Sorbonne Université
Master Mathématiques et Applications
Master 1, UM4MA379

Année 2025/2026
Second Semestre

Statistique Computationnelle

Arnaud Guyader

Table des matières

Chapitre 1

Simulation et intégration Monte-Carlo

Introduction

Si l'on suppose disposer d'un générateur de variables i.i.d. uniformes sur $[0,1]$, les méthodes d'inversion et de rejet permettent de simuler des variables ou vecteurs aléatoires suivant d'autres lois. Ceci s'applique en particulier à l'estimation de quantités pouvant s'exprimer sous formes d'espérances. Ces principes généraux sont illustrés en fin de chapitre dans le cadre de la statistique bayésienne, où la loi a posteriori joue un rôle central et n'est généralement accessible que par la simulation et l'estimation.

1.1 Simulation Monte-Carlo

1.1.1 La loi uniforme

Le but est de générer une suite de variables aléatoires (U_n) indépendantes et de loi uniforme sur $[0,1]$. Ceci est bien entendu impossible au sens strict sur une machine : 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) est bien i.i.d. est une question très délicate. Cette problématique est notamment liée à la théorie de la complexité de Kolmogorov (voir par exemple [5], Chapitre 7, pour une introduction à ce sujet).

En pratique, les logiciels se basent sur une suite dite pseudo-aléatoire, de la forme $X_{n+1} = f(X_n)$, où X_n prend ses valeurs dans un ensemble fini très grand, typiquement de taille bien supérieure à 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 (voir par exemple la Section 1.4 de [2], la Section 1.2 de [6] ou l'article [11]).

Mentionnons simplement un générateur classique, à savoir le Mersenne Twister MT-19937, qui s'écrit

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

où X_n est un vecteur de 19937 bits et A une matrice savamment choisie. La variable pseudo-aléatoire U_n correspond alors aux 32 derniers bits de X_n. Proposé par Matsumoto et Nishimura en 1998 [14], 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), qui est un nombre premier de Mersenne.

Dans tout ce cours, nous supposerons donc disposer d'un générateur pseudo-aléatoire "satisfaisant" pour la loi uniforme. La suite de cette section 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 Devroye [7].

1.1.2 Méthode d'inversion

La loi d'une variable aléatoire est complètement caractérisée par sa fonction de répartition. Même si celle-ci n'est pas bijective, on peut toujours définir son inverse généralisée comme suit.

Définition 1.1 (Inverse généralisée)
Soit F une fonction de répartition. On appelle inverse généralisée de F, ou fonction quantile, la fonction définie pour tout $u \in [0,1]$ par

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

avec les conventions $\inf \mathbb{R} = -\infty$ *et* $\inf \emptyset = +\infty$.

Remarque : Ainsi, on peut noter que $F^{-1}(0) = -\infty$ tandis que $F^{-1}(1)$ est la borne supérieure du support de la loi de X lorsque cette variable a pour fonction de répartition F.

Si F est bijective, il est clair que cette fonction quantile coïncide avec l'inverse classique de F (au sens de fonction réciproque), avec les conventions $F(-\infty) = 0$ et $F(\infty) = 1$. C'est en particulier le cas pour Φ, fonction de répartition de la loi normale standard (voir Figure 1.1).

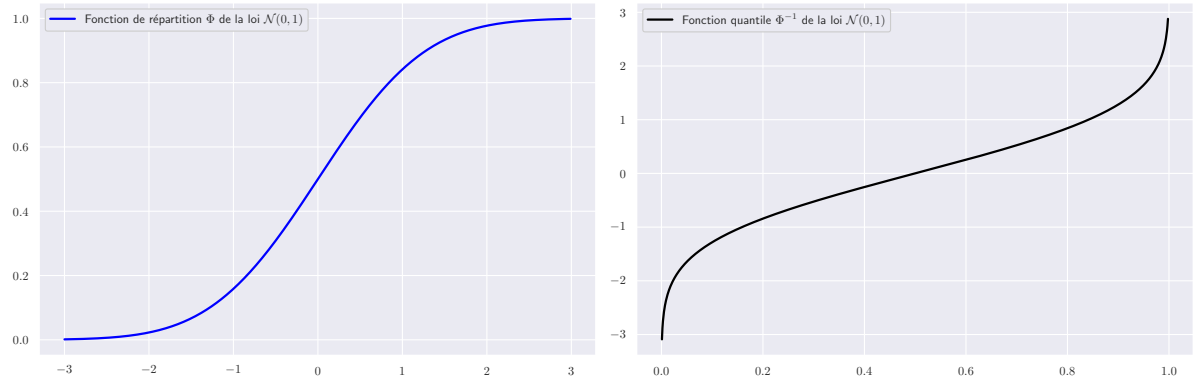

FIGURE 1.1 – Fonction de répartition Φ et fonction quantile Φ^{-1} de la loi normale standard.

Si F est la fonction de répartition d'une loi discrète, il existe une suite strictement croissante de réels $(x_n)_{n\geq 0}$ et une suite de poids $(p_n)_{n\geq 0}$ telles que $F(x) = \sum_{n=0}^{\infty} p_n \mathbf{1}_{x_n \leq x}$. L'inverse généralisée s'écrit alors, pour tout $u \in]0,1]$,

$$F^{-1}(u) = \sum_{n=0}^{\infty} x_n \mathbf{1}_{p_0 \cdots + p_{n-1} < u \leq p_0 \cdots + p_n},$$

avec la convention selon laquelle une somme vide fait 0. Outre que, tout comme F, cette fonction quantile est croissante et en escalier, on notera que, contrairement à F, elle est continue à gauche.

Convention : Dans toute la suite, nous conviendrons que $F(-\infty) = 0$ et $F(+\infty) = 1$ afin de définir sans ambiguïté la fonction composée $F \circ F^{-1}$ sur $[0,1]$.

Propriétés 1.2
Soit F une fonction de répartition et F^{-1} son inverse généralisée. Alors :

1. *Valeur en 0 :* $F^{-1}(0) = -\infty$.
2. *Monotonie :* F^{-1} *est croissante.*
3. *Régularité :* F^{-1} *est continue à gauche.*
4. *Equivalence :* $\forall u \in [0,1]$,

$$F(x) \geq u \iff x \geq F^{-1}(u). \tag{1.1}$$

5. *Inversibilité :* $\forall u \in [0,1]$, *on a* $(F \circ F^{-1})(u) \geq u$. *De plus :*
   - *si* F *est continue alors* $F \circ F^{-1} = Id$, *mais si elle n'est pas injective il existe* x_0 *tel que* $(F^{-1} \circ F)(x_0) < x_0$;
   - *si* F *est injective alors* $F^{-1} \circ F = Id$, *mais si elle n'est pas continue il existe* u_0 *tel que* $(F \circ F^{-1})(u_0) > u_0$;
   - *il y a équivalence entre* $F \circ F^{-1} = F^{-1} \circ F = Id$ *et l'inversibilité de* F *au sens usuel.*

Preuve : Les deux premiers points découlent de la définition de F^{-1}. Etablissons l'équivalence (1.1) : avec la convention $F^{-1}(0) = -\infty$, il n'y a rien à montrer pour $u = 0$, donc on peut considérer $u \in]0,1]$. Par définition de $F^{-1}(u)$, si $F(x) \geq u$, alors $x \geq F^{-1}(u)$. Inversement, si $F^{-1}(u) \leq x$, alors pour tout $\varepsilon > 0$ on a $F^{-1}(u) < x + \varepsilon$, donc par définition de $F^{-1}(u)$, il vient $u \leq F(x + \varepsilon)$. Puisque F est continue à droite, on en déduit que $u \leq F(x)$ et l'équivalence (1.1) est établie.

La continuité à gauche en découle : puisqu'il n'y a rien à prouver pour $u = 0$, il suffit en effet de montrer, grâce à la croissance de F^{-1}, que pour tout $u \in]0,1]$ et tout $\varepsilon > 0$, on peut trouver $\delta > 0$ tel que $F^{-1}(u - \delta) > F^{-1}(u) - \varepsilon =: x'$. Puisque $x' < F^{-1}(u)$, (1.1) assure que $F(x') < u$ donc $F(x') < u - \delta$ pour δ assez petit. Ceci implique $x' < F^{-1}(u - \delta)$, c'est-à-dire précisément ce qu'il fallait établir.

Pour le dernier point, il n'y a rien à prouver si $u = 0$. Si $u \in]0,1]$, d'après (1.1), on a

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

Supposons maintenant F continue. Alors, pour tout $u \in]0,1]$ et pour tout $\varepsilon > 0$, on a, toujours par (1.1),

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

Etant donné que $u \in]0,1]$ et que F est supposée continue, 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 $u \in]0,1]$, $(F \circ F^{-1})(u) = u$. Avec les conventions prises pour F et F^{-1}, ceci est encore vrai pour $u = 0$. Supposons F non injective, ce qui signifie qu'il existe $x_0' < x_0$ tels que $F(x_0') = F(x_0) = u_0$, donc

$$(F^{-1} \circ F)(x_0) = F^{-1}(u_0) \leq x_0' < x_0.$$

Dans le même ordre d'idée, si F est injective, alors quel que soit le réel x, il n'existe pas de réel $x' < x$ tel que $F(x') = F(x)$, donc

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

Si F n'est pas continue en un point x_0, il existe u_0 tel que $F(x_0^-) < u_0 < F(x_0)$, auquel cas

$$(F \circ F^{-1})(u_0) = F(F^{-1}(u_0)) = F(x_0) > u_0.$$

Quant au dernier point, il correspond exactement à la définition de la réciproque d'une fonction bijective, de sorte qu'il n'y a rien à démontrer.

■

Remarque : La preuve ci-dessus montre que si F est continue en $F^{-1}(u_0)$ alors $(F \circ F^{-1})(u_0) = u_0$.

Exemples : Illustrons le dernier point des Propriétés 1.2.

1. Si X suit une loi uniforme sur $[0,1]$, alors sa fonction de répartition F est continue mais pas injective. De fait, on a
$$(F^{-1} \circ F)(2) = F^{-1}(1) = 1 < 2.$$
2. Soit $Y \sim \mathcal{N}(0,1)$, $B \sim \mathcal{B}(1/2)$, avec Y et B indépendantes, et $X = BY$, alors la fonction de répartition de X présente un saut en 0 puisque $F(0^-) = 1/4$ tandis que $F(0) = 3/4$ (voir Figure 1.2). Elle est injective mais pas continue, et on voit que
$$(F \circ F^{-1})(1/2) = F(0) = \frac{3}{4} > \frac{1}{2}.$$

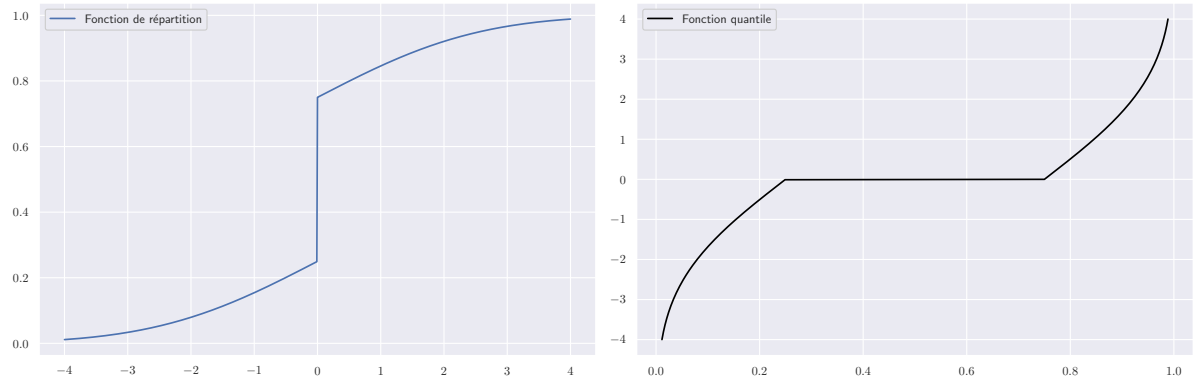

FIGURE 1.2 – Fonction de répartition de la variable $X = 2BY$ et son inverse généralisée.

Le résultat suivant est utile tant d'un point de vue pratique, par exemple pour les méthodes Monte-Carlo, que théorique, par exemple pour l'étude du processus empirique.

