en proche P^n, avec la convention $P^0 = I_d$, et le théorème de Chapman-Kolmogorov assure que, pour tout $n \geq 0$, la variable X_n a pour loi $\mu_n = \mu_0 P^n$.

On dit que la loi de probabilité π est stationnaire (ou invariante, ou d'équilibre) pour le noyau de transition P si $\pi P = \pi$, i.e.

$$\forall y \in E, \qquad \pi(dy) = \int_E \pi(dx) P(x, dy).$$

Si $E = \mathbb{R}^d$, si $\pi(dx) = \pi(x)dx$ et $P(x, dy) = p(x, y)dy$, ceci s'écrit encore

$$\forall y \in E, \qquad \pi(y) = \int_E \pi(x) p(x, y) dx.$$

On dit que le noyau de transition P est réversible pour la loi de probabilité π si les équations d'équilibre détaillé sont vérifiées :

$$\forall (x, y) \in E^2, \qquad \pi(dx)P(x, dy) = \pi(dy)P(y, dx). \tag{3.15}$$

Si $E = \mathbb{R}^d$, $\pi(dx) = \pi(x)dx$ et $P(x, dy) = p(x, y)dy$, ceci s'écrit encore

$$\forall (x, y) \in E^2, \qquad \pi(x)p(x, y) = \pi(y)p(y, x).$$

Comme dans le cas discret, il est facile de voir que si P est réversible pour π, alors π est stationnaire pour P.

Exemple : noyau gaussien. Soit le noyau de transition P défini pour tout couple $(x, y) \in \mathbb{R}^2$ par

$$P(x, dy) = \frac{1}{\sqrt{\pi}} e^{-\left(y - \frac{x}{\sqrt{2}}\right)^2} dy.$$

Dit autrement, si (X_n) est une chaîne de Markov de noyau de transition P, elle admet la représentation suivante sous forme de récurrence aléatoire :

$$X_{n+1} = \frac{X_n}{\sqrt{2}} + \frac{1}{\sqrt{2}} W_{n+1},$$

où $(W_n)_{n \geq 1}$ est une suite de variables i.i.d. gaussiennes standards et indépendantes de X_0. Il est clair que si $X_0 \sim \mathcal{N}(0, 1)$, alors il en va de même pour tout X_n, c'est-à-dire que $\pi = \mathcal{N}(0, 1)$ est une loi invariante pour cette chaîne. On vérifie en fait sans problème que la chaîne est réversible pour π.

3.4.2 Algorithme de Metropolis-Hastings

Le but est de simuler suivant une loi de probabilité π sur E. Les hypothèses requises pour l'algorithme de Metropolis-Hastings dans le cas général sont alors la traduction pure et simple de celles du cas discret. Pour simplifier, considérons le cas d'une loi cible absolument continue $\pi(dx) = \pi(x)dx$ et un noyau Q à densité, i.e. $Q(x, dy) = q(x, y)dy$, et supposons que :

— pour tout point x, on sait simuler selon le noyau $Q(x, dy)$;
— pour tout couple (x, y), on a $q(x, y) > 0$ implique $q(y, x) > 0$;
— pour tout couple (x, y) tel que $q(x, y) > 0$, on sait calculer le rapport de Metropolis-Hastings

$$r(x, y) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}.$$

L'algorithme de Metropolis-Hastings fonctionne alors comme dans le cas discret :

1. partant de $X_n = x$, simuler $Y_{n+1} = y$ selon la loi $Q(x, dy)$;
2. calculer le rapport de Metropolis-Hastings $r(x, y)$;

3. tirer une loi uniforme $U_{n+1} \sim \mathcal{U}_{[0,1]}$, poser $X_{n+1} = Y = y$ si $U_{n+1} \leq r(x,y)$, et $X_{n+1} = X_n = x$ sinon.

Remarque. En notant $\alpha(x,y) = \min(1, r(x,y))$, la représentation sous forme de récurrence aléatoire

$$X_{n+1} = Y_{n+1}\mathbf{1}_{U_{n+1}\leq\alpha(X_n,Y_{n+1})} + X_n\mathbf{1}_{U_{n+1}>\alpha(X_n,Y_{n+1})}$$

assure que (X_n) est une chaîne de Markov homogène. Bien que Q soit à densité, ce n'est pas le cas du noyau de transition P de la chaîne (X_n) et ceci en raison de la probabilité de rester sur place, laquelle implique une mesure de Dirac. Plus précisément, on a

$$P(x,dy) = \alpha(x,y)q(x,y)dy + \left(1 - \int_E \alpha(x,x')q(x,x')dx'\right)\delta_x(dy). \tag{3.16}$$

En effet, on cherche à déterminer pour toute fonction test φ le noyau P tel que

$$(P\varphi)(X_n) = \mathbb{E}[\varphi(X_{n+1})|X_n] = \int \varphi(y)P(X_n,dy).$$

Par définition de l'algorithme, en notant $Y_{n+1} = Y$ et $U_{n+1} = U$ pour alléger les notations, on peut écrire

$$(P\varphi)(X_n) = \mathbb{E}[\mathbb{E}[\varphi(X_{n+1})|X_n,Y]|X_n] = \mathbb{E}[\mathbb{E}[\varphi(Y)\mathbf{1}_{U_{\leq}\alpha(X_n,Y)} + \varphi(X_n)\mathbf{1}_{U>\alpha(X_n,Y)}|X_n,Y]|X_n],$$

d'où, en tenant compte du fait que U est uniforme et indépendante de X_n et Y,

$$(P\varphi)(X_n) = \mathbb{E}[\varphi(Y)\alpha(X_n,Y) + \varphi(X_n)(1-\alpha(X_n,Y))|X_n].$$

Puisque, de façon générale,

$$\mathbb{E}[h(X_n,Y)|X_n] = \int_E h(X_n,y)q(X_n,y)dy,$$

ceci s'écrit encore

$$(P\varphi)(X_n) = \int_E \varphi(y)\alpha(X_n,y)q(X_n,y)dy + \varphi(X_n)\int_E (1-\alpha(X_n,y))q(X_n,y)dy = \int_E \varphi(y)P(X_n,dy).$$

avec P défini en (3.16).

Lemme 6
La chaîne (X_n) est réversible pour la loi π.

Preuve. Les équations d'équilibre détaillé (3.15) étant toujours vérifiées pour $x = y$, il suffit de vérifier que

$$\pi(dx)P(x,dy) = \pi(dy)P(y,dx)$$

pour $x \neq y$, or dans ce cas tout est simple puisqu'on a des densités :

$$\pi(dx)P(x,dy) = \pi(x)\alpha(x,y)q(x,y)dxdy = \pi(x)q(x,y)\min\left(1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right)dxdy,$$

c'est-à-dire, par symétrie,

$$\pi(dx)P(x,dy) = \min\left(\pi(x)q(x,y), \pi(y)q(y,x)\right)dxdy = \pi(dy)P(y,dx).$$

■

Exemple : noyau de proposition gaussien. Considérons l'exemple caricatural où $E = \mathbb{R}$ et où l'on souhaite appliquer l'algorithme de Metropolis-Hastings pour simuler selon la loi cible $\pi = \mathcal{N}(0,1)$. Pour ce faire, soit $\sigma^2 > 0$ fixé et le noyau de transition gaussien

$$Q_\sigma(x, dy) = q_\sigma(x, y)dy = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-x)^2}{2\sigma^2}} dy.$$

Comme nous l'avons vu, partant de $X_n = x$, l'algorithme commence par proposer

$$Y_{n+1} = X_n + \sigma W_{n+1} = x + \sigma W_{n+1},$$

avec W_{n+1} gaussienne standard indépendante de tout le reste. Si $Y_{n+1} = y$, puisque $q_\sigma(x, y) = q_\sigma(y, x)$, c'est la version Metropolis de l'algorithme :

$$r(x, y) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} = \frac{\pi(y)}{\pi(x)} = e^{-\frac{1}{2}(y^2-x^2)}.$$