Lemme 1.3 (Universalité de la loi uniforme)
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 $[0,1]$.*

Preuve : Soit $X = F^{-1}(U)$ et x réel fixé, alors par (1.1) 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 $u \in [0,1]$, $P(U \leq u) = u$. Le premier point est donc établi. 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 F est continue, donc par le dernier point des Propriétés 1.2, $F \circ F^{-1} = Id$, donc $F(Y) = U$ et $F(X)$ est de loi uniforme sur $[0,1]$.

■

Remarque : Concernant le second point, il est clair que si X présente un atome en x_0, la variable $F(X)$ va hériter d'un atome en $F(x_0)$, donc ne sera certainement pas distribuée selon une loi uniforme. Par exemple, si $X \sim \mathcal{B}(1/3)$, alors $F(X)$ est une variable discrète prenant les valeurs $F(0) = 2/3$ et $F(1) = 1$ avec les probabilités respectives $2/3$ et $1/3$.

Conséquence : Méthode d'inversion. C'est le premier point du résultat précédent qui est connu sous le nom de méthode d'inversion en simulation Monte-Carlo. Concrètement, si l'on dispose d'une expression pour F^{-1} et d'un générateur de variables uniformes, alors on sait simuler une variable X de fonction de répartition F.

Exemples :

1. Loi exponentielle. On veut générer une variable X selon la loi exponentielle de paramètre $\lambda > 0$ fixé connu. Pour tout $x > 0$, $F(x) = 1 - e^{-\lambda x}$, bijective de $]0, \infty[$ vers $]0, 1[$. Il s'ensuit que pour tout $u \in]0, 1[$, $F^{-1}(u) = -(\log(1-u))/\lambda$. Dès lors, si U suit une loi uniforme sur $[0, 1]$, la variable $X = -(\log(1-U))/\lambda$ suit une loi exponentielle de paramètre 1. Puisque U a la même loi que $1 - U$, on peut même aller plus vite en considérant $X = -(\log U)/\lambda$.
2. Loi de Cauchy. On veut générer une variable X selon la loi de Cauchy standard, c'est-à-dire de densité $f(x) = 1/(\pi(1+x^2))$, donc de fonction de répartition $F(x) = (\pi/2 + \arctan x)/\pi$, bijective de $\mathbb{R}$ vers $]0, 1[$. Par la méthode d'inversion, si U suit une loi uniforme sur $]0, 1[$, $X = \tan(\pi(U - 1/2))$ suit une loi de Cauchy.
3. Loi de Bernoulli. Si $0 < p < 1$, alors $X = \mathbf{1}_{U \leq p} \sim \mathrm{Ber}(p)$.
4. Loi Géométrique. Puisque $1 - U \sim \mathrm{Unif}([0, 1])$ il vient

$$X = \sum_{n=0}^{\infty} n \mathbf{1}_{(1-p)^{n+1} < U < (1-p)^n} \sim \mathrm{Geo}_{\mathbb{N}}(p). \tag{1.2}$$

   On retrouve le fait que si $T \sim \mathrm{Exp}(\lambda)$, alors $X = \lfloor T \rfloor \sim \mathrm{Geo}_{\mathbb{N}}(1 - e^{-\lambda})$.

Remarque : Dans l'expression (1.2), le fait d'écrire des inégalités larges ou strictes n'a aucune espèce d'importance puisque la loi uniforme ne charge pas les singletons.

1.1.3 Méthode de rejet

Comme illustré en section précédente, la méthode d'inversion est un outil puissant pour simuler selon une loi donnée, à condition qu'il existe une formule explicite pour la fonction quantile F^{-1}. Malheureusement, ce n'est pas toujours le cas (cf. la loi normale standard). Dans ce type de situation, la méthode de rejet peut constituer une alternative pertinente.

Dans un cadre général, supposons la loi de X de densité f par rapport à une mesure[1] μ sur un espace mesurable $(E, \mathcal{E})$. Bien souvent, μ sera la mesure de Lebesgue sur $\mathbb{R}$ ou la mesure de comptage sur $\mathbb{N}$. Les conditions d'application de la méthode de rejet sont les suivantes.

Hypothèse 1.4
Il existe une densité g par rapport à μ telle que :
(a) On sait simuler selon g.
(b) Il existe un nombre réel m tel que $\forall y \in E, \quad f(y) \leq m g(y)$.
(c) pour tout y, on sait évaluer le rapport $r(y) := \frac{f(y)}{m g(y)}$, avec la convention $0/0 = 0$.

Remarques :

1. Dans ce cours, toutes les mesures sont supposées σ-finies.

1. Puisque f et g sont des densités, il est clair que
$$m = \int_E mg(x)\mu(dx) \geq \int_E f(x)\mu(dx) = 1.$$
2. Le fait que la formule de la densité f ne soit connue qu'à une constante de normalisation près n'est pas un obstacle. Supposons en effet que [2] $f(x) = cf_u(x)$ avec $c^{-1} = \int_E f_u(x)\mu(dx)$ inconnue. Si g est une densité répondant aux Hypothèses 1.4 avec f_u en lieu et place de f, alors ces mêmes conditions sont satisfaites pour f en remplaçant m par mc puisque le rapport
$$r(y) = \frac{f(y)}{mcg(y)} = \frac{f_u(y)}{mg(y)}$$
ne requiert pas la connaissance de c.
3. Il importe de faire le distinguo entre savoir évaluer une densité en tout point et savoir simuler selon cette densité. Par exemple, la densité $f(x) = \frac{2}{\Gamma(1/4)}\exp(-x^4)$ a une formule très simple, évaluable en tout point. Cependant, il n'y a pas de méthode élémentaire pour simuler selon cette densité [3]. Inversement, il n'y a rien de plus facile que de simuler la variable $X = (U_1 + 2U_2)/(3U_3 + 4U_4)$ où les U_i sont i.i.d. uniformes sur $[0, 1]$, mais la formule de sa densité n'a certainement rien d'élémentaire.

Soient (Y_n) et (U_n) deux suites i.i.d. et indépendantes entre elles, où Y_n a pour densité g par rapport à μ et U_n est uniforme sur $[0, 1]$. L'idée de la méthode de rejet est de générer des propositions Y_n jusqu'à ce que le critère $U_n \leq r(Y_n)$ soit satisfait.

Proposition 1.5 (Échantillonnage par rejet)
Soit $N := \min\{n \geq 1,\ U_n \leq r(Y_n)\}$ le premier instant où la proposition est acceptée. Alors :
- *$N \sim \mathrm{Geo}_{\mathbb{N}^*}(1/m)$, en particulier N est presque sûrement fini.*
- *La variable aléatoire $X = Y_N$ a pour densité f par rapport à μ.*
- *Les variables X et N sont indépendantes.*

Preuve : Pour tout $n \in \mathbb{N}$, on écrit
$$\mathbb{P}(N > n) = \mathbb{P}(U_1 > r(Y_1), \ldots, U_n > r(Y_n)) = \mathbb{P}(U_1 > r(Y_1))^n.$$
Puisque U_1 et Y_1 sont indépendantes, le théorème de Fubini-Tonelli donne
$$\mathbb{P}(U_1 > r(Y_1)) = \int_E \left(\int_{[0,1]} \mathbf{1}_{u>r(y)} du \right) g(y)\mu(dy) = \int_E (1 - r(y))g(y)\mu(dy) = 1 - \frac{1}{m},$$
ce qui prouve que $N \sim \mathrm{Geo}_{\mathbb{N}^*}(1/m)$. Soit maintenant $B \in \mathcal{E}$, alors
$$\mathbb{P}(X \in B) = \sum_{n=1}^{\infty} \mathbb{P}(Y_n \in B, N = n) = \sum_{n=1}^{\infty} \mathbb{P}(U_1 > r(Y_1), \ldots, U_{n-1} > r(Y_{n-1}), U_n \leq r(Y_n), Y_n \in B).$$
Par indépendance et d'après le résultat précédent,
$$\begin{aligned} \mathbb{P}(X \in B) &= \sum_{n=1}^{\infty} \mathbb{P}(U_1 > r(Y_1))^{n-1} \mathbb{P}(U_n \leq r(Y_n), Y_n \in B) \\ &= \sum_{n=1}^{\infty} \left(1 - \frac{1}{m}\right)^{n-1} \mathbb{P}(U_n \leq r(Y_n), Y_n \in B). \end{aligned}$$

2. Le u en indice de la notation f_u signifie *unnormalized*.
3. Le plus efficace serait sans doute d'appliquer une méthode de type Ziggourat [13], comme pour la gaussienne.

Une nouvelle application du théorème de Fubini-Tonelli donne

$$\mathbb{P}(U_n \leq r(Y_n), Y_n \in B) = \int_E \left(\int_{[0,1]} \mathbf{1}_{u \leq r(y)} du \right) \mathbf{1}_{y \in B} g(y) \mu(dy) = \frac{1}{m} \int_E \mathbf{1}_{y \in B} f(y) \mu(dy),$$

d'où

$$\mathbb{P}(X \in B) = \int_E \mathbf{1}_{y \in B} f(y) \mu(dy),$$

ce qui prouve que X a pour densité f par rapport à μ. Enfin, pour montrer que X et N sont indépendantes, on a vu précédemment que

$$\mathbb{P}(X \in B, N = n) = \mathbb{P}(Y_n \in B, N = n) = \frac{1}{m} \left(1 - \frac{1}{m}\right)^{n-1} \int_E \mathbf{1}_{y \in B} f(y) \mu(dy),$$

donc

$$\mathbb{P}(X \in B, N = n) = \mathbb{P}(X \in B) \mathbb{P}(N = n),$$

ce qui conclut la preuve.

■

Remarque : La moyenne d'une loi géométrique sur $\mathbb{N}^*$ étant égale à l'inverse de son paramètre, il s'ensuit que $\mathbb{E}[N] = m$. Ainsi, le résultat précédent montre que plus m est proche de 1, plus cet algorithme est efficace. Le cas $m = 1$ correspond à la situation où l'on sait simuler directement selon f. Par ailleurs, la probabilité de rejet de la proposition Y est $\mathbb{P}(N > 1) = 1 - \frac{1}{m}$.

Exemples :

1. Rejet basique. Supposons que l'on souhaite simuler uniformément dans un borélien non négligeable $B \subset [0,1]^d$, c'est-à-dire selon la densité $f(x) = \lambda(B)^{-1} \mathbf{1}_B(x)$. Supposons également que pour tout $y \in [0,1]^d$, on puisse décider si y appartient à B ou non. Une application triviale de la Proposition 1.5 consiste à simuler des points uniformément dans $[0,1]^d$ et à conserver ceux qui tombent dans B. Notons que, dans ce cas, la constante $m = 1/\lambda(B)$ peut être inconnue, mais ce n'est pas problématique puisque le rapport $f(Y)/(mg(Y)) = \mathbf{1}_{Y \in B}$ ne nécessite pas sa connaissance. De plus, comme ce rapport vaut 0 ou 1, il n'est pas nécessaire de simuler des variables aléatoires uniformes. Bien sûr, si $\lambda(B) \ll 1$, la probabilité de rejet $\mathbb{P}(N > 1) = 1 - \frac{1}{m} = 1 - \lambda(B)$ est très grande et il ne faut pas appliquer cette méthode, mais plutôt utiliser des techniques spécifiques de simulation d'événements rares.
2. Loi de Poisson. Supposons que $0 < \lambda < 1$ et que l'on souhaite simuler $X \sim \text{Poi}(\lambda)$, qui a pour densité $f(x) = e^{-\lambda} \lambda^x / x!$ par rapport à la mesure de comptage sur $\mathbb{N}$. Comme vu précédemment, on peut facilement générer $Y \sim \text{Geo}_{\mathbb{N}}(1 - \lambda)$. En posant

$$m = \sup_{y \in \mathbb{N}} \frac{f(y)}{g(y)} = \sup_{y \in \mathbb{N}} \frac{e^{-\lambda}}{(1-\lambda) y!} = \frac{e^{-\lambda}}{1 - \lambda},$$

   il suffit d'appliquer la méthode de rejet.
3. Loi Gamma. Supposons que $r > 1$ et que l'on souhaite simuler $X \sim \text{Gamma}(r, 1)$, qui a pour densité $f(x) = \Gamma(r)^{-1} x^{r-1} e^{-x} \mathbf{1}_{x \geq 0}$ par rapport à la mesure de Lebesgue sur $\mathbb{R}$. On sait facilement générer $Y \sim \text{Exp}(1/r)$. De plus, une optimisation simple révèle que

$$m = \sup_{y \in \mathbb{R}} \frac{f(y)}{g(y)} = \Gamma(r)^{-1} r^r e^{-(r-1)} \implies r(y) = \frac{f(y)}{mg(y)} = \left(\frac{ey}{r}\right)^{r-1} e^{-\frac{r-1}{r} y}.$$

   L'intuition derrière la distribution instrumentale $\text{Exp}(1/r)$ est qu'elle a la même espérance que la loi $\text{Gamma}(r, 1)$. En fait, un calcul direct mais fastidieux révèle que c'est le choix optimal parmi les lois $\text{Exp}(\lambda)$ au sens où c'est celui qui donne la constante m la plus petite.