Avec des mots, si Y_{n+1} est plus proche de l'origine que X_n (i.e. plus vraisemblable pour π), alors cette transition est systématiquement acceptée, sinon elle l'est seulement avec probabilité $e^{-\frac{1}{2}(y^2-x^2)}$, donc d'autant moins que $|Y_{n+1}|$ est supérieur à $|X_n|$. Or nous avons vu que le pas moyen de la transition proposée est

$$\mathbb{E}[|Y_{n+1} - X_n|] = \sigma\mathbb{E}[|W_{n+1}|] = \sigma\sqrt{\frac{2}{\pi}}.$$

Ceci illustre l'importance cruciale du paramètre de réglage σ dans l'efficacité de l'algorithme :

- si σ est trop petit (e.g. $\sigma = 0.01$, pour forcer le trait), alors $r(X_n, Y_{n+1})$ est proche de 1, donc la transition sera presque toujours acceptée et $X_{n+1} = Y_{n+1}$: la chaîne bouge, ce qui est une bonne chose, mais chaque transition est minuscule et il faudra beaucoup de temps avant de réussir à explorer l'ensemble de la zone d'intérêt pour π, soit en gros l'intervalle $[-3, 3]$;
- si σ est trop grand (e.g. $\sigma = 10$, toujours pour forcer le trait), alors $r(X_n, Y_{n+1})$ a de grandes chances d'être proche de 0 (penser par exemple à la condition initiale $X_0 = 0$), donc la transition sera presque toujours refusée, auquel cas $X_{n+1} = Y_{n+1}$: la chaîne ne bouge pas, ou très rarement, ce qui est tout aussi désastreux.

Interprétation. De façon générale, le noyau de proposition Q doit réaliser un compromis entre le fait de bouger suffisamment pour bien explorer le support de la loi π et ne pas proposer des "pas" trop grands qui risquent d'être systématiquement refusés. Un indicateur à cet égard, aisément implémentable, est donné par le taux moyen de rejet mis à jour au fur et à mesure du déroulement de l'algorithme. Il n'existe pas énormément de résultats théoriques sur le meilleur réglage de ce taux de rejet. Une règle empirique est d'essayer de le conserver autour de 0.5.

Sous des hypothèses raisonnables sur le noyau de transition Q, la chaîne (X_n) a pour unique loi stationnaire π. L'étude théorique dépasse cependant le cadre de ce cours. L'article de Tierney [33] propose une introduction accessible à ce domaine. Afin d'en donner une idée, nous nous contenterons d'énoncer une conséquence du Corollaire 3 qui s'y trouve. Au préalable, définissons de façon générale la distance en variation totale entre deux mesures de probabilités μ et ν sur $\mathbb{R}^d$:

$$\|\nu - \mu\|_{vt} = \sup_{A \in \mathcal{B}(\mathbb{R}^d)} |\nu(A) - \mu(A)|.$$

Si $\mu(dx) = f(x)dx$ et $\nu(dx) = g(x)dx$, on retrouve le lien avec la distance L^1, à savoir

$$\|\nu - \mu\|_{vt} = \frac{1}{2}\int_{\mathbb{R}^d} |g(x) - f(x)|dx = \frac{1}{2}\|g - f\|_{L^1},$$

mais cette interprétation ne tient plus dès lors que, comme dans le cadre Metropolis-Hastings, l'une des mesures est un mélange de lois avec une partie absolument continue et un Dirac.

Proposition 14
Si π est à support compact S, si π et q sont strictement positives et continues sur ce compact, alors le noyau de Metropolis-Hastings est uniformément ergodique, c'est-à-dire qu'il existe des constantes $C > 0$ et $0 \leq r < 1$ telles que

$$\forall n \in \mathbb{N}, \qquad \sup_{x \in S} \|P^n(x, \cdot) - \pi\|_{vt} \leq C r^n.$$

Bien entendu, l'hypothèse très forte de ce résultat est la compacité du support de la loi cible π. Par exemple, il ne s'applique pas à l'exemple jouet/caricatural ci-dessus de la simulation d'une gaussienne standard via un noyau de proposition gaussien... Il existe cependant des résultats plus fins permettant de traiter le cas où π n'est pas supposée à support compact mais seulement à queues sous-exponentielles (cf. par exemple le point (ii) de la Proposition 1.3.33 du polycopié Méthodes de Monte Carlo de Benjamin Jourdain).

3.4.3 Echantillonneur de Gibbs

Dans sa version balayage aléatoire, l'échantillonneur de Gibbs correspond à un cas particulier de Metropolis-Hastings où toutes les transitions sont acceptées ! Il requiert néanmoins beaucoup plus de connaissance sur la loi cible π.

Plaçons-nous par exemple dans le cas d'un espace d'états continu $\mathbb{R}^d$ ou un sous-ensemble de $\mathbb{R}^d$. La densité cible s'écrit $\pi(x) = \pi(x_1, \ldots, x_d)$. Pour tout indice ℓ entre 1 et d, on note

$$x_{-\ell} = (x_1, \ldots, x_{\ell-1}, x_{\ell+1}, \ldots, x_d)$$

le d-uplet correspondant à x privé de sa ℓ-ème composante. On convient alors d'écrire la densité jointe $\pi(x) = \pi(x_\ell, x_{-\ell})$, tandis que $\pi(\cdot|x_{-\ell})$ est la densité conditionnelle correspondant à $(d-1)$ coordonnées gelées et $\pi(x_{-\ell})$ la densité jointe de ces $(d-1)$ coordonnées.

L'hypothèse **cruciale** est la suivante : pour tout ℓ et tout $(d-1)$-uplet $x_{-\ell}$, on sait simuler suivant la densité conditionnelle $\pi(\cdot|x_{-\ell})$. L'échantillonneur de Gibbs à balayage aléatoire fonctionne alors comme suit :

1. partant de $X_n = x$, tirer au hasard uniformément une coordonnée ℓ entre 1 et d ;
2. simuler X'_ℓ selon la loi $\pi(\cdot|x_{-\ell})$;
3. poser $X_{n+1} = (X'_\ell, x_{-\ell})$.

Lemme 7 (Echantillonneur de Gibbs à balayage aléatoire)
L'échantillonneur de Gibbs à balayage aléatoire correspond à un algorithme de Metropolis-Hastings où toutes les transitions sont acceptées, c'est-à-dire $r(x, y) = 1$ à chaque étape. Par conséquent, la chaîne (X_n) est π-réversible et π est invariante.

Preuve. Si l'on reprend les notations de Metropolis-Hastings, l'algorithme précédent revient à proposer $Y = y = (x'_\ell, x_{-\ell})$ partant de $X_n = x = (x_\ell, x_{-\ell})$ avec la densité de transition

$$q(x, y) = \frac{1}{d}\, \pi(x'_\ell | x_{-\ell}) = \frac{1}{d} \times \frac{\pi(x'_\ell, x_{-\ell})}{\pi(x_{-\ell})} = \frac{1}{d} \times \frac{\pi(y)}{\pi(x_{-\ell})}.$$

Réciproquement, la transition de y vers x a donc pour densité

$$q(y, x) = \frac{1}{d} \times \frac{\pi(x)}{\pi(x_{-\ell})}.$$

Ceci étant établi, le rapport de Metropolis-Hastings vaut donc dans ce cas

$$r(x,y) = \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} = 1,$$

et il est donc normal d'accepter toutes les transitions proposées.

■

Remarque : Interprétation en termes de noyaux de transition. Si on note Q_ℓ le noyau de transition correspondant au changement de la coordonnée ℓ, c'est-à-dire permettant de passer de $x = (x_\ell, x_{-\ell})$ à $y = (y_\ell, y_{-\ell}) = (y_\ell, x_{-\ell})$, ce noyau n'a pas de densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$ puisque $(d-1)$ coordonnées sont inchangées. Il s'écrit

$$Q_\ell(x, dy) = \mathbf{1}_{y_{-\ell}=x_{-\ell}} \delta_{x_{-\ell}}(dy_{-\ell}) \pi(y_\ell | x_{-\ell}) dy_\ell.$$

Le calcul précédent montre que, pour tout couple (x, y),

$$\pi(dx)Q_\ell(x, dy) = \pi(dy)Q_\ell(y, dx),$$

les deux membres valant 0 si $y_{-\ell} \neq x_{-\ell}$. Puisque chaque noyau Q_ℓ est choisi avec probabilité $1/d$, le noyau P correspondant à l'échantillonneur de Gibbs à balayage aléatoire est

$$P = \frac{1}{d}(Q_1 + \cdots + Q_d).$$

En d'autres termes, P est un mélange de noyaux de transition. Le noyau P n'est pas à densité : partant de x, on ne peut aller en y que si x et y ne diffèrent que par une coordonnée. Néanmoins, puisque les équations d'équilibre détaillé sont vérifiées par chaque noyau Q_ℓ, c'est encore vrai pour le noyau P : pour tout couple (x, y),

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

Cette méthode est appelée échantillonnage de Gibbs par balayage aléatoire car l'indice ℓ est tiré au hasard à chaque étape. On comprend que ce n'est pas complètement satisfaisant puisque, si la coordonnée ℓ est tirée deux fois de suite, le second tirage ne sert à rien en terme d'évolution de la chaîne.

Ce qu'on appelle plus couramment échantillonnage de Gibbs (sous-entendu : à balayage déterministe ou systématique) est la méthode consistant à tirer les d coordonnées successivement :

1. partant de $X_n = x$,
   pour ℓ allant de 1 à d, simuler X'_ℓ selon la loi $\pi(\cdot | X'_1, \ldots, X'_{\ell-1}, x_{\ell+1}, \ldots, x_d)$;
2. poser $X_{n+1} = X' = (X'_1, \ldots, X'_d)$.

Ainsi, lors d'une étape du Gibbs systématique, i.e. pour le passage de X_n à X_{n+1}, toutes les coordonnées ont été modifiées. A contrario, pour le passage de X_n à X_{n+1} dans le Gibbs à balayage aléatoire, une seule d'entre elles a changé.

Lemme 8 (Echantillonneur de Gibbs)
L'échantillonneur de Gibbs (sous-entendu : à balayage systématique) admet π comme loi invariante.

Preuve. Notons $R(x, dy)$ le noyau de transition de l'échantillonneur de Gibbs à balayage systématique. Il revient à composer les noyaux de transition Q_ℓ introduits ci-dessus, i.e. $R = Q_1 \ldots Q_d$. On veut montrer que $\pi R = \pi$, or on a vu que $\pi Q_\ell = \pi$ pour tout ℓ, donc le résultat est clair.

■

Remarques :

1. C'est bien entendu cette version de l'échantillonneur de Gibbs qui est utilisée en pratique, non celle à balayage aléatoire.
2. Si π a une densité, alors le noyau R de l'échantillonneur de Gibbs (sous-entendu : à balayage systématique) est à densité : pour tout couple $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$, celle-ci vaut
$$R(x, y) = \pi(y_1|x_2, \ldots, x_d)\pi(y_2|y_1, x_3, \ldots, x_d) \cdots \pi(y_d|y_1, \ldots, y_{d-1}).$$
3. Contrairement au cas du balayage aléatoire, la chaîne (X_n) n'est en général pas π-réversible. En dimension 2, il faudrait en effet que pour tous couples (x_1, x_2) et (y_1, y_2),
$$\pi(x_1, x_2)R((x_1, x_2), (y_1, y_2)) = \pi(y_1, y_2)R((y_1, y_2), (x_1, x_2)),$$
c'est-à-dire, d'après l'expression précédente du noyau R,
$$\pi(x_1, x_2)\pi(y_1|x_2)\pi(y_2|y_1) = \pi(y_1, y_2)\pi(x_1|y_2)\pi(x_2|x_1),$$
ou encore
$$\pi(y_1|x_2)\pi(y_2) = \pi(y_1)\pi(y_2|x_1).$$
Cette relation est vérifiée si les composantes de π sont indépendantes, mais n'a aucune raison de l'être en général. Il suffit en effet de voir que le membre de gauche dépend de x_2, contrairement à celui de droite.
4. Comme le montre l'Exercice 3.12, il ne faut surtout pas remettre toutes les coordonnées à jour simultanément, sans quoi on risque de ne plus converger vers la loi cible !

3.5 Exercices

Exercice 3.1 (Modèle de diffusion d'Ehrenfest)
On considère deux urnes A et B, contenant M boules à elles deux, numérotées de 1 à M. A chaque instant, on choisit un numéro $j \in \{1, \ldots, M\}$ de façon équiprobable et on change d'urne à la boule numéro j. L'état X_n de la chaîne est le nombre de boules à l'instant n dans l'urne A.

1. Donner le graphe de transition de la chaîne (X_n).
2. Cette chaîne est-elle irréductible ? apériodique ?
3. Montrer que (X_n) admet pour loi stationnaire π une loi binomiale dont on précisera les paramètres.
4. En fixant par exemple $M = 10$, retrouver π par simulation grâce à une seule réalisation (très longue) de la chaîne. En d'autres termes, illustrer la loi des grands nombres du cours.
5. Toujours pour $M = 10$, retrouver π grâce à plusieurs réalisations de la chaîne. En d'autres termes, illustrer la convergence en loi du cours.
6. Supposons M grand. Approcher la loi π par une loi normale dont on précisera les paramètres. En déduire un intervalle $[m_1, m_2]$ tel qu'asymptotiquement, la chaîne passe 95% de son temps entre ces deux valeurs.
7. Vérifier ce résultat par simulation.

Exercice 3.2 (Bistochasticité et Monopoly)

1. On dit qu'une matrice de transition (ou matrice stochastique) P est bistochastique si la somme de chaque colonne est aussi égale à 1. Soit (X_n) une chaîne de Markov dont la matrice de transition est bistochastique : vérifier que la loi uniforme est une loi stationnaire. Est-elle nécessairement unique ?

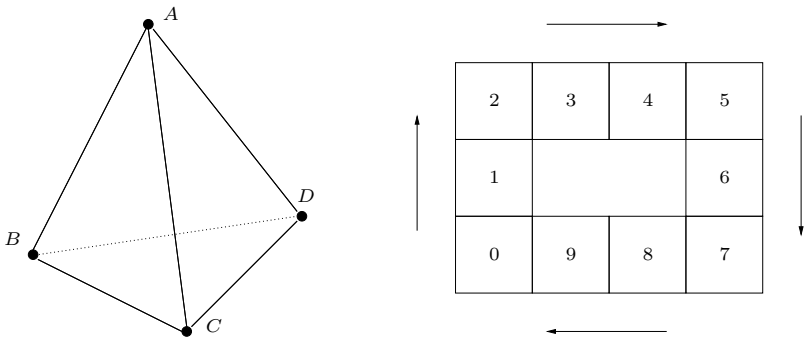

FIGURE 3.2 – Tétraèdre et Monopoly

2. Un jeu du genre Monopoly a dix cases (voir figure 3.2 à droite). On part de la case 0 et on lance un dé équilibré à six faces pour avancer le pion. X_n est la position du pion après le n-ème lancer.
   (a) Déterminer la matrice de transition de la chaîne de Markov (X_n).
   (b) La chaîne est-elle irréductible ? apériodique ?
   (c) Déterminer la (ou les) loi(s) stationnaire(s).

Exercice 3.3 (Le scarabée)
Un scarabée se déplace sur les arêtes d'un tétraèdre régulier (voir figure 3.2 à gauche). Quel que soit le sommet où il se trouve à un instant donné, il choisit au hasard et de façon équiprobable le sommet vers lequel il va se diriger. Il lui faut une unité de temps pour l'atteindre. On suppose de plus que le scarabée se déplace en continu, c'est-à-dire qu'il ne s'arrête jamais en un sommet. X_n est la position du scarabée à l'instant n.

1. Déterminer la matrice de transition de la chaîne de Markov (X_n). Loi(s) stationnaire(s) ?
2. A-t-on convergence en loi de (X_n) ?
3. Illustrer cette convergence par simulation.
4. Le scarabée paye 1€ chaque fois qu'il passe au sommet A, 2€ chaque fois qu'il passe au sommet B, 3€ chaque fois qu'il passe au sommet C, 4€ chaque fois qu'il passe au sommet D. Soit C_n le coût de sa trajectoire jusqu'à l'instant n. Que dire de la convergence de C_n/n ? L'illustrer par simulation.
5. Supposons maintenant qu'en chaque sommet, le scarabée reste sur place avec probabilité 7/10 ou parte vers un des autres sommets de façon équiprobable. Quid de la convergence en loi ?
6. On considère maintenant que les déplacements du scarabée sont régis par la matrice de transition :
$$P = \begin{bmatrix} 0 & 2/3 & 0 & 1/3 \\ 1/3 & 0 & 2/3 & 0 \\ 0 & 1/3 & 0 & 2/3 \\ 2/3 & 0 & 1/3 & 0 \end{bmatrix}.$$
   (a) Vérifier la loi des grands nombres.
   (b) Que dire de la convergence en loi ?
7. Tirer "au hasard" une matrice de transition P à l'aide de la fonction `runif`. Vérifier que la loi des grands nombres et la convergence en loi permettent de trouver un même vecteur probabilité ligne π. Retrouver cette loi d'équilibre grâce à la fonction `eigen` et en utilisant

sa propriété caractéristique : π est un vecteur propre à gauche de P associé à la valeur propre 1.