Remarques :

1. Outre la loi uniforme, celle dont on a le plus souvent besoin en simulation est la loi gaussienne standard. À ce jour, la méthode la plus souvent implémentée est l'algorithme de Ziggourat [13], que l'on peut voir comme une combinaison des méthodes de rejet et d'inversion.
2. De façon générale, comme le but est d'aller le plus vite possible pour simuler selon une loi donnée, les méthodes réellement mises en place dans les logiciels sont rarement purement du rejet ou purement de l'inversion, mais plutôt un savant dosage des deux, voire une technique spécifique à la loi en question. De plus, cette loi est en général paramétrique et le choix de la méthode de simulation peut dépendre de la valeur du paramètre. C'est ce qui se passe par exemple en python pour la loi de Poisson de paramètre λ : deux méthodes différentes sont utilisées, selon que λ est inférieur ou supérieur à 10 (voir [10]). C'est aussi le cas pour simuler selon la loi Gamma.

1.1.4 Conditionnement et mélanges

Le résultat suivant est clair.

Lemme 1.6 (Simulation par conditionnement)
Soit (X, Y) un couple de variables aléatoires. Si l'on peut simuler selon la loi de Y et si, pour tout y dans son support, on peut simuler selon $\mathrm{Loi}(X|Y = y)$*, alors on peut simuler* $(X, Y) \sim \mathrm{Loi}(X, Y)$ *et, par conséquent,* $X \sim \mathrm{Loi}(X)$.

Une application classique du conditionnement est la simulation selon une loi de mélange.

Définition 1.7 (Loi de mélange)
Soient $(E, \mathcal{E})$ et $(F, \mathcal{F}, \nu)$ deux espaces mesurables et (X, Y) un couple aléatoire à valeurs dans $E \times F$. Notons P_Y la loi de Y sur $(E, \mathcal{E})$ et supposons que, pour tout y, $\mathrm{Loi}(X|Y = y)$ *admet une densité $f(x|y)$ par rapport à ν. Alors la loi de X est absolument continue par rapport à ν avec la densité dite de mélange*

$$f(x) = \int_E f(x|y) P_Y(dy).$$

La loi de X est appelée une loi de mélange.

D'après le Lemme 1.6, pour simuler X, il suffit donc de simuler $Y = y$ selon P_Y, puis X grâce à la densité $f(x|y)$.

Exemple : Modèle de mélange gaussien. Ici $E = \{0, 1\}$, $Y \sim \mathrm{Ber}(p)$ c'est-à-dire $P_Y = (1-p)\delta_0 + p\delta_1$, $(F, \mathcal{F}, \nu) = (\mathbb{R}, \mathcal{B}, dx)$, et la loi de X sachant $Y = y$ est la loi normale de moyenne m_y et de variance σ_y^2. Ainsi, la loi de X est un mélange de deux lois gaussiennes et admet la densité suivante par rapport à la mesure de Lebesgue (voir Figure 1.3) :

$$f(x) = \frac{1-p}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(x-m_0)^2}{2\sigma_0^2}} + \frac{p}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-m_1)^2}{2\sigma_1^2}}. \tag{1.3}$$

Pour simuler X, on commence par le tirage de $Y \sim \mathrm{Ber}(p)$ et suivant le résultat on simule ou bien $X_0 \sim \mathcal{N}(m_0, \sigma_0^2)$, ou bien $X_1 \sim \mathcal{N}(m_1, \sigma_1^2)$, toutes les variables étant indépendantes, c'est-à-dire : $X = (1-Y)X_0 + YX_1$. Ce type de distribution est très populaire en probabilités et statistique, par exemple en modélisation de populations.

Remarque : Statistique bayésienne. Le modèle de mélange apparaît naturellement en statistique bayésienne. Le contexte suivant sera détaillé en Section 1.3.1 : le paramètre aléatoire inconnu $\boldsymbol{\theta}$ a une loi a priori Π sur Θ et, sachant $\boldsymbol{\theta} = \theta$, l'observation X a une densité (ou vraisemblance)

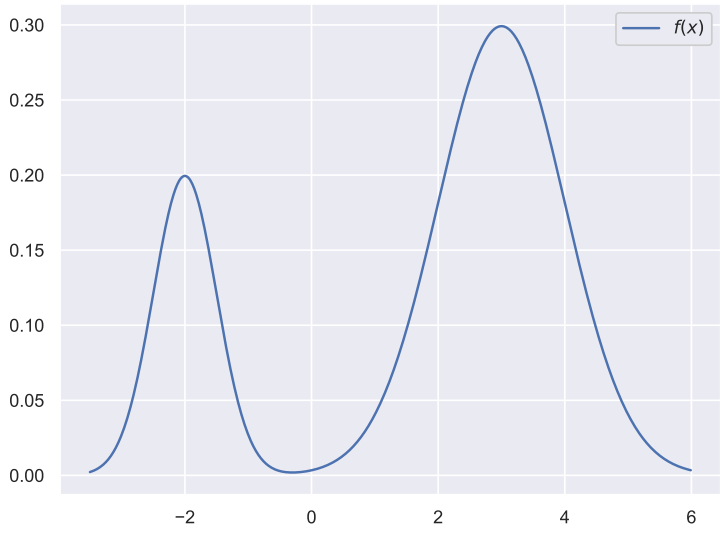

FIGURE 1.3 – Mélange de gaussiennes (1.3) avec $p = \frac{3}{4}$, $(m_0, \sigma_0) = (-2, \frac{1}{2})$ et $(m_1, \sigma_1) = (3, 1)$.

$p_\theta(x) = f(x|\theta)$ par rapport à une mesure ν sur $(E, \mathcal{E})$. Par conséquent, la loi de X est un mélange et sa densité par rapport à ν, appelée densité marginale, s'écrit

$$f(x) = \int_\Theta p_\theta(x)\Pi(d\theta).$$

1.1.5 Vecteurs aléatoires

L'énoncé suivant garantit que, au moins théoriquement, la simulation de vecteurs aléatoires peut être réduite à la simulation de variables aléatoires i.i.d. uniformes.

Lemme 1.8
Pour toute loi μ sur $\mathbb{R}^d$, il existe une application $\varphi_\mu : \mathbb{R}^d \to \mathbb{R}^d$ telle que si $U_1, \dots, U_d$ sont i.i.d. de loi uniforme sur $[0, 1]$, alors le vecteur aléatoire $\varphi_\mu(U_1, \dots, U_d)$ a pour loi μ.

En d'autres termes, μ est la mesure image de $\mathrm{Unif}([0,1]^d) = \mathrm{Unif}([0,1])^{\otimes d}$ par φ_μ. Ainsi, si l'on connaît φ_μ, le problème de la simulation selon μ est résolu. Cependant, φ_μ n'est généralement pas explicite, comme on peut déjà le voir dans le cas $d = 1$: si F désigne la fonction de répartition de μ, alors $\varphi_\mu = F^{-1}$. Ainsi le Lemme 1.8 n'est-il que la généralisation de la méthode d'inversion du Lemme 1.3. De même qu'il n'existe pas de technique universelle pour simuler une variable aléatoire de distribution donnée, il n'en existe pas non plus pour un vecteur aléatoire.

Un cas classique est celui où $\mathbf{X} = \phi(\mathbf{Y})$ pour une fonction ϕ donnée et un vecteur aléatoire $\mathbf{Y}$ que l'on sait simuler. Rappelons que dans le cas où $\phi : \mathbb{R}^d \to \mathbb{R}^d$ est un C^1-difféomorphisme et $\mathbf{Y}$ a une densité g par rapport à la mesure de Lebesgue, alors la densité de $\mathbf{X} = \phi(\mathbf{Y})$ est

$$f(\mathbf{x}) = g(\phi^{-1}(\mathbf{x}))|\det(J_{\phi^{-1}}(\mathbf{x}))| = \frac{g(\phi^{-1}(\mathbf{x}))}{|\det(J_\phi(\phi^{-1}(\mathbf{x}))|}. \tag{1.4}$$

A cet égard, l'un des exemples les plus populaires est sans doute le suivant.

Lemme 1.9 (Box-Muller)
Les coordonnées d'un vecteur aléatoire (X, Y) sont des gaussiennes standards i.i.d. si et seulement si $X = R\cos\boldsymbol{\theta}$ et $Y = R\sin\boldsymbol{\theta}$, où le module R et l'argument $\boldsymbol{\theta}$ sont indépendants, avec $R^2 \sim \mathrm{Exp}(1/2)$ et $\boldsymbol{\theta} \sim \mathrm{Unif}([0, 2\pi])$.

Preuve : L'application $\phi : U = (0,\infty)\times(0,2\pi) \to V = \mathbb{R}^2 \setminus ([0,\infty)\times\{0\})$ définie par $(x,y) = \phi(r,\theta) = (r\cos\theta, r\sin\theta)$ est un C^1-difféomorphisme avec $\det(J_\phi(r,\theta)) = r$, donc

$$\det(J_{\phi^{-1}}(x,y)) = \frac{1}{\det(J_\phi(\phi^{-1}(x,y)))} = \frac{1}{\sqrt{x^2+y^2}}.$$

Via ce difféomorphisme, l'existence d'une densité $f(x,y)$ pour le couple $(X,Y) = \phi(R,\boldsymbol{\theta})$, équivaut à l'existence d'une densité $g(r,\theta)$ pour le couple $(R,\boldsymbol{\theta})$, avec d'après (1.4)

$$f(x,y) = \frac{g(\phi^{-1}(x,y))}{\sqrt{x^2+y^2}} \quad \text{et} \quad g(r,\theta) = rf(\phi(r,\theta)) = rf(r\cos\theta, r\sin\theta).$$

Si $R^2 \sim \text{Exp}(1/2)$, on voit facilement que la densité de R est $g_R(r) = r\exp(-\frac{r^2}{2})\mathbf{1}_{r>0}$, donc la densité de $(R,\boldsymbol{\theta})$ est, par indépendance de R et $\boldsymbol{\theta}$,

$$g(r,\theta) = g_R(r)g_{\boldsymbol{\theta}}(\theta) = re^{-\frac{r^2}{2}}\mathbf{1}_{r\geq 0}\frac{1}{2\pi}\mathbf{1}_{[0,2\pi]}(\theta).$$

Puisque la densité d'un vecteur gaussien standard (X,Y) est

$$f(x,y) = \frac{1}{2\pi}e^{-\frac{x^2+y^2}{2}},$$

l'équivalence du Lemme 1.9 est établie. ■

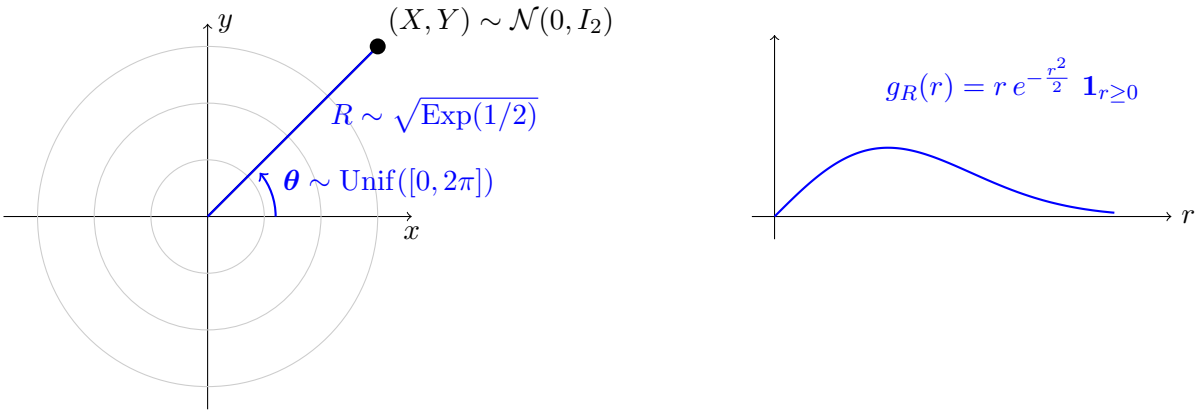

FIGURE 1.4 – Décomposition polaire d'un vecteur gaussien et loi de Rayleigh du rayon.