Exercice 3.4 (Le coup du parapluie)
Un employé lambda, Franz Kafka, se rend chaque matin de son appartement à son bureau et fait le contraire le soir. Il dispose en tout de 3 parapluies, certains chez lui, les autres au bureau. A Prague, il pleut 2 fois sur 3 lorsqu'il fait le trajet, et ce indépendamment du passé. Soit X_n le nombre de parapluies à son domicile durant la nuit.

1. Déterminer la matrice de transition de la chaîne de Markov associée.
2. Quelle est la proportion du temps où Kafka est mouillé en arrivant au bureau le matin ?
3. Généraliser avec n parapluies.

Exercice 3.5 (Metropolis : exemple jouet)
On veut simuler une loi normale centrée réduite grâce à l'algorithme de Metropolis. La loi de proposition est uniforme sur $[-1/2, 1/2]$. Par rapport au cours, nous sommes donc dans un espace d'états continu.

1. L'équivalent de la matrice de transition q_{ij} est le noyau de transition $q(x, y)$ qui correspond à la densité de $Y = x + U$, densité de probabilité d'aller en y sachant qu'on part du point x. Déterminer $q(x, y)$. En déduire $q(y, x)$.
2. Traduire l'algorithme de Metropolis-Hastings dans ce cadre.
3. En partant de X_0 suivant une loi uniforme sur $[-3, 3]$, simuler alors une chaîne de Markov $(X_0, \ldots, X_n)$ par l'algorithme de Metropolis-Hastings.
4. Pour la réalisation de cette chaîne, déterminer la moyenne empirique, l'écart-type empirique, représenter un histogramme des X_i auquel on superposera la densité de la loi normale centrée réduite.
5. Grâce à la fonction `unique`, donner le taux d'acceptation du Metropolis-Hastings.
6. Observer ce qui se passe lorsque la condition initiale est $X_0 = 10$.
7. Revenons à $X_0 \sim \mathcal{U}_{[-3,3]}$. Plutôt qu'une loi de proposition uniforme sur $[-1/2, 1/2]$, considérer une loi uniforme sur $[-10, 10]$. Qu'observez-vous ? Et avec une loi uniforme sur $[-0.1, 0.1]$?

Exercice 3.6 (Metropolis vs Rejet)
On considère sur $\mathbb{R}^2$ la densité $g(u, v) = c \times f(u, v)$, où c est une constante de normalisation et

$$f(u, v) = (\cos u)^2 (\sin v)^2 \exp\left(-0.05(u^2 + v^2)\right).$$

1. Sur $[-5, 5] \times [-5, 5]$, donner une représentation 3d de f (par exemple par `persp(u,v,z)`) et une représentation 2d par lignes de niveau de f (par exemple par `image(u,v,z)`).
2. On veut simuler selon la densité g en utilisant l'algorithme de Metropolis-Hastings. Partant de $x = (u, v)$, on considère le noyau de transition $q(x, x')$ qui correspond à la densité de $X' = x + \sigma W$, où $\sigma > 0$ est un paramètre de réglage et $W \sim \mathcal{N}([0, 0]', I_2)$, loi gaussienne centrée réduite dans $\mathbb{R}^2$. Expliquer pourquoi $q(x, x') = q(x', x)$, c'est-à-dire qu'on est dans le cadre de l'algorithme de Metropolis.
3. Implémenter l'algorithme de Metropolis. On utilisera le noyau de transition $q(x, x')$ ci-dessus avec $\sigma = 1$. On pourra par exemple prendre comme initialisation $X_1 = (0, 0)$ et considérer une chaîne $(X_k)_{1 \leq k \leq n}$ de longueur $n = 10^4$.
4. Sur le graphe de f par lignes de niveau (via `image(u,v,z)`), superposer les points de la chaîne par exemple par la commande `points(X,pch=3,cex=.1)`. Faire la même chose avec σ grand, par exemple $\sigma = 10$, et commenter. Idem avec σ petit, par exemple $\sigma = 0.1$.

5. Proposer une méthode de rejet pour simuler suivant g à partir d'une loi instrumentale gaussienne. Comme en question précédente, superposer un échantillon de grande taille simulé par cette méthode aux niveaux de la fonction f pour vérifier visuellement le bon fonctionnement de l'algorithme.
6. Des deux méthodes, laquelle vous semble préférable ?

Exercice 3.7 (Metropolis-Hastings : méthode de Roberts et Rosenthal (1998))
On veut maintenant simuler un mélange équiprobable de deux lois normales réduites, centrées respectivement en $m = 3$ et $-m = -3$.

1. Rappeler la densité f correspondante. Quelle est sa moyenne ? déterminer son écart-type, à la main ou grâce à la fonction `integrate`.
2. Simuler une chaîne de Markov de densité d'équilibre f par la même méthode que dans l'exercice précédent, mais avec une loi de proposition gaussienne centrée d'écart-type $\sigma = 1$. Donner la moyenne empirique, l'écart-type empirique, représenter un histogramme des X_i auquel on superposera la densité f.
3. Observer ce qui se passe lorsque $X_0 = m$. D'où vient le problème ?
4. Essayer d'améliorer les choses en modifiant l'écart-type σ.
5. On ne considère plus un noyau de transition symétrique. Partant de x, la proposition est désormais
$$Y = x + \frac{\sigma^2}{2} \times \frac{f'(x)}{f(x)} + \sigma W,$$
où W suit une loi normale centrée réduite.
   (a) Quelles transitions sont favorisées par ce noyau de transition ?
   (b) Expliciter le rapport de Metropolis-Hastings $r(x, y)$.
   (c) Simuler la chaîne de Markov correspondant à cet algorithme de Metropolis-Hastings.

Exercice 3.8 (Metropolis indépendant et optimisation)
On veut trouver le maximum sur $[0, 1]$ de la fonction
$$f_u(x) = (\cos(50x) + \sin(20x))^2.$$

1. Représenter f_u et déterminer ce maximum numériquement grâce à la fonction `optimize`.
2. Une autre idée, plus artisanale, est de construire une chaîne de Markov ayant pour loi stationnaire la densité f proportionnelle à f_u. Quelle est l'idée sous-jacente ?
3. Dans un premier temps, prenons le noyau de transition correspondant à $Y = x+U \pmod 1$ où $U \sim \mathcal{U}_{[0,1]}$. Quelle est la loi de Y ? Construire l'algorithme de Metropolis correspondant : on parle de Metropolis indépendant, pourquoi ?
4. Donner une estimation du maximum de f grâce à une longue réalisation de la chaîne.
5. Estimer de même le minimum en considérant la fonction $\exp(-f_u(x))$.

Exercice 3.9 (Recuit simulé)
On veut estimer par recuit simulé le minimum sur $\mathbb{R}$ de la fonction $V(x) = x^2(2 + (\sin(100x))^2)$.

1. Représenter $V(x)$ pour x variant entre -2 et 2. Quel est le minimum de V sur $\mathbb{R}$?
2. Pour une température $T > 0$ fixé, on considère sur $\mathbb{R}$ la densité $f_T(x) = \frac{1}{Z_T} \exp(-V(x)/T)$ où
$$Z_T = \int_{\mathbb{R}} \exp(-V(x)/T)dx.$$
En quel(s) point(s) f_T admet-elle son maximum sur $\mathbb{R}$?

3. Sur une même figure, pour x variant entre -1 et 1, représenter à gauche $x \mapsto \exp(-V(x)/T)$ pour $T = 10$, et à droite pour $T = 0.1$.
4. On note U une loi uniforme sur $[-1/2, 1/2]$. Partant d'un point x, on considère le noyau de transition $q(x, y)$ qui correspond à la densité de $Y = x + U$. Déterminer $q(x, y)$. En déduire $q(y, x)$ et le rapport de Metropolis-Hastings $r(x, y)$ pour une température $T > 0$ fixée.
5. Implémenter l'algorithme du recuit simulé pour le noyau de transition q et le schéma de température $T_k = 1/(1 + \log k)$ pour k allant de 1 à $n = 1000$. On pourra prendre comme initialisation $X_1 \sim \mathcal{U}_{[-10,10]}$. On donnera en sortie la valeur $V(X_n)$ ainsi que le minimum observé sur toutes les valeurs $V(X_k)$ pour $k \in \{1, \ldots, n\}$.

Exercice 3.10 (Le voyageur de commerce)
Un commercial doit passer par N villes et revenir à son point de départ, il se déplace en avion et il y a des vols directs entre toutes les villes. Le but est de trouver le trajet optimal, c'est-à-dire la distance minimale à parcourir.

1. Pourquoi la solution exacte est-elle difficile à trouver quand N n'est pas petit ?
2. Simuler $N = 10$ points $(M_1, \ldots, M_N)$ uniformément dans le carré $[0, 1] \times [0, 1]$. Les représenter, ainsi que les trajets $(M_1, M_2), \ldots, (M_{N-1}, M_N), (M_N, M_1)$.
3. Construire la matrice $N \times N$ des distances entre toutes les villes.
4. On considère le groupe $\mathcal{S}_N$ des permutations de l'ensemble $\{1, \ldots, N\}$. A chaque permutation $\sigma = (\sigma(1), \ldots, \sigma(N))$ est associée la distance totale parcourue par le voyageur de commerce, c'est-à-dire en notant $d(A, B)$ la distance euclidienne entre A et B :

$$D_\sigma = \sum_{\ell=1}^{N} d(M_{\sigma(\ell)}, M_{\sigma(\ell+1)}),$$

   avec la convention $\sigma(N + 1) = \sigma(1)$. Dans ce qui précède, on avait donc $\sigma = (1, 2, \ldots, N - 1, N)$. Reformuler le problème d'optimisation avec ces notations.
5. La permutation σ étant donnée, on note $\sigma^{\ell,\ell'}$ la permutation consistant à inverser l'ordre de visite des villes ℓ et ℓ', autrement dit :

$$\sigma = (\sigma(1), \ldots, \sigma(\ell), \ldots, \sigma(\ell'), \ldots, \sigma(N)) \implies \sigma^{\ell,\ell'} = (\sigma(1), \ldots, \sigma(\ell'), \ldots, \sigma(\ell), \ldots, \sigma(N)).$$

   Partant de σ, comment tirer uniformément parmi les $\sigma^{\ell,\ell'}$?
6. Implémenter l'algorithme de Metropolis correspondant aux transitions $Q(\sigma, \sigma^{\ell,\ell'})$ ci-dessus et de loi stationnaire proportionnelle à $\exp(-D_\sigma)$ sur l'espace d'états $\mathcal{S}_N$.
7. Déterminer le meilleur trajet obtenu par cette méthode et le représenter.
8. On adopte la technique du recuit simulé : par rapport à ce qui précède, il suffit donc de tenir compte de la température T_n à chaque étape. Pour comparer les schémas de température, on utilisera la fonction `set.seed` (permettant de fixer en début de simulation la graine du générateur aléatoire) et on comparera les performances sur un nombre représentatif de simulations, par exemple une centaine. Le nombre N de villes ne devra pas être pris trop grand, contrairement à la longueur n de la chaîne...
   (a) Implémenter le recuit simulé avec un schéma de température en $T_n = 1/\log n$. Comparer à ce qui était obtenu avec la température fixe $T = 1$ de la question 7.
   (b) Même question avec un schéma de température à décroissance polynomiale, c'est-à-dire de type $T_n = n^{-\gamma}$, pour plusieurs valeurs de γ.

Exercice 3.11 (Refroidissement trop rapide)
Cet exercice est tiré du polycopié de Thierry Bodineau [2] et illustre sur un exemple élémentaire la raison pour laquelle un refroidissement trop rapide nuit à la convergence du recuit simulé vers le minimum global. Soit $E = \{1, 2, 3\}$ et la fonction $V : E \to \mathbb{R}$ définie par $V(1) = 0$, $V(2) = 1$ et $V(3) = -1$. L'exploration se fait comme suit : partant des états 1 ou 3, on propose systématiquement d'aller en 2, tandis que partant de 2, on propose d'aller en 1 ou 3 de façon équiprobable. On considère l'algorithme du recuit simulé avec un schéma de température noté (T_n).

1. Minimum local de V ? Minimum global ? Donner le graphe de transition de la matrice de transition Q.
2. On part du point $X_0 = 1$ et on veut contrôler $\mathbb{P}(\exists k \leq n,\ X_k = 3)$. Majorer cette quantité en fonction de la probabilité que la chaîne ne bouge pas de sa position initiale.
3. Grâce à la propriété de Markov, déterminer $\mathbb{P}(\forall k \leq n,\ X_k = 1)$ en fonction des T_k.
4. On rappelle que si les x_k sont des nombres entre 0 et 1, alors $\prod_{k=0}^{n-1}(1 - x_k) \geq 1 - \sum_{k=0}^{n-1} x_k$. En déduire une majoration de $\mathbb{P}(\exists n,\ X_n = 3)$ par la somme d'une série dépendant du schéma de température. Que se passe-t-il si $T_n \leq c/\log n$ avec c trop petit ?

Exercice 3.12 (Gibbs simultané = Gibbs vérolé)
On considère sur $E = \{(0,0), (0,1), (1,0), (1,1)\}$ la distribution $\pi = [2/5, 1/5, 1/5, 1/5]$. Sur cet espace d'états, on considère la chaîne de Markov (X_n) régie par les transitions suivantes : partant de $X = (x_1, x_2)$, le point suivant est $X' = (x_1', x_2')$ où x_1' est tiré suivant la loi conditionnelle $\pi(\cdot|x_2)$ et, indépendamment, x_2' est tiré suivant la loi conditionnelle $\pi(\cdot|x_1)$. Cette façon de se déplacer dans l'espace d'états s'apparente donc à un échantillonneur de Gibbs où toutes les coordonnées seraient remises à jour simultanément.

1. Donner la matrice de transition P de la chaîne de Markov (X_n). Cette chaîne est-elle irréductible ? apériodique ?
2. Montrer que que l'algorithme proposé ne converge pas vers ce qu'on veut.
3. Donner la matrice de transition P_a de l'échantillonneur de Gibbs par balayage aléatoire. Vérifier que π est stationnaire pour P_a. La chaîne est-elle réversible ?
4. Mêmes questions pour la matrice P_d de l'échantillonneur de Gibbs par balayage déterministe.

Exercice 3.13 (Somme d'exponentielles contrainte)
Soit $X_1, \ldots, X_d$ des variables exponentielles indépendantes de paramètres $\lambda_1, \ldots, \lambda_d$ donnés. Soit $S = X_1 + \cdots + X_d$ leur somme et $m > 0$ une constante fixée.

1. Donner la densité f_m du vecteur aléatoire $(X_1, \ldots, X_d)$ sachant $S > m$.
2. Soit $(x_2, \ldots, x_d)$ fixé. Comment simuler X_1 sachant que $X_1 + x_2 + \cdots + x_d > m$?
3. En déduire un échantillonneur de Gibbs pour simuler un échantillon $(X_1, \ldots, X_d)$ conditionnellement au fait que $S > m$.