Ce résultat est intuitivement clair (voir Figure 1.4) : par isotropie du vecteur gaussien (X,Y), l'angle polaire $\boldsymbol{\theta}$ est uniforme, tandis que la loi du rayon R, appelée loi de Rayleigh, provient de [4]

$$R = \sqrt{X^2+Y^2} \overset{\text{Loi}}{=} \sqrt{\chi^2_2} \overset{\text{Loi}}{=} \sqrt{\text{Gamma}(1,1/2)} \overset{\text{Loi}}{=} \sqrt{\text{Exp}(1/2)}.$$

1.2 Intégration Monte-Carlo

Soit X un objet aléatoire à valeurs dans un espace mesurable $(E,\mathcal{E})$, et $\varphi : E \to \mathbb{R}$ une application mesurable telle que l'on puisse calculer $\varphi(x)$ pour tout x. Sous réserve d'existence, l'objectif est d'estimer

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

4. Rappelons que $\chi^2_d \overset{\text{Loi}}{=} \text{Gamma}(d/2,1/2)$ et $\text{Gamma}(1,\lambda) \overset{\text{Loi}}{=} \text{Exp}(\lambda)$.

La méthode la plus simple est celle du Monte-Carlo standard. Différentes techniques ont très vite été proposées pour obtenir des estimateurs plus précis : c'est ce qu'on appelle les méthodes de réduction de variance. Elles sont nombreuses, mais nous ne présenterons que deux d'entre elles (échantillonnage préférentiel et conditionnement). Pour la culture, nous dirons aussi un mot des méthodes numériques et des méthodes Quasi-Monte-Carlo.

De façon générale, si l'on veut comparer équitablement des méthodes Monte-Carlo, il faut prendre en compte à la fois les variances des estimateurs et leurs coûts algorithmiques respectifs. Un critère en ce sens est celui d'efficacité, correspondant à l'inverse du produit de la variance par le coût. Cependant, ceci n'est pas toujours facile à quantifier. Sur machine, une approche fruste mais raisonnable consiste, pour un même temps de calcul, à comparer les précisions obtenues.

1.2.1 Monte-Carlo standard

Si l'on sait simuler des variables X_i i.i.d. selon la loi de X, alors l'estimateur Monte-Carlo standard (ou naïf) est simplement la moyenne empirique des variables aléatoires i.i.d. $\varphi(X_i)$, c'est-à-dire

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

Les propriétés suivantes découlent de la linéarité de l'espérance, de la loi des grands nombres et du théorème central limite.

Propriétés 1.10 (Estimateur Monte-Carlo standard)
L'estimateur Monte-Carlo standard a les propriétés suivantes :

1. *Non-biais :* $\mathbb{E}[\hat{I}_n] = I$.
2. *Consistance forte :*
$$\hat{I}_n \xrightarrow[n\to\infty]{\text{p.s.}} I.$$
3. *Normalité asymptotique : si* $\sigma^2 := \mathrm{Var}(\varphi(X)) < \infty$, *alors*
$$\sqrt{n}\left(\hat{I}_n - I\right) \xrightarrow[n\to\infty]{\text{d}} \mathcal{N}(0, \sigma^2).$$

Puisque σ^2 est généralement inconnu, il faut l'estimer lui aussi. Cela peut être fait **avec le même échantillon** via la variance empirique

$$\hat{\sigma}_n^2 := \frac{1}{n}\sum_{i=1}^{n}\left(\varphi(X_i) - \hat{I}_n\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\varphi(X_i)^2 - \hat{I}_n^2.$$

La loi des grands nombres assure que $\hat{\sigma}_n$ est un estimateur fortement consistant de σ^2, et le Lemme de Slutsky donne

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

ce qui permet de construire des intervalles de confiance asymptotiques. Pour cela, soit $0 < \alpha < 1$ et notons Φ^{-1} la fonction quantile de la loi gaussienne standard. Par exemple, si $\alpha = 0.05$, alors $\Phi^{-1}(1-\alpha/2) = -\Phi^{-1}(\alpha/2) \approx 1.96$.

Lemme 1.11 (Intervalles de confiance asymptotiques)
Avec les notations précédentes,

$$\mathbb{P}\left(-\Phi^{-1}(1-\alpha/2) \le \sqrt{n}\frac{\hat{I}_n - I}{\hat{\sigma}_n} \le \Phi^{-1}(1-\alpha/2)\right) \xrightarrow[n\to\infty]{} 1-\alpha,$$

ce qui signifie qu'un intervalle de confiance de niveau asymptotique $(1-\alpha)$ *pour* I *est*

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

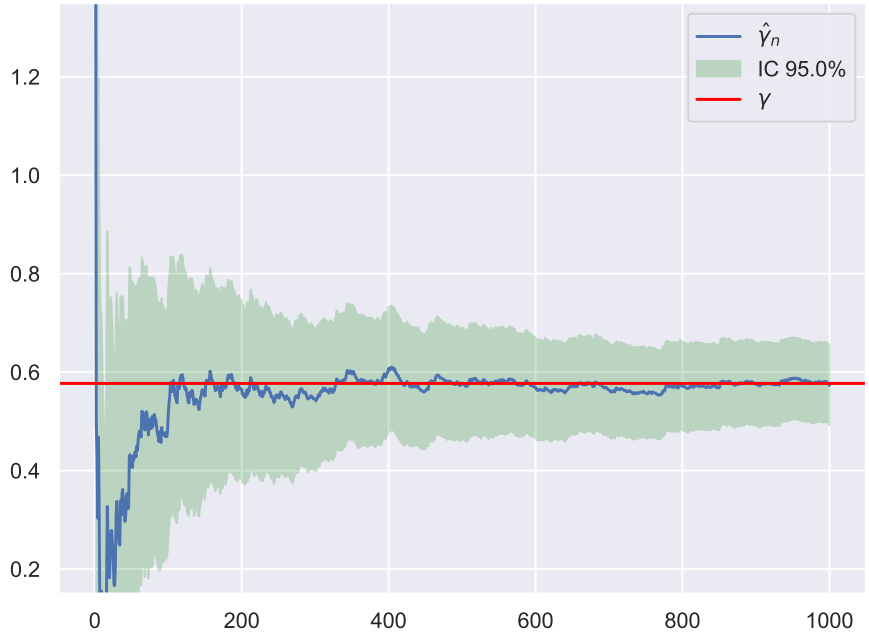

FIGURE 1.5 – Estimation de la constante d'Euler γ par Monte-Carlo standard.

Exemples :

1. Constante d'Euler. Soit X une variable aléatoire de loi de Gumbel standard, dont la fonction de répartition s'écrit $F(x) = \exp(-\exp(-x))$. On peut montrer que son espérance $\mathbb{E}[X] = \gamma \approx 0.577$ est la constante d'Euler. Par la méthode d'inversion, on peut simuler X via $X = F^{-1}(U) = -\log(-\log U)$ avec $U \sim \text{Unif}([0,1])$, et appliquer les résultats précédents pour obtenir une estimation $\hat{\gamma}_n$ de γ ainsi que des intervalles de confiance asymptotiques (voir Figure 1.5).

2. Estimation de π. Soit $X \sim \text{Unif}([0,1]\times[0,1])$ et $\varphi(x) = \mathbf{1}_{\|x\|\leq 1}$, où $\|x\|$ est la norme euclidienne de $x \in \mathbb{R}^2$, si bien que $I = \mathbb{E}[\varphi(X)] = \mathbb{P}(\|X\| \leq 1) = \pi/4$. Un estimateur de π est donc

$$\hat{\pi}_n := 4 \times \hat{I}_n = 4\sum_{i=1}^{n} \mathbf{1}_{\|X_i\|\leq 1},$$

avec

$$\sqrt{n}\,(\hat{\pi}_n - \pi) \xrightarrow[n\to\infty]{\text{d}} \mathcal{N}(0,\sigma^2) = \mathcal{N}(0,\pi(4-\pi)).$$

Remarque : Si $\mathbb{E}[|\varphi(X)|] < \infty$ mais $\sigma^2 = \text{Var}(\varphi(X)) = \infty$, l'estimateur $\hat{I}_n$ est sans biais et consistant mais on ne peut pas construire d'intervalles de confiance asymptotiques. La Figure 1.6 illustre ce qui se passe lorsque X suit une loi de Pareto de densité $f(x) = 2x^{-3}\mathbf{1}_{x\geq 1}$ donc de moyenne $I = \mathbb{E}[X] = 2$ mais de variance $\sigma^2 = \text{Var}(X) = \infty$.

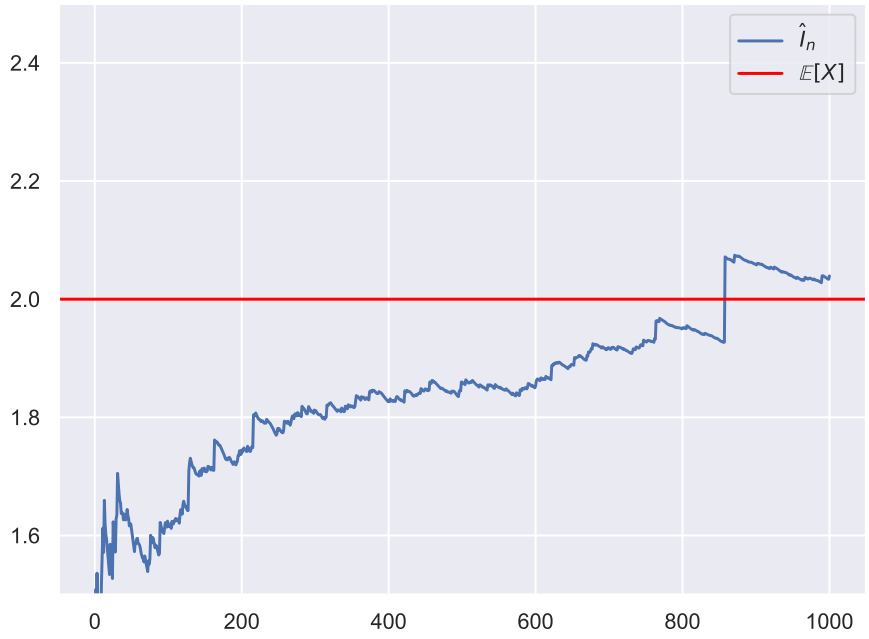

FIGURE 1.6 – Estimation de la moyenne d'une loi de Pareto de variance infinie.

1.2.2 Échantillonnage préférentiel

L'objectif est toujours d'estimer

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

Les Propriétés 1.10 indiquent que lorsque n est grand, la précision du Monte-Carlo naïf est de l'ordre $\sigma/\sqrt{n}$, où $\sigma^2 = \mathrm{Var}(\varphi(X))$. Il s'avère que si X prend la plupart de ses valeurs là où φ est petit, alors cette méthode n'est pas satisfaisante au sens où σ est grand par rapport à I.

Exemples :

1. Evénement rare. Soit X une variable aléatoire que l'on peut facilement simuler, et définissons $\varphi(x) = \mathbf{1}_{x>q}$, où q est un nombre réel donné. On cherche à estimer $p = \mathbb{E}[\varphi(X)] = \mathbb{P}(X > q)$. Lorsque p est petit, l'écart-type $\sigma = \sqrt{p(1-p)}$ est beaucoup plus grand que p lui-même. Par conséquent, pour obtenir une précision relative satisfaisante, il faut simuler un échantillon de taille $n \gg 1/p$. Dans cet exemple, φ est non nul uniquement dans la queue de la distribution de Y.
2. Exemple jouet. Soient $m > 0$, $X \sim \mathcal{N}(m, 1)$, $\varphi(x) = \exp(-mx + \frac{m^2}{2})$, et on veut estimer $I = \mathbb{E}[\varphi(X)]$. Pour tout m, un calcul élémentaire montre que $I = 1$ et $\sigma^2 = \exp(m^2) - 1$, donc σ devient beaucoup plus grand que I lorsque m augmente. Ceci est clair intuitivement (voir Figure 1.7) : environ 95% des X_i tombent dans l'intervalle $[m-2, m+2]$, intervalle dans lequel $\varphi(x)$ est très petit si m est grand, ce qui numériquement implique un estimateur désastreux.