Exercice 3.14 (Cas d'école pour Gibbs)
On considère la densité

$$f(x, y) = C\ \exp\left(-\frac{y^2}{2} - \frac{x^2(1 + y + y^2)}{2}\right).$$

1. Loi de X sachant $Y = y$? Loi de Y sachant $X = x$?
2. En déduire un échantillonneur de Gibbs de loi cible f.

Exercice 3.15 (Simulation d'un couple)
Soit (X, Y) un couple aléatoire de densité jointe

$$f(x, y) = e^{-y}\mathbf{1}_{0 \leq x \leq y}.$$

1. Déterminer la densité marginale de X, notée $f(x)$. Quelle loi reconnaissez-vous ?
2. Sachant $X = x \geq 0$, déterminer la densité conditionnelle $f(y|x)$. Quelle loi reconnaissez-vous ?
3. En déduire une méthode pour simuler une réalisation du couple aléatoire (X, Y). L'implémenter pour simuler un échantillon de couples $(X_1, Y_1), \ldots, (X_n, Y_n)$ de taille $n = 1000$. Représenter le nuage de points ainsi obtenu.
4. Sachant $Y = y \geq 0$, déterminer la densité conditionnelle $f(x|y)$. Quelle loi reconnaissez-vous ?
5. En partant par exemple du point $(x_0, y_0) = (0, 1)$, implémenter un échantillonneur de Gibbs à balayage déterministe pour obtenir un échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ de taille $n = 1000$ de loi cible f. Représenter le nuage de points ainsi obtenu.
6. Des deux méthodes proposées, laquelle choisiriez-vous pour simuler selon la densité $f(x, y)$?

Exercice 3.16 (Balayage aléatoire vs balayage déterministe)
On considère un espace d'états E fini de cardinal M, une densité cible $\pi = [\pi_1, \ldots, \pi_M]$ avec tous les π_i strictement positifs, et une matrice de transition $Q = [q_{ij}]$ avec tous les q_{ij} strictement positifs.

1. Dans l'algorithme de Metropolis-Hastings, quelle est la probabilité α_{ij} qu'une loi uniforme soit inférieure à r_{ij} ?
2. En déduire le coefficient p_{ij} de la matrice P correspondant aux transitions de la chaîne (X_n) construite par l'algorithme (attention aux termes diagonaux).
3. Vérifier que P est réversible pour π. En déduire que π est stationnaire pour P.
4. Vérifier que c'est encore vrai si au lieu de prendre $\alpha(r) = \min(1, r)$ comme ci-dessus, on considère n'importe quelle fonction $\alpha :]0, \infty[\to]0, 1]$ telle que $\alpha(r) = r \times \alpha(1/r)$.
5. Reformuler tout ce qui précède dans le cadre d'un espace d'états continu.
6. Dans l'échantillonneur de Gibbs, notons $P_1, \ldots, P_d$ les noyaux de transition en jeu, c'est-à-dire que la transition de x vers x' peut avoir lieu seulement si x et x' ne diffèrent que par une coordonnée, disons ℓ, c'est-à-dire $x_{-\ell} = x'_{-\ell}$, auquel cas ceci arrive avec la densité de probabilité $P_\ell(x, x') = f(x'_\ell | x_{-\ell})$.
   (a) Vérifier que le noyau P_ℓ est réversible pour la densité f. En déduire que f est stationnaire pour le noyau P_ℓ.
   (b) Pour le balayage aléatoire, exprimer le noyau de transition P en fonction des P_ℓ. En déduire que le noyau P est réversible pour la densité f.
   (c) Pour le balayage déterministe, exprimer le noyau de transition P en fonction des P_ℓ. En déduire que f est stationnaire pour P. La chaîne est-elle réversible ?

Exercice 3.17 (Modèle d'Ising)
Soit N entier naturel fixé. On considère le carré $\mathcal{C} = \{(i, j),\ 1 \leq i, j \leq N\}$, chaque site (i, j) ne pouvant prendre que deux valeurs $S_{ij} = \pm 1$ (son spin). A une configuration de spins $S = (S_{ij})_{1 \leq i,j \leq N}$ donnée est alors associée l'énergie

$$V(S) = -\sum_{(i,j) \sim (k,\ell)} S_{ij} S_{k\ell} \qquad \text{avec} \qquad (i, j) \sim (k, \ell) \ \Leftrightarrow\ |i - k| + |j - \ell| = 1.$$

Autrement dit, dans la somme définissant V interviennent toutes les paires de sites voisins dans le carré $\mathcal{C}$.

1. Quelles sont les configurations minimisant l'énergie ?

2. Pour tenir compte de la température $T > 0$, on associe à V la mesure de Gibbs

$$\mu_T(S) = \frac{1}{Z_T} \exp\left(-\frac{1}{T}V(S)\right),$$

où Z_T est la constante de normalisation. Que dire de cette mesure de probabilité dans les deux situations extrêmes où $T \to 0$ et où $T \to +\infty$?

3. On considère T fixé et on veut simuler S suivant la mesure μ_T. Donner en fonction de N le nombre de configurations $S = (S_{ij})_{1\leq i,j\leq N}$ possibles.

4. On applique la méthode de Metropolis-Hastings pour simuler suivant μ_T. Pour cela, partant d'une configuration $S = (S_{ij})_{1\leq i,j\leq N}$, on commence par tirer un site (i_0, j_0) au hasard (uniforme) et on propose de changer $S_{i_0j_0}$ en $-S_{i_0j_0}$. On note $S^{(i_0,j_0)}$ cette nouvelle configuration où seul un spin a changé de signe. Que vaut $Q(S, S^{(i_0,j_0)})$? Et $Q(S^{(i_0,j_0)}, S)$? Calculer le rapport de Metropolis-Hastings $r(S, S^{(i_0,j_0)})$.

5. Expliquer brièvement pourquoi la chaîne de Markov ainsi construite est irréductible et apériodique.

6. Pour $N = 40$ et $T = 0.1$, implémenter l'algorithme de Metropolis-Hastings permettant de simuler une chaîne de mesure stationnaire μ_T. On ne demande pas de garder tout l'historique de la chaîne, mais seulement la mise à jour de $S = (S_{ij})_{1\leq i,j\leq N}$. Représenter la configuration S obtenue après 10^4 étapes via la commande `image(S)`.

7. Faire de même pour $T = 10$. Interpréter la différence entre les deux figures obtenues.

Exercice 3.18 (Débruitage d'image)
Soit N entier naturel fixé. On considère la grille $\mathcal{C} = \{(i, j),\ 1 \leq i, j \leq N\}$, chaque pixel (i, j) ne pouvant prendre que deux valeurs $S_{ij} = \pm 1$ (blanc pour -1, noir pour +1). L'image initiale $\Sigma = (\sigma_{ij})_{1\leq i,j\leq N}$, supposée connue, va être modifiée en une nouvelle image $S = (S_{ij})_{1\leq i,j\leq N}$ visant à minimiser l'énergie

$$V(S) = \alpha \sum_{1\leq i,j\leq N} (S_{ij}-\sigma_{ij})^2 - \beta \sum_{(i,j)\sim(k,\ell)} S_{ij}S_{k\ell} \qquad \text{avec} \qquad (i, j) \sim (k, \ell) \Leftrightarrow |i-k|+|j-\ell| = 1,$$

où α et β sont deux constantes positives. Dans la seconde somme définissant V interviennent donc toutes les paires de sites voisins dans le carré $\mathcal{C}$.

1. Si $\alpha > 0$ et $\beta = 0$, quelle(s) configuration(s) minimise(nt) V ? Même question si $\alpha = 0$ et $\beta > 0$.

2. On suppose dans toute la suite que α et β sont tous deux strictement positifs fixés et permettent d'effectuer un compromis entre deux contraintes : l'attache à l'image initiale Σ et l'obtention de contours nets (élimination du bruit). On veut alors minimiser V, en supposant ce minimum atteint pour une unique configuration $S^\star = (S^\star_{ij})_{1\leq i,j\leq N}$. On introduit pour cela un paramètre de température $T > 0$ et on associe à V la mesure de Gibbs

$$\mu_T(S) = \frac{1}{Z_T} \exp\left(-\frac{1}{T}V(S)\right),$$

où Z_T est la constante de normalisation. Que dire de cette mesure de probabilité dans les deux situations extrêmes où $T \to 0$ et où $T \to +\infty$?

3. Soit $T > 0$ fixé. Comme dans le modèle d'Ising, on applique la méthode de Metropolis-Hastings pour simuler suivant μ_T. Pour cela, partant d'une configuration $S = (S_{ij})_{1\leq i,j\leq N}$, on commence par tirer un site (i_0, j_0) au hasard (uniforme) et on propose de changer $S_{i_0j_0}$ en $-S_{i_0j_0}$. On note $S^{(i_0,j_0)}$ cette nouvelle configuration où seul un pixel a changé. Que vaut $Q(S, S^{(i_0,j_0)})$? Et $Q(S^{(i_0,j_0)}, S)$? Expliciter le rapport de Metropolis-Hastings $r(S, S^{(i_0,j_0)})$.