L'objectif des techniques dites de **réduction de variance** est d'atténuer ce problème. Parmi celles-ci, l'échantillonnage préférentiel (ou *Importance Sampling*) est certainement la plus populaire.

Hypothèse 1.12
Soit g soit une densité par rapport à μ sur $(E, \mathcal{E})$ telle que :

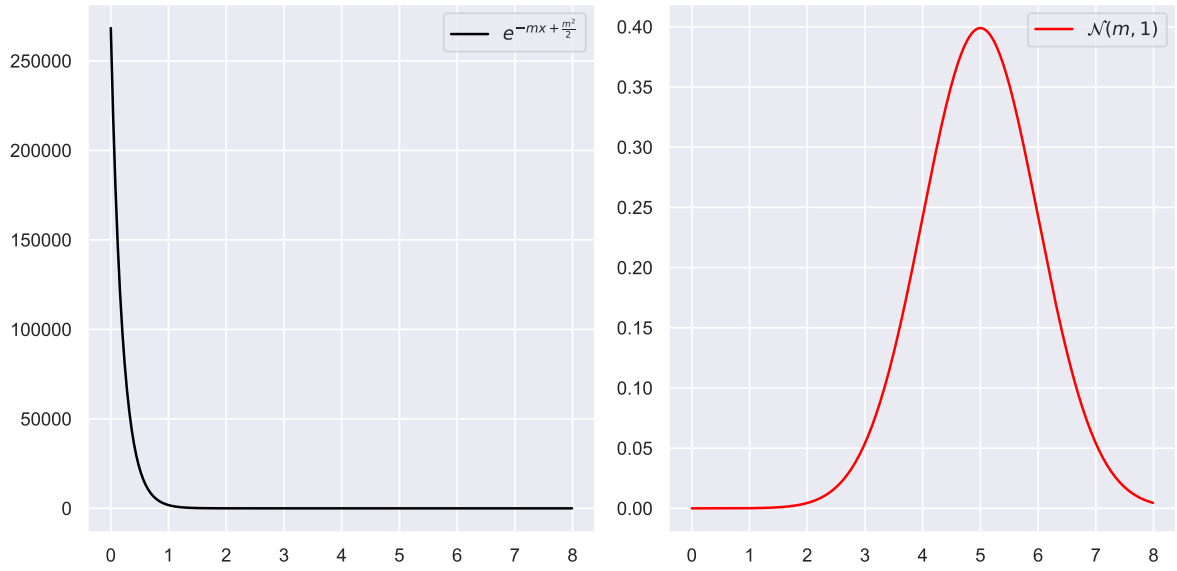

FIGURE 1.7 – Monte-Carlo naïf en échec : $I = \mathbb{E}\left[e^{-mX+\frac{m^2}{2}}\right]$ avec $X \sim \mathcal{N}(m,1)$ et $m = 5$.

(a) On sait simuler selon g.
(b) $f \cdot \mu \ll g \cdot \mu$ sur l'ensemble $E_0 = \{\varphi \neq 0\}$.
(c) On sait calculer $w(y) = \frac{f(y)}{g(y)}$ pour tout y dans E_0.

Commençons par écrire

$$I = \mathbb{E}[\varphi(X)] = \int_E \varphi(x) f(x) \mu(dx) = \int_{E_0} \varphi(x) f(x) \mu(dx).$$

Puisque $f \cdot \mu \ll g \cdot \mu$ sur l'ensemble E_0, le Théorème de Radon-Nikodym assure qu'il existe une fonction positive mesurable $\frac{d(f \cdot \mu)}{d(g \cdot \mu)}$ définie sur E_0 telle que

$$I = \int_{E_0} \varphi(y) \frac{d(f \cdot \mu)}{d(g \cdot \mu)}(y) g(y) \mu(dy),$$

or sur E_0 cette dérivée de Radon-Nikodym n'est rien d'autre que la fonction w ci-dessus, si bien que

$$I = \int_{E_0} \varphi(y) w(y) g(y) \mu(dy) = \int_E \varphi(y) w(y) g(y) \mu(dy) = \mathbb{E}[w(Y)\varphi(Y)],$$

où Y est une variable de densité g. L'estimateur par échantillonnage préférentiel est tout bonnement l'estimateur Monte-Carlo naïf de cette dernière expression. Les propriétés suivantes sont donc immédiates.

Propriétés 1.13 (Échantillonnage préférentiel)
Avec les notations précédentes, en considérant des variables Y_i i.i.d. selon la densité g, l'estimateur par échantillonnage préférentiel

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

est sans biais et fortement consistant. Si $s^2 := \mathrm{Var}(w(Y)\varphi(Y)) < \infty$*, alors* $\tilde{I}_n$ *est asymptotiquement gaussien :*

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

Dans ce cas, un estimateur fortement consistant de s^2 *est*

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

et un intervalle de confiance de niveau asymptotique $(1-\alpha)$ *pour* I *est*

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

Remarque : Si l'hypothèse "$f\cdot\mu \ll g\cdot\mu$ sur l'ensemble $E_0 = \{\varphi \neq 0\}$" n'est pas satisfaite, alors $\tilde{I}_n$ est biaisé puisque $\varphi(X)$ met une partie de sa masse en dehors du support de la loi de $\varphi(Y)$.

Exemples :

1. Prenons l'exemple caricatural où $X \sim \mathrm{Exp}(1)$, $\varphi(x) = \mathbf{1}_{x\geq 0}$, donc $I = 1$. Maintenant, si $Y \sim 1 + \mathrm{Exp}(1)$, le rapport $w(Y) = f(Y)/g(Y) = 1/e$ est p.s. bien défini puisque $Y \geq 1$. Cependant
$$\mathbb{E}[w(Y)\varphi(Y)] = \frac{1}{e}\mathbb{P}(Y \geq 0) = \frac{1}{e} < 1 = I.$$
Ceci est dû au fait que sur $E_0 = \{\varphi \neq 0\} = \mathbb{R}_+$, la loi de X n'est pas absolument continue par rapport à celle de Y : en particulier $\mathbb{P}(0 < X < 1) = 1 - 1/e > 0$ alors que $\mathbb{P}(0 < Y < 1) = 0$.
2. Une condition suffisante est que la loi de X soit absolument continue par rapport à celle de Y, mais ceci est inutilement restrictif car peu importe ce qui se passe là où la fonction φ est nulle. Restons sur l'exemple précédent et considérons $\varphi(x) = \mathbf{1}_{x\geq 1}$, donc $E_0 = \{\varphi \neq 0\} = [1, \infty[$. Cette fois on a bien "$f\cdot\mu \ll g\cdot\mu$ sur l'ensemble $E_0 = \{\varphi \neq 0\}$" et l'échantillonnage préférentiel s'applique.

Remarque : Si $s^2 = \infty$, alors $\tilde{I}_n$ est consistant mais sans intérêt, car on ne peut pas construire d'intervalles de confiance. Cela se produit en particulier lorsque g est à queue plus légère que f.

Exemple : Considérons la situation caricaturale suivante. Soit $q > 1$ connu fixé et $p := \mathbb{P}(X > q)$ où X suit une loi de Pareto de densité $f(x) = x^{-2}\mathbf{1}_{x\geq 1}$, donc $p = 1/q$. Comme mentionné précédemment, la méthode Monte-Carlo naïve a pour variance asymptotique $\sigma^2 = p(1-p)$. Supposons maintenant qu'on applique l'échantillonnage préférentiel en utilisant comme loi de proposition la loi exponentielle translatée de q, c'est-à-dire $Y \sim q + \mathrm{Exp}(1)$ de densité $g(y) = \exp(-(y-q))\mathbf{1}_{y\geq q}$. La condition "$g\cdot\mu \ll f\cdot\mu$ sur l'ensemble $E_0 = \{\varphi \neq 0\}$" est bien vérifiée donc l'estimateur est non biaisé et fortement consistant. Cependant, sa variance asymptotique $s^2 = \mathrm{Var}(w(Y)\varphi(Y))$ est infinie puisque

$$\mathbb{E}\left[(w(Y)\varphi(Y))^2\right] = \int_q^\infty y^{-4}e^{y-q}dy = \infty.$$

Numériquement, le problème sera donc le même que celui illustré Figure 1.6.

Plus généralement, l'échantillonnage préférentiel n'est intéressant que si la variance associée est inférieure à celle du Monte-Carlo naïf, c'est-à-dire $\mathrm{Var}(w(Y)\varphi(Y)) < \mathrm{Var}(\varphi(X))$. Par conséquent, une question naturelle est celle du meilleur choix pour la densité de proposition g.

Lemme 1.14 (Choix optimal en échantillonnage préférentiel)
Pour toute densité instrumentale g, on a

$$s^2 = \mathrm{Var}(w(Y)\varphi(Y)) \geq \mathbb{E}\left[|\varphi(X)|\right]^2 - I^2 =: s_\star^2,$$

et la borne inférieure est atteinte pour

$$g^\star(y) := \frac{|\varphi(y)|f(y)}{\int_E |\varphi(y)|f(y)\mu(dy)} = \frac{|\varphi(y)|f(y)}{\mathbb{E}\left[|\varphi(X)|\right]}.$$

Preuve : Comme $\mathrm{Var}(w(Y)|\varphi(Y)|) \geq 0$, il vient

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

avec égalité si et seulement si $\mathrm{Var}(w(Y)|\varphi(Y)|) = 0$, c'est-à-dire si et seulement si la variable aléatoire $w(Y)|\varphi(Y)|$ est presque sûrement constante. Cette constante ne peut être que son espérance, notée

$$c := \mathbb{E}\left[w(Y)|\varphi(Y)|\right] = \mathbb{E}\left[|\varphi(X)|\right].$$

Dire que $w(Y)|\varphi(Y)| = c$ presque sûrement équivaut à dire que

$$0 = \mathbb{E}\left[|w(Y)|\varphi(Y)| - c|\right] = \int_E \left|\frac{f(y)}{g(y)}|\varphi(y)| - c\right| g(y)\mu(dy) = \int_E |f(y)|\varphi(y)| - cg(y)|\,\mu(dy),$$

ce qui n'est possible que si $cg(y) = f(y)|\varphi(y)|$ μ-presque partout. Ceci conclut la preuve puisqu'une densité par rapport à la mesure de référence μ n'est définie que μ-presque partout.

■

Si φ est de signe constant, alors $s_\star^2 = 0$, donc $\mathrm{Var}(\tilde{I}_n) = s_\star^2/n = 0$, c'est-à-dire un estimateur de variance nulle, donc p.s. constant et $n = 1$ suffit pour estimer I ! Cependant, dans ce cas, $w^\star(Y)\varphi(Y) = I$ est impossible à calculer à moins de connaître déjà I. De fait, ce résultat n'est pas utile. Le seul message est qu'une "bonne" densité de proposition g doit mettre une partie de sa masse là où le produit de f par φ est grand. A ce sujet, revenons aux exemples déjà mentionnés.

Exemples :

1. Evénement rare. La densité optimale est la restriction de f à l'événement rare, ce qui signifie que $Y \sim \mathrm{Loi}(X|X > q)$ et $g^\star(y) = p^{-1}f(y)\mathbf{1}_{y>q}$.
2. Exemple jouet. Un calcul simple révèle que $g^\star$ est la densité d'une variable gaussienne standard, c'est-à-dire $Y \sim \mathcal{N}(0,1)$. Cette loi offre en quelque sorte le meilleur compromis entre la loi de X et la fonction φ.

1.2.3 Conditionnement

On cherche toujours à estimer $I = \mathbb{E}[\varphi(X)]$. L'idée dans cette section a été vue en cours de probabilités, à savoir : le conditionnement ne change pas la moyenne, mais réduit l'incertitude.

Hypothèse 1.15
On considère une variable aléatoire Y telle que :
(a) On sait simuler Y.
(b) On sait calculer $\psi(Y) := \mathbb{E}[\varphi(X)|Y]$.

En remarquant que $I = \mathbb{E}[\varphi(X)] = \mathbb{E}[\mathbb{E}[\varphi(X)|Y]] = \mathbb{E}[\psi(Y)]$, l'estimateur par conditionnement n'est rien d'autre que l'estimateur Monte-Carlo standard de cette dernière espérance.

Proposition 1.16 (Estimation par conditionnement)
Avec les notations précédentes, en considérant des variables Y_i i.i.d. selon la loi de Y, l'estimateur de I par conditionnement est défini par

$$\tilde{I}_n := \frac{1}{n}\sum_{i=1}^{n} \psi(Y_i).$$

Il est sans biais et fortement consistant. Si $\sigma^2 := \mathrm{Var}(\varphi(X)) < \infty$, alors $\tilde{I}_n$ est asymptotiquement normal :

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

avec

$$s^2 := \mathrm{Var}(\psi(Y)) \leq \sigma^2.$$

Un estimateur fortement consistant de s^2 est

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

et un intervalle de confiance de niveau asymptotique $(1-\alpha)$ pour I est

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

Preuve : Le non-biais vient de la propriété classique de l'espérance conditionnelle déjà mentionnée :

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

La consistance forte découle de ce point et de la loi des grands nombres, tandis que la normalité asymptotique est une conséquence du TCL. Enfin, puisque $\varphi(X)$ est de carré intégrable, l'espérance conditionnelle $\psi(Y)$ est la projection orthogonale de $\varphi(X)$ sur le sous-espace[5] fermé $\sigma(Y)$ correspondant à l'ensemble des variables aléatoires $h(Y)$ où h est borélienne et telle que $\mathbb{E}[h(Y)^2] < \infty$ (cf. Figure 1.8). Par le Théorème de Pythagore, on a donc

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

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

$$\sigma^2 = \mathrm{Var}(\varphi(X)) \geq \mathrm{Var}(\psi(Y)) = s^2.$$

■

Remarques :

1. Une autre façon de voir que $s^2 \leq \sigma^2$ est d'appliquer la formule dite de la variance totale, c'est-à-dire que la variance est la somme de l'espérance de la variance conditionnelle et de la variance de l'espérance conditionnelle :

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

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

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

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

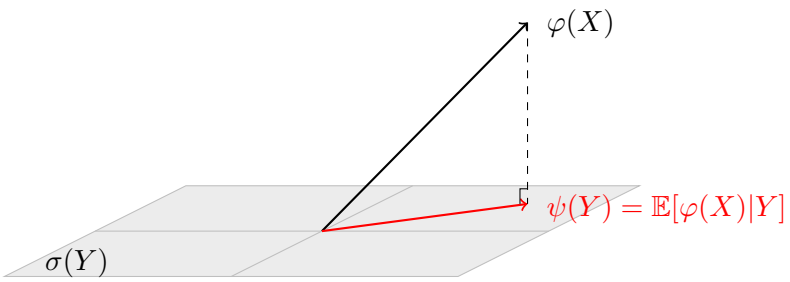

FIGURE 1.8 – Interprétation de l'espérance conditionnelle comme projection orthogonale.

2. Les deux situations extrêmes sont les suivantes : si l'on choisit $Y = X$, on retombe sur le Monte-Carlo naïf, tandis que si Y est indépendante de X, alors une seule donnée suffit puisque $\psi(Y) = \mathbb{E}[\varphi(X)|Y] = I$, mais ceci suppose que l'on connaît déjà le résultat. Bref, la méthode par conditionnement n'a d'intérêt que si l'on peut identifier une variable Y non indépendante de X, facilement simulable et pour laquelle on sait calculer $\psi(Y) = \mathbb{E}[\varphi(X)|Y]$. Ceci arrive en particulier lorsque X est un vecteur aléatoire et que l'on conditionne par une partie de celui-ci.

Exemple : On revient à la surface du quart de disque unité D c'est-à-dire, en notant $X = (U, V)$ avec U et V i.i.d. uniformes sur $[0, 1]$,

$$I = \frac{\pi}{4} = \mathbb{P}\left(\|X\| \leq 1\right) = \mathbb{P}\left(U^2 + V^2 \leq 1\right) = \mathbb{E}\left[\mathbf{1}_{U^2+V^2\leq 1}\right] = \mathbb{E}[\mathbf{1}_D(X)].$$

Par Monte-Carlo naïf, l'écart-type asymptotique de $\hat{I}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{U_i^2+V_i^2\leq 1}$ est $\sigma = (\frac{\pi}{4}(1-\frac{\pi}{4}))^{1/2} \approx 0.41$. Appliquons la méthode de conditionnement avec $Y = V$. Les variables U et V étant indépendantes, il vient [6]

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

ce qui conduit à l'estimateur

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

Sa variance se calcule facilement :

$$s^2 = \mathbb{E}[\psi^2(V)] - I^2 = \int_0^1 (1 - v^2)dv - (\pi/4)^2 = 2/3 - (\pi/4)^2 \implies s \approx 0.22.$$

Par rapport au Monte-Carlo naïf, on gagne environ 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 U_i (cf. Figure 1.9).

1.2.4 Quelques mots sur l'intégration numérique

Supposons pour simplifier que, dans la quantité d'intérêt $I = \mathbb{E}[\varphi(X)]$, la variable X suit une loi uniforme sur $[0, 1]^d$, c'est-à-dire que

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

6. Rappel : si U et V sont indépendantes, alors $\mathbb{E}[\varphi(U, V)|V] = h(V)$ où $h(v) = \mathbb{E}[\varphi(U, v)]$ pour tout v.

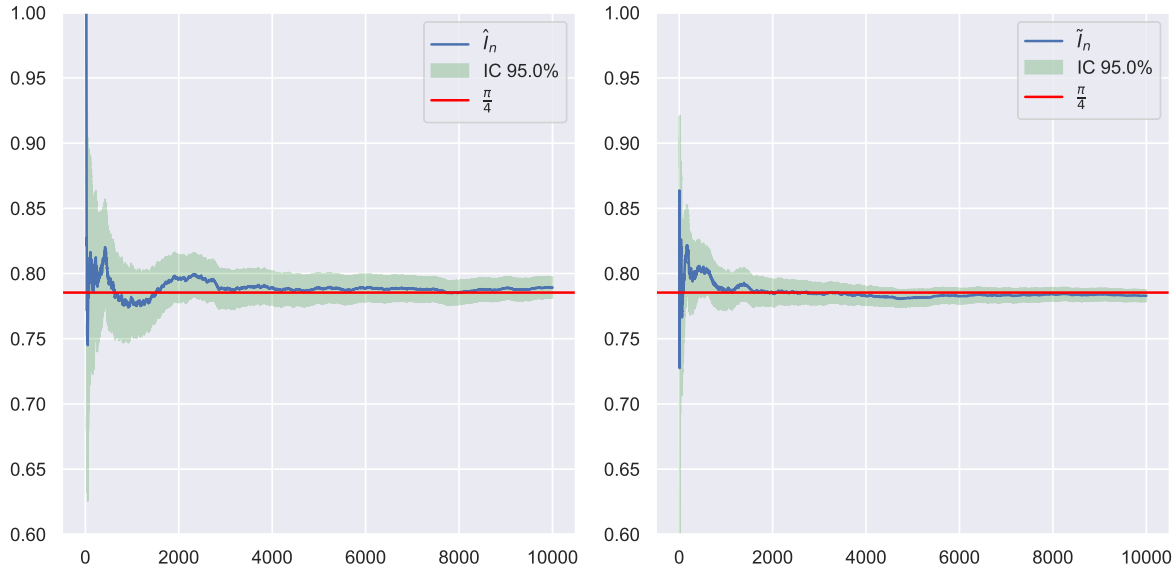

FIGURE 1.9 – Estimateurs naïf $\hat{I}_n$ et par conditionnement $\tilde{I}_n$ du quart de disque unité.

Dans ce cas, l'estimateur Monte-Carlo standard de I est

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

où les X_i sont des vecteurs i.i.d. uniformes sur $[0,1]^d$. Cependant, 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 :

— Si φ est de classe $\mathcal{C}^1$, la méthode des rectangles sur la subdivision $\{1/n, \dots, (n-1)/n, 1\}$ a une précision en $\mathcal{O}(1/n)$. Plus précisément, si on note $\|\varphi'\|_\infty = \sup_{x\in[0,1]} |\varphi'(x)|$ et R_n l'approximation obtenue par les rectangles, alors

$$|R_n - I| \leq \frac{\|\varphi'\|_\infty}{2n}.$$

— 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 T_n l'approximation obtenue, alors

$$|T_n - I| \leq \frac{\|\varphi''\|_\infty}{12n^2}.$$

— 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 S_n l'approximation obtenue, alors

$$|S_n - I| \leq \frac{\|\varphi^{(4)}\|_\infty}{2880n^4}.$$

— 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 a donc le même type de borne qu'en dimension 1 :

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

Néanmoins, pour comparer ce qui est comparable, il faut considérer que les deux méthodes font le même nombre n d'appels[7] 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 standard 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 se détériore avec d est symptomatique du fléau de la dimension.

Avec une vitesse en $1/\sqrt{n}$, la méthode Monte-Carlo est quant à elle insensible à la dimension et peut donc s'avérer avantageuse dès que l'on travaille en dimension grande ou sur une fonction irrégulière. En fin de chapitre, Section 1.4, nous dirons un mot des méthodes Quasi-Monte-Carlo, qui représentent un compromis entre les méthodes déterministes et les méthodes Monte-Carlo.

1.3 Illustration en statistique bayésienne

Cette section rappelle brièvement le cadre bayésien et présente certaines applications des méthodes Monte-Carlo pour la simulation de lois a posteriori et l'estimation de quantités relatives à celles-ci. Pour plus de détails sur le paradigme bayésien et de nombreuses questions afférentes, on pourra se reporter au polycopié d'Anna Ben-Hamou [1] ainsi qu'au livre de Christian Robert [20].

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

1.3.1 Simulation selon la loi a posteriori

Le cadre bayésien est le suivant :

— Le paramètre aléatoire $\boldsymbol{\theta}$ vit dans l'espace mesurable $(\Theta, \mathcal{F})$ et a pour loi Π, supposée absolument continue par rapport à une mesure μ :

$$\Pi(d\theta) = \pi(\theta)\mu(d\theta).$$

Cette loi Π est appelée loi a priori (ou *prior* en anglais).

— Sachant $\boldsymbol{\theta} = \theta$, l'observation aléatoire X, vivant dans l'espace $(E, \mathcal{E})$, a pour loi P_θ et le modèle $(P_\theta)_{\theta\in\Theta}$ est dominé par une mesure ν :

$$P_\theta(dx) = p_\theta(x)\nu(dx).$$

La fonction p_θ est appelée la vraisemblance (ou *likelihood*).

Pour résumer, on écrira

$$\boldsymbol{\theta} \sim \Pi \qquad \text{et} \qquad X|\boldsymbol{\theta} \sim P_{\boldsymbol{\theta}}.$$

Il en découle que la loi de X (ou *marginal*) est absolument continue par rapport à ν et a pour densité

$$f(x) = \int_\Theta p_\theta(x)\Pi(d\theta) = \int_\Theta p_\theta(x)\pi(\theta)\mu(d\theta).$$

Dans ce cadre, on appelle loi a posteriori (ou *posterior*) la loi de $\boldsymbol{\theta}$ sachant X, notée $\Pi(\cdot|X)$ ou plus explicitement $\text{Loi}(\boldsymbol{\theta}|X)$. Celle-ci est absolument continue par rapport à μ et a pour densité

$$\forall\theta \in \Theta, \qquad \pi(\theta|X) = \frac{p_\theta(X)\pi(\theta)}{f(X)},$$

ce que l'on résume parfois en disant

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal}}.$$

Hormis dans quelques cas d'école[8], même si l'on a des formules explicites pour la densité a priori et la vraisemblance, ce n'est plus le cas pour la densité a posteriori, et ce en raison de l'intégrale définissant la densité marginale $f(x)$. De plus, même lorsque l'on dispose d'une formule explicite pour la densité a posteriori $\pi(\theta|X)$, cela n'implique pas que l'on sache simuler facilement selon celle-ci.

Remarques :

1. Les objets aléatoires $\boldsymbol{\theta}$ et X ne jouent pas le même rôle. On connaît la loi a priori Π mais on n'observe pas le paramètre $\boldsymbol{\theta}$. On connaît le modèle $(P_\theta)_{\theta\in\Theta}$ et on observe X. Cette observation permet d'affiner la connaissance du paramètre inconnu et aléatoire $\boldsymbol{\theta}$, ce qui se traduit grosso modo par le fait que la loi a posteriori est plus concentrée que la loi a priori.
2. Dans toute la suite, il convient de garder en tête que l'observation X est **fixée** ! Par conséquent, tout est effectué "sachant X". De plus, sachant $\boldsymbol{\theta}$, l'observation X consiste généralement en un N-uplet i.i.d., ce que l'on note

$$X|\boldsymbol{\theta} = (X_1, \ldots, X_N)|\boldsymbol{\theta} \sim P_{\boldsymbol{\theta}}^{\otimes N}.$$

8. Par exemple si Π appartient à une famille de lois conjuguée par rapport au modèle $(P_\theta)_{\theta\in\Theta}$.

Supposons maintenant qu'étant donnée l'observation X, on souhaite simuler selon la loi a posteriori. Dans les situations les plus simples, celle-ci appartient à une famille connue de lois selon lesquelles on sait simuler facilement. C'est en particulier le cas du modèle le plus classique.