4. On veut appliquer un recuit simulé pour minimiser V. Comme paramètres, on pourra prendre $N = 40$, $\alpha = 1/3$, $\beta = 2/3$ et le schéma de température $T_k = 1/k$ pour k allant de 1 à $n = 10^4$. L'image à retrouver est Σ_0 dont tous les pixels sont à -1 sauf le carré central de côté 20 dont tous les pixels sont à 1. L'image initiale est sa version bruitée $\Sigma = (\sigma_{ij})_{1\leq i,j\leq N}$ où, partant de Σ_0, chaque pixel est changé de signe avec probabilité 1/4, indépendamment les uns des autres. Pour l'implémentation du recuit simulé, on pourra s'inspirer du code vu pour le modèle d'Ising.

Exercice 3.19 (Echantillonnage par tranches)
On considère la fonction $f(u,x) = \mathbf{1}_{0<u<\frac{1}{2}\exp(-\sqrt{x})}$.

1. Grâce à un changement de variable, montrer que $f(x) = \frac{1}{2}\exp(-\sqrt{x})\mathbf{1}_{x>0}$ est une densité sur $\mathbb{R}$. En déduire que $f(u,x)$ est une densité sur $\mathbb{R}^2$.
2. Montrer que les densités conditionnelles $f(u|x)$ et $f(x|u)$ sont celles de lois uniformes que l'on précisera.
3. En partant du point $(U_1, X_1) = (1/(4e), 1)$, implémenter un échantillonneur de Gibbs à balayage déterministe pour obtenir un échantillon $(U_1, X_1), \ldots, (U_n, X_n)$ de taille $n = 1000$ et de loi cible $f(u,x)$.
4. Sur un même graphe, représenter la densité $f(x)$ et un estimateur de la densité obtenu à partir de l'échantillon $X_1, \ldots, X_n$.

Exercice 3.20 (Euclidean matching)
Pour $N \in \mathbb{N}^*$ fixé, on considère 2 ensembles de N points $(A_1, \ldots, A_N)$ et $(B_1, \ldots, B_N)$ dans le carré $[0,1] \times [0,1]$. On veut relier chaque point du premier ensemble à exactement un point du second ensemble de façon à minimiser la somme des longueurs des segments $[A_i, B_j]$. Autrement dit, on cherche σ dans $\mathcal{S}_N$, ensemble des permutations de $\{1, \ldots, N\}$, qui minimise la quantité

$$V(\sigma) = \sum_{i=1}^{N} d(A_i, B_{\sigma(i)}),$$

où d représente la distance euclidienne usuelle dans le plan.

1. Pour $N = 10$, simuler des points $(A_1, \ldots, A_N)$ et $(B_1, \ldots, B_N)$ uniformément dans le carré $[0,1] \times [0,1]$.
2. Construire la matrice D de taille $N \times N$ des distances entre tous les couples de points (A_i, B_j), c'est-à-dire que $D_{i,j} = d(A_i, B_j)$.
3. Coder une fonction **V** qui prend comme paramètres d'entrée la matrice D et une permutation σ et renvoie la valeur $V(\sigma)$.
4. La permutation σ étant donnée, on note $\sigma^{\ell,\ell'}$ la permutation consistant à inverser les appariements $(A_\ell, B_{\sigma(\ell)})$ et $(A_{\ell'}, B_{\sigma(\ell')})$ en $(A_\ell, B_{\sigma(\ell')})$ et $(A_{\ell'}, B_{\sigma(\ell)})$. Autrement dit :

   $$\sigma = (\sigma(1), \ldots, \sigma(\ell), \ldots, \sigma(\ell'), \ldots, \sigma(N)) \implies \sigma^{\ell,\ell'} = (\sigma(1), \ldots, \sigma(\ell'), \ldots, \sigma(\ell), \ldots, \sigma(N)).$$

   Partant de σ, comment tirer uniformément parmi les $\sigma^{\ell,\ell'}$?
5. Pour minimiser V, on adopte la technique du recuit simulé avec le schéma de température $T_n = 1/(1+\log n)$ et des transitions $Q(\sigma, \sigma^{\ell,\ell'})$ comme ci-dessus. Implémenter cet algorithme pour une chaîne de longueur $n = 10^3$. On pourra s'inspirer du code du voyageur de commerce et prendre comme condition initiale $\sigma = (1, \ldots, N)$.
6. Déterminer le meilleur $\sigma^\star$ obtenu. Sur un même graphique, représenter à gauche les couplages initiaux, c'est-à-dire les segments $(A_1, B_1), \ldots, (A_N, B_N)$, et à droite les couplages optimaux, c'est-à-dire les segments $(A_1, B_{\sigma^\star(1)}), \ldots, (A_N, B_{\sigma^\star(N)})$. En ggplot, on pourra utiliser la fonction **geom_segment**.

Exercice 3.21 (Statistique bayésienne)
Dans un cadre bayésien, on considère que la loi a priori sur θ est une loi uniforme sur $[0,1]$ et que, sachant θ, les variables $(X_i)_{1\leq i\leq N}$ sont i.i.d. selon un mélange pondéré par θ de deux gaussiennes réduites de moyennes respectives $+1$ et -1 (densités respectives ϕ et ψ), c'est-à-dire

$$f(x_1|\theta) = \theta \times \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_1-1)^2}{2}} + (1-\theta) \times \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_1+1)^2}{2}} = \theta \times \phi(x_1) + (1-\theta) \times \psi(x_1).$$

1. Donner la formule de la loi a posteriori $\pi(\theta|\mathbf{x}) = \pi(\theta|x_1,\ldots,x_N)$ et l'estimateur de Bayes pour la perte quadratique, i.e. la moyenne a posteriori $\hat{\theta}_N = \hat{\theta}_N(\mathbf{X})$.
2. Pour $\theta^* = 3/4$ et $N = 100$, générer $X_1,\ldots,X_N$ selon le mélange ci-dessus pondéré par θ^* pour obtenir une réalisation $\mathbf{x} = (x_1,\ldots,x_N)$.
3. En déduire un estimateur Monte-Carlo $\hat{\theta}_N^n(\mathbf{x})$ de la réalisation $\hat{\theta}_N(\mathbf{x})$ de l'estimateur de Bayes (moyenne a posteriori). L'implémenter avec par exemple $n = 500$.
4. Toujours pour la même réalisation $\mathbf{x} = (x_1,\ldots,x_N)$, on souhaite générer un échantillon suivant la loi a posteriori $\pi(\theta|\mathbf{x})$. On adopte pour cela la méthode de Metropolis-Hastings avec comme noyau de proposition $q(\theta,\theta') = \mathbf{1}_{[0,1]}(\theta')$, c'est-à-dire que la loi de proposition est tout simplement une loi uniforme sur $[0,1]$. Que vaut le rapport de Metropolis-Hastings $r(\theta,\theta')$? Avec la condition initiale $\theta_1 = 1/2$, implémenter l'algorithme pour une chaîne de longueur $n = 10^4$ et représenter un estimateur de la densité de l'échantillon $(\theta_1,\ldots,\theta_n)$. Donner le taux global d'acceptation sur l'ensemble des mutations proposées.
5. Mêmes questions en considérant le noyau de transition correspondant à $\theta' = \theta + U/\sqrt{N}$ (modulo 1), avec U uniforme sur $[-1,1]$. Indication : on pourra utiliser %% pour obtenir le modulo 1 en R : $1.2\%\%1 = 0.2$ et $-0.1\%\%1 = 0.9$.