Exemple : Modèle fondamental. On considère le modèle

$$\boldsymbol{\theta} \sim \mathcal{N}(0,1) \qquad \text{et} \qquad X = (X_1, \ldots, X_N)|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, 1)^{\otimes N}.$$

Un calcul simple montre que la loi a posteriori est encore gaussienne :

$$\Pi(\cdot|X) = \text{Loi}(\boldsymbol{\theta}|X) = \mathcal{N}\left(\frac{N\bar{X}_N}{N+1}, \frac{1}{N+1}\right),$$

loi selon laquelle on sait simuler. Il n'y a donc aucun problème pour obtenir un échantillon selon la loi a posteriori. Cet exemple permet également d'illustrer la remarque ci-dessus selon laquelle la loi a posteriori est plus concentrée que la loi a priori (voir Figure 1.10).

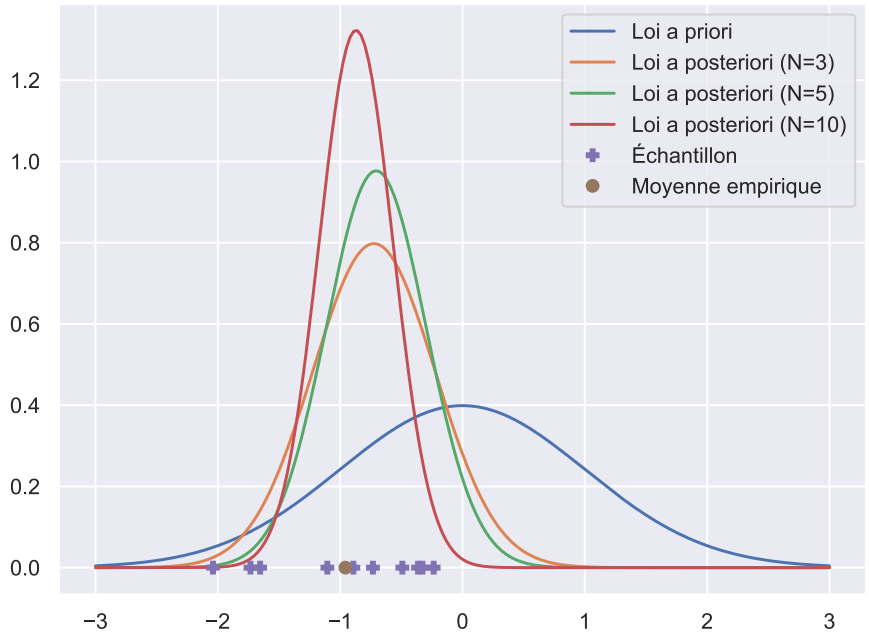

FIGURE 1.10 – Loi a priori et lois a posteriori pour le modèle fondamental gaussien.

Si le paramètre $\boldsymbol{\theta}$ est une variable aléatoire réelle, notons F_X la fonction de répartition de la loi a posteriori, c'est-à-dire

$$F_X(\theta) := \mathbb{P}(\boldsymbol{\theta} \leq \theta|X).$$

Si celle-ci admet une inverse généralisée explicite, alors la simulation selon la loi a posteriori ne pose pas problème : il suffit d'appliquer la méthode d'inversion du Lemme 1.3.

Si l'on n'est pas dans l'une de ces situations, on peut éventuellement s'en sortir grâce à la méthode de rejet. Pour cela, il faut trouver une densité conditionnelle $g(\theta) = g(\theta|X)$ selon laquelle on sait simuler et pour laquelle il existe une constante $m = m(X)$ telle que

$$\forall \theta \in \Theta, \qquad \pi(\theta|X) = \frac{p_\theta(X)\pi(\theta)}{f(X)} \leq mg(\theta), \tag{1.5}$$

et de sorte que, pour tout θ, on sache évaluer le rapport $r(\theta) = \pi(\theta|X)/(mg(\theta))$.

Cas particulier : Rejet et Maximum de Vraisemblance. Supposons les points suivants :

— On peut simuler selon la loi a priori Π, c'est-à-dire selon la densité π ;
— On sait calculer la vraisemblance $p_\theta(X)$ pour tout θ ;
— L'estimateur du maximum de vraisemblance $\hat{\theta}_N = \hat{\theta}_N(X)$ admet une formule explicite.

Alors, en posant

$$g(\theta) := \pi(\theta) \qquad \text{et} \qquad m := \frac{p_{\hat{\theta}_N}(X)}{f(X)},$$

la condition (1.5) est bien vérifiée et on peut donc appliquer la méthode de rejet. Le rapport d'intérêt vaut tout simplement $r(\theta) = \frac{p_\theta(X)}{p_{\hat{\theta}_N}(X)}$, qui est bien inférieur ou égal à 1.

Exemple : Modèle Cauchy-Gauss. Considérons la situation où

$$\boldsymbol{\theta} \sim \text{Cauchy} \qquad \text{et} \qquad X|\boldsymbol{\theta} = (X_1, \dots, X_N)|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, 1)^{\otimes N}.$$

Avec les notations précédentes, nous pouvons donc prendre pour mesures dominantes les mesures de Lebesgue $\mu = \lambda$ et $\nu = \lambda_N$ et les densités

$$\pi(\theta) = \frac{1}{\pi(1+\theta^2)} \qquad \text{et} \qquad p_\theta(X) = \prod_{j=1}^N \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_j-\theta)^2}{2}} = \frac{e^{-\frac{N}{2}\{(\theta-\bar{X}_N)^2+V_X\}}}{(2\pi)^{\frac{N}{2}}}$$

où

$$V_X := \frac{1}{N}\sum_{j=1}^N X_i^2 - (\bar{X}_N)^2$$

désigne la variance empirique des observations. La marginale s'écrit donc

$$f(X) = \int_{\mathbb{R}} p_t(X)\pi(t)dt = \frac{e^{-\frac{N}{2}V_X}}{\pi(2\pi)^{\frac{N}{2}}} \int_{\mathbb{R}} \frac{e^{-\frac{N}{2}(t-\bar{X}_N)^2}}{1+t^2} dt,$$

qui n'est pas triviale. La loi a posteriori n'est pas une loi classique, sa densité étant donc :

$$\pi(\theta|X) \propto \frac{e^{-\frac{N}{2}(\theta-\bar{X}_N)^2}}{1+\theta^2}.$$

Cependant, simuler selon la loi a priori, c'est-à-dire la loi de Cauchy, est très simple par la méthode d'inversion : si $U \sim \text{Unif}([0,1])$, alors comme nous l'avons vu

$$\boldsymbol{\theta} = \tan\left(\frac{\pi}{2}\left(U - \frac{1}{2}\right)\right) \sim \text{Cauchy}.$$

De plus, dans ce modèle, il est facile de voir que l'EMV est $\hat{\theta}_N = \bar{X}_N$. Par conséquent, pour toute proposition $\boldsymbol{\theta}$ selon la loi de Cauchy, on peut facilement calculer le rapport

$$r(\boldsymbol{\theta}) = \frac{p_{\boldsymbol{\theta}}(X)}{p_{\bar{X}_N}(X)} = e^{-\frac{N}{2}\{(\boldsymbol{\theta}-\bar{X}_N)^2\}}.$$

Cette proposition $\boldsymbol{\theta}$ aura donc d'autant moins de chances d'être acceptée qu'elle est éloignée de la moyenne empirique $\bar{X}_N$ des données.

Remarque : Tout comme pour le rejet basique vu en Section 1.1.3, on notera que la connaissance de $m = p_{\hat{\theta}_N}(X)/f(X)$ n'est nullement requise pour pouvoir appliquer cette méthode.

1.3.2 Estimations pour la loi a posteriori

On veut estimer la moyenne

$$I = I(X) := \mathbb{E}[\varphi(\boldsymbol{\theta})|X]$$

d'une fonction test par rapport à la loi a posteriori. Si l'on sait simuler selon celle-ci, alors on peut toujours appliquer la méthode Monte-Carlo standard, c'est-à-dire considérer l'estimateur

$$\hat{I}_n := \frac{1}{n}\sum_{i=1}^{n} \varphi(\boldsymbol{\theta}_i) \qquad \text{avec} \qquad \boldsymbol{\theta}_i \overset{\text{i.i.d.}}{\sim} \text{Loi}(\boldsymbol{\theta}|X).$$

Celui-ci hérite naturellement des Propriétés 1.10. Par exemple si $\sigma^2(X) := \text{Var}(\varphi(\boldsymbol{\theta})|X) < \infty$ p.s., alors sachant X on a

$$\sqrt{n}\left(\hat{I}_n - I\right) \xrightarrow[n\to\infty]{\text{d}} \mathcal{N}(0, \sigma^2(X)).$$

Exemple : Modèle Cauchy-Gauss. Supposons que l'on veuille estimer l'estimateur de Bayes pour la perte quadratique, qui n'est autre que la moyenne a posteriori $m(X) := \mathbb{E}[\boldsymbol{\theta}|X]$. Nous venons de voir que la méthode de rejet permet d'obtenir un échantillon $(\boldsymbol{\theta}_i)_{1\le i\le n}$ i.i.d. selon la loi a posteriori. Ainsi, un estimateur de $m(X)$ est

$$\hat{m}_n(X) := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\theta}_i.$$

En notant $\sigma^2(X) := \text{Var}(\boldsymbol{\theta}|X)$ la variance a posteriori, il est facile de vérifier [9] que $\sigma^2(X) < \infty$ p.s., donc

$$\sqrt{n}\left(\hat{m}_n(X) - m(X)\right) \xrightarrow[n\to\infty]{\text{d}} \mathcal{N}(0, \sigma^2(X)).$$

On notera au passage un inconvénient de cette technique : puisqu'elle est basée sur la méthode de rejet, pour obtenir n variables $(\boldsymbol{\theta}_i)_{1\le i\le n}$ i.i.d. selon la loi a posteriori, il faut en moyenne simuler mn couples aléatoires $(\boldsymbol{\theta}, U)$, où $(\boldsymbol{\theta}, U) \sim \Pi \otimes \text{Unif}([0,1])$ et $m = p_{\bar{X}_N}(X)/f(X)$.

En fait, on peut s'en sortir sans savoir simuler selon la loi a posteriori. Il suffit de savoir :
- simuler selon la loi a priori Π ;
- calculer la vraisemblance $p_\theta(X)$ pour toute valeur θ.

En effet, si ces conditions sont réunies, puisque

$$I = \mathbb{E}[\varphi(\boldsymbol{\theta})|X] = \int_\Theta \varphi(\theta)\pi(\theta|X)\mu(d\theta) = \frac{\int_\Theta \varphi(\theta)p_\theta(X)\pi(\theta)\mu(d\theta)}{\int_\Theta p_\theta(X)\pi(\theta)\mu(d\theta)}, \tag{1.6}$$

il suffit d'estimer haut et bas par Monte-Carlo standard : la simulation de $(\boldsymbol{\theta}_i)_{1\le i\le n}$ i.i.d. selon Π permet en effet de construire l'estimateur

$$\tilde{I}_n := \frac{\frac{1}{n}\sum_{i=1}^{n} \varphi(\boldsymbol{\theta}_i)p_{\boldsymbol{\theta}_i}(X)}{\frac{1}{n}\sum_{i=1}^{n} p_{\boldsymbol{\theta}_i}(X)}.$$

Par la loi des grands nombres, conditionnellement à X, le haut tend p.s. vers l'intégrale du haut en (1.6), le bas vers l'intégrale du bas, d'où par le Théorème de continuité

$$\hat{I}_n \xrightarrow[n\to\infty]{\text{p.s.}} I = \mathbb{E}[\varphi(\boldsymbol{\theta})|X].$$

9. Montrer que l'intégrale définissant le moment d'ordre 2 pour la loi a posteriori est p.s. finie.

Exemple : Modèle Cauchy-Gauss. Pour estimer la moyenne a posteriori, il suffit donc de simuler des variables $(\boldsymbol{\theta}_i)_{1\leq i\leq n}$ i.i.d. selon la loi de Cauchy (élémentaire) et de calculer

$$\tilde{m}_n(X) := \frac{\sum_{i=1}^n \boldsymbol{\theta}_i \; e^{-\frac{N}{2}(\boldsymbol{\theta}_i-\bar{X}_N)^2}}{\sum_{i=1}^n e^{-\frac{N}{2}(\boldsymbol{\theta}_i-\bar{X}_N)^2}},$$

pour en déduire que, sachant X,

$$\tilde{m}_n(X) \xrightarrow[n\to\infty]{\text{p.s.}} m(X).$$

Remarque : Si la consistance de $\tilde{m}_n(X)$ est immédiate par la loi des grands nombres et le théorème de continuité, il n'en va pas de même de la normalité asymptotique. On peut cependant montrer que cette propriété est vraie en passant par la méthode Delta et le TCL multidimensionnel.