Exercice 3.22 (Champ libre gaussien)
Soit $N \in \mathbb{N}^\star$ fixé. On considère le carré $\mathcal{C} = \{(i,j),\ 1 \leq i,j \leq N\}$, chaque site (i,j) pouvant prendre une valeur réelle x_{ij}. A une configuration $x = (x_{ij})_{1\leq i,j\leq N}$ donnée est alors associée l'énergie

$$V(x) = \sum_{(i,j)\sim(k,\ell)} (x_{ij} - x_{k\ell})^2 \qquad \text{avec} \qquad (i,j) \sim (k,\ell) \ \Leftrightarrow\ |i-k| + |j-\ell| = 1.$$

On adopte la convention selon laquelle le carré $\mathcal{C}$ est entouré de 0, ainsi tout site $(i,j)_{1\leq i,j\leq N}$ a exactement 4 voisins. A une température $T > 0$, on associe la mesure de Gibbs

$$\mu_T(x) = \frac{1}{Z_T} \exp\left(-\frac{1}{T} V(x)\right),$$

où Z_T est la constante de normalisation. On veut simuler X suivant la mesure μ_T.

1. Pour un site (i_0,j_0) fixé et une configuration x donnée, on note $x_{-(i_0,j_0)}$ le vecteur x privé de la coordonnée $x_{i_0j_0}$. Quelle est la loi conditionnelle $\mu_T(x_{i_0j_0}|x_{-(i_0,j_0)})$? On reconnaîtra une loi classique.
2. Pour $N = 40$ et $T = 10^{-3}$, implémenter un échantillonneur de Gibbs à balayage aléatoire pour simuler selon μ_T. Via la commande **image(X)**, représenter la configuration initiale, où chaque site est i.i.d. selon une loi normale centrée réduite, et la configuration finale obtenue après 10^5 étapes. On pourra s'inspirer du code pour le modèle d'Ising.
3. Faire de même pour $T = 10^3$. Interpréter la différence entre les deux figures obtenues.
4. Implémenter un échantillonneur de Gibbs à balayage déterministe, c'est-à-dire en balayant intégralement le carré $\mathcal{C}$ du point $(i,j) = (1,1)$ au point $(i,j) = (N,N)$ un certain nombre n de fois (par exemple $n = 100$).

Exercice 3.23 (Maximum de vraisemblance)
Soit $\theta \in \mathbb{R}$ et la loi de Cauchy translatée de θ, c'est-à-dire de densité

$$f_\theta(x) = 1/(\pi(1 + (x - \theta)^2)).$$

1. Si $X_1, \cdots, X_N$ sont i.i.d. selon f_θ, donner la vraisemblance $L_N(\theta) = L_N(\theta; (X_1, \cdots, X_N))$ puis la log-vraisemblance $\ell_N(\theta)$.
2. Pour $N = 10$ et $\theta^* = 0$, simuler $X_1, \cdots, X_N$ i.i.d. selon f_{θ^*}. Représenter la log-vraisemblance $\theta \mapsto \ell_N(\theta)$ associée à cet échantillon pour $\theta \in [X_{(1)} - 10; X_{(N)} + 10]$, où comme d'habitude $X_{(1)} = \min_{1 \leq j \leq N} X_j$ et $X_{(N)} = \max_{1 \leq j \leq N} X_j$.
3. Toujours pour le même échantillon, on veut maintenant déterminer l'estimateur du maximum de vraisemblance par recuit simulé. On pose $V(\theta) = -\ell_N(\theta)$ et à tout $T > 0$ on associe la densité
$$h_T(\theta) = \frac{1}{Z_T} \exp\left(-\frac{1}{T} V(\theta)\right),$$
où Z_T est la constante de normalisation. Partant d'un point θ, on considère la proposition $Y = \theta + \mathcal{N}(0, \sigma^2)$.
   (a) Ecrire le noyau de proposition $q(\theta, y)$ associé et en déduire le rapport de Metropolis-Hastings $r(\theta, y)$.
   (b) Implémenter l'algorithme du recuit simulé pour la condition initiale $\theta_1 = 1$, la variance $\sigma^2 = 1$ et le schéma de température $T_n = 1/n$ avec $n \in \{1, \cdots, 100\}$.
   (c) Grâce au code de la fonction précédente, représenter $V(\theta_n)$ en fonction de n. Donner en particulier la valeur θ_{n^*} maximisant la vraisemblance. Vérifier que ceci est cohérent avec le graphe de la question 2.

3.6 Corrigés

Voir la page du cours.

Bibliographie

[1] Søren Asmussen and Peter W. Glynn. *Stochastic simulation : algorithms and analysis*, volume 57 of *Stochastic Modelling and Applied Probability*. Springer, New York, 2007.

[2] Thierry Bodineau. *Modélisation de phénomènes aléatoires*. Format électronique, 2015.

[3] Raphaël Cerf. The dynamics of mutation-selection algorithms with large population sizes. *Ann. Inst. H. Poincaré Probab. Statist.*, 32(4) :455–508, 1996.

[4] Raphaël Cerf. A new genetic algorithm. *Ann. Appl. Probab.*, 6(3) :778–817, 1996.

[5] Raphaël Cerf. Asymptotic convergence of genetic algorithms. *Adv. in Appl. Probab.*, 30(2) :521–550, 1998.

[6] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. John Wiley & Sons Inc., 1991.

[7] Bernard Delyon. *Simulation et Modélisation*. Format électronique, 2014.

[8] Luc Devroye. *Nonuniform random variate generation*. Springer-Verlag, New York, 1986.

[9] Marie Duflo. *Algorithmes stochastiques*. Springer, 1996.

[10] Rick Durrett. *Essentials of Stochastic Processes*. Springer Texts in Statistics. Springer-Verlag, New York, 1999.

[11] Jean-François Delmas et Benjamin Jourdain. *Modèles aléatoires*. Springer, 2006.

[12] Pierre Del Moral et Christelle Vergé. *Modèles et méthodes stochastiques*. Springer, 2014.

[13] Bernard Bercu et Djalil Chafaï. *Modélisation stochastique et simulation*. Dunod, 2007.

[14] Michel Benaïm et Nicole El Karoui. *Promenade aléatoire*. Editions de l'Ecole Polytechnique, 2004.

[15] Jean Jacod et Philip Protter. *L'essentiel en théorie des probabilités*. Cassini, 2003.

[16] Nathalie Bartoli et Pierre Del Moral. *Simulation et algorithmes stochastiques*. Cépaduès-Editions, 2001.

[17] Olivier François. Global optimization with exploration/selection algorithms and simulated annealing. *Ann. Appl. Probab.*, 12(1) :248–271, 2002.

[18] Emmanuel Gobet. *Méthodes de Monte-Carlo et processus stochastiques : du linéaire au non-linéaire*. Editions de l'Ecole Polytechnique, 2013.

[19] Bruce Hajek. Cooling schedules for optimal annealing. *Math. Oper. Res.*, 13(2) :311–329, 1988.

[20] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1) :97–109, April 1970.

[21] Benjamin Jourdain. *Méthodes de Monte Carlo*. Format électronique, 2021.

[22] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov chains and mixing times*. Format électronique, 2009.

[23] Makoto Matsumoto and Takuji Nishimura. Mersenne twister : A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Model. Comput. Simul.*, 8(1) :3–30, January 1998.

[24] Alexander J. McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative risk management*. Princeton Series in Finance. Princeton University Press, Princeton, NJ, 2005. Concepts, techniques and tools.

[25] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6) :1087–1092, 1953.

[26] Roger B. Nelsen. *An introduction to copulas*. Springer Series in Statistics. Springer, New York, second edition, 2006.

[27] Harald Niederreiter. *Random number generation and quasi-Monte Carlo methods*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992.

[28] James R. Norris. *Markov Chains*. Cambridge University Press, 1997.

[29] Christian P. Robert and George Casella. *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2004.

[30] Christian P. Robert and George Casella. *Introducing Monte Carlo methods with R*. Use R ! Springer, New York, 2010.

[31] Sheldon M. Ross. *Simulation*. Statistical Modeling and Decision Science. Academic Press, Inc., San Diego, CA, second edition, 1997.

[32] Justin Salez. *Temps de mélange des chaînes de Markov*. Format électronique, 2021.

[33] Luke Tierney. Markov chains for exploring posterior distributions. *Ann. Statist.*, 22(4) :1701–1762, 1994. With discussion and a rejoinder by the author.

[34] Bruno Tuffin. *La simulation de Monte Carlo*. Hermes Science Publications, 2010.

[35] Bernard Ycart. *Modèles et algorithmes markoviens*. Springer-Verlag, Berlin, 2002.