1.4 Complément : méthodes Quasi-Monte-Carlo

Revenons au problème d'intégration par rapport à la loi uniforme évoqué en Section 1.2.4, autrement dit l'estimation de

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

Si $d = 1$, pour la méthode Monte-Carlo standard, c'est-à-dire lorsque $(X_n)_{n\geq 1}$ est une suite de variables aléatoires 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 $\{1/n, \ldots, (n-1)/n, 1\}$, 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 1.17 (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 infinie entre la fonction de répartition de la loi discrète $\frac{1}{n}\sum_{i=1}^n \delta_{\xi_i}$ 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 :

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

   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 :

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

   d'où, retraduit en écriture usuelle, la suite de van der Corput de base 2 :

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

   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 $b_1, \ldots, b_d$ et on construit les d suites de van der Corput $\xi^{(1)}, \ldots, \xi^{(d)}$ de bases respectives $b_1, \ldots, b_d$. La suite de Halton associée est alors

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

   Par exemple, en dimension 2 avec $b_1 = 2$ et $b_2 = 3$, les premiers termes de la suite de Halton associée sont (voir aussi Figure 1.11)

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

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

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.

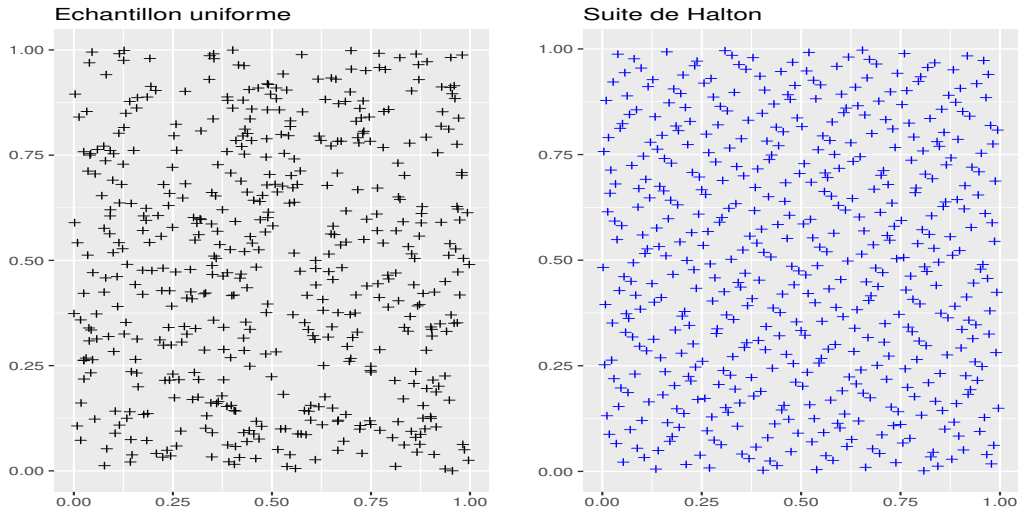

FIGURE 1.11 – Echantillon uniforme de taille $n = 500$ vs n premiers points d'une suite de Halton.

Définition 1.18 (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

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

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, etc. Néanmoins, cette notion intervient de façon cruciale pour majorer la qualité d'un estimateur basé sur une séquence $(\xi_n)_{n \geq 1}$.

Théorème 1.19 (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'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 1.20 (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 la quantité

$$I_n = \frac{1}{n}\sum_{i=1}^{n} \varphi(\xi_i)$$

est une approximation 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 $\mathcal{O}(n^{-s/d})$. De ce fait, les méthodes Quasi-Monte-Carlo sont considérées comme compétitives en dimension "intermédiaire" et pour des fonctions suffisamment régulières.

Remarques :

1. Contrairement aux méthodes Monte-Carlo, les méthodes Quasi-Monte-Carlo n'ont rien d'aléatoire. Leur nom vient de la formule d'approximation

   $$I_n = \frac{1}{n}\sum_{i=1}^{n} \varphi(\xi_i),$$

   qui est donc la même que pour un estimateur Monte-Carlo standard. En particulier, elle présente l'intérêt d'obéir elle aussi à la formule de récurrence

   $$I_{n+1} = \frac{n}{n+1} I_n + \frac{n}{n+1}\varphi(\xi_{n+1}).$$

2. Sur toute cette section, on trouvera compléments et exemples en Chapitre 7 de [22] ainsi qu'en Chapitre 3 de [17].

Chapitre 2

Monte-Carlo par Chaînes de Markov

Introduction

Etant donnée une loi π selon laquelle on souhaite simuler et estimer, le principe des méthodes de Monte-Carlo par Chaînes de Markov (MCMC) est de construire une chaîne de Markov (X_n) ayant π comme unique loi invariante et telle que la convergence en variation totale et le théorème ergodique (équivalent de la loi des grands nombres) soient vérifiés. Comme on ne veut pas se limiter aux espaces d'états dénombrables, il convient au préalable de présenter la notion de chaîne de Markov en espace d'états général. Ce préambule se veut plus pragmatique que théorique : le but est avant tout de fixer un cadre et des notations pour les algorithmes qui vont suivre.

2.1 Chaînes de Markov

Nous considérons un espace mesurable général $(E, \mathcal{E})$, par exemple un espace polonais[1]. Cette hypothèse permet par exemple de parler de lois conditionnelles. Par ailleurs, comme précédemment, toutes les mesures sur $(E, \mathcal{E})$ seront supposées σ-finies, ce qui assure que des résultats fondamentaux comme les théorèmes de Fubini et de Radon-Nikodym sont valides.

On gardera en tête les deux situations suivantes :

— E est dénombrable[2] et $\mathcal{E} = \mathcal{P}(E)$ est l'ensemble de toutes les parties de E.

— $E = \mathbb{R}^d$ ou un pavé de $\mathbb{R}^d$ et $\mathcal{E}$ est la tribu borélienne associée, notée $\mathcal{B}(E)$.

Le premier cas nous servira surtout à faire le lien avec les résultats déjà étudiés pour les chaînes de Markov en espaces d'états discrets. La seconde situation est celle qui nous intéressera en fin de chapitre lorsque nous étudierons l'application des méthodes MCMC à la statistique bayésienne.

Sur les chaînes de Markov en espace d'états dénombrable, nous nous contenterons de rappeler certains résultats classiques, prouvés notamment dans les polycopiés de Thierry Lévy [12], Thierry Bodineau [3] ou Djalil Chafaï [4]. Le livre de Norris [18] est une référence classique en anglais. En espace d'états général, la théorie est nettement plus ardue et des références telles que Meyn and Tweedie [16] ou Douc *et al.* [8] dépassent le niveau M1. Mentionnons enfin que l'ouvrage de Robert et Casella [19] est l'un des rares à aborder à la fois les aspects théoriques et pratiques.

2.1.1 Noyau de transition, récurrence aléatoire

En espace d'états général, l'équivalent de la matrice de transition s'appelle un noyau de transition et est défini comme suit.

1. Un espace polonais est un espace topologique qui admet une métrique compatible avec sa topologie, pour laquelle il est séparable et complet.

2. Dans toute la suite, "dénombrable" signifiera "au plus dénombrable".

Définition 2.1 (Noyau de transition)
L'application $P : E \times \mathcal{E} \to [0,1]$ est un noyau de transition, ou noyau markovien, si :
- *Pour tout x de E, $P(x,\cdot)$ est une mesure de probabilité sur $(E,\mathcal{E})$;*
- *Pour tout B de $\mathcal{E}$, l'application $x \mapsto P(x,B)$ est mesurable de $(E,\mathcal{E})$ vers $(\mathbb{R},\mathcal{B}(\mathbb{R}))$.*

Notation : Pour tout x fixé, puisque $P(x,\cdot)$ est une mesure sur l'espace mesurable $(E,\mathcal{E})$, la théorie de l'intégration "à la Lebesgue" s'applique. En outre, puisque c'est une mesure de probabilité, elle somme à 1, ce qui s'écrit

$$\int_E P(x,dy) = 1.$$

De même, pour x fixé, si Y est aléatoire à valeurs dans $(E,\mathcal{E})$ et de loi $P(x,\cdot)$, alors pour toute fonction test [3] $\varphi : E \to \mathbb{R}$, $\varphi(Y)$ est une variable aléatoire dont l'espérance s'écrit, par le Théorème de transfert,

$$\mathbb{E}[\varphi(Y)] = \int_E P(x,dy)\varphi(y).$$

Exemples :

1. Si E est dénombrable, le noyau P correspond à une matrice de transition, éventuellement de taille infinie, c'est-à-dire $P = [p(x,y)]_{(x,y)\in E\times E}$ avec $0 \le p(x,y) \le 1$ pour tout couple (x,y) et $\sum_y p(x,y) = 1$ pour tout x. Autrement dit, chaque ligne de la matrice P est une loi de probabilité discrète.
2. Si $(E,\mathcal{E}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, on dira que le noyau P est absolument continu ou à densité, sous-entendu par rapport à la mesure de Lebesgue $\lambda(dx) = dx$, si pour tout x on a $P(x,\cdot) \ll \lambda$, auquel cas on notera $p(x,\cdot) = \frac{dP(x,\cdot)}{d\lambda}$ cette densité, fonction borélienne de $\mathbb{R}^d$ dans $\mathbb{R}_+$ telle que
$$\int_{\mathbb{R}^d} p(x,y)dy = 1.$$
On note donc $P(x,dy) = p(x,y)dy$ et, pour toute fonction test $\varphi : \mathbb{R}^d \to \mathbb{R}$, on a
$$\int_{\mathbb{R}^d} \varphi(y)P(x,dy) = \int_{\mathbb{R}^d} \varphi(y)p(x,y)dy.$$
3. Noyau gaussien : un exemple classique de noyau à densité est le noyau dit gaussien où, pour tout réel x, $P(x,\cdot)$ est la loi normale $\mathcal{N}(x,1)$, c'est-à-dire
$$p(x,y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-x)^2}{2}} = f(y-x),$$
où f est la densité de la gaussienne standard.
4. Noyau gaussien normalisé : soit $-1 < r < 1$ un réel fixé, un autre exemple de noyau à densité est celui où, pour tout réel x, $P(x,\cdot)$ est la loi normale $\mathcal{N}(rx, 1-r^2)$. La densité s'écrit donc
$$p(x,y) = \frac{1}{\sqrt{2\pi(1-r^2)}} e^{-\frac{(y-rx)^2}{2(1-r^2)}}.$$
5. Mélange : toujours sur l'espace d'états $(E,\mathcal{E}) = (\mathbb{R},\mathcal{B}(\mathbb{R}))$, soit $0 < p < 1$ un réel fixé. Un exemple typique de noyau non absolument continu est fourni par le mélange entre un Dirac en x et une loi normale standard, ce qui s'écrit
$$P(x,dy) = p\delta_x(dy) + (1-p)f(y)dy. \tag{2.1}$$

3. Une fonction test désignera une fonction mesurable qui est positive ou bornée, bref une fonction pour laquelle l'intégration par rapport à n'importe quelle mesure positive a un sens.

La loi $P(x, \cdot)$ n'est ni discrète ni absolument continue par rapport à la mesure de Lebesgue et sa fonction de répartition F_x est

$$F_x(y) = \int_{-\infty}^{y} P(x, ds) = p\mathbf{1}_{x \leq y} + (1-p)\Phi(y),$$

où Φ désigne la fonction de répartition de la gaussienne standard (voir Figure 2.1). C'est ce genre de noyau qui apparaîtra dans l'algorithme de Metropolis en Section 2.2.1.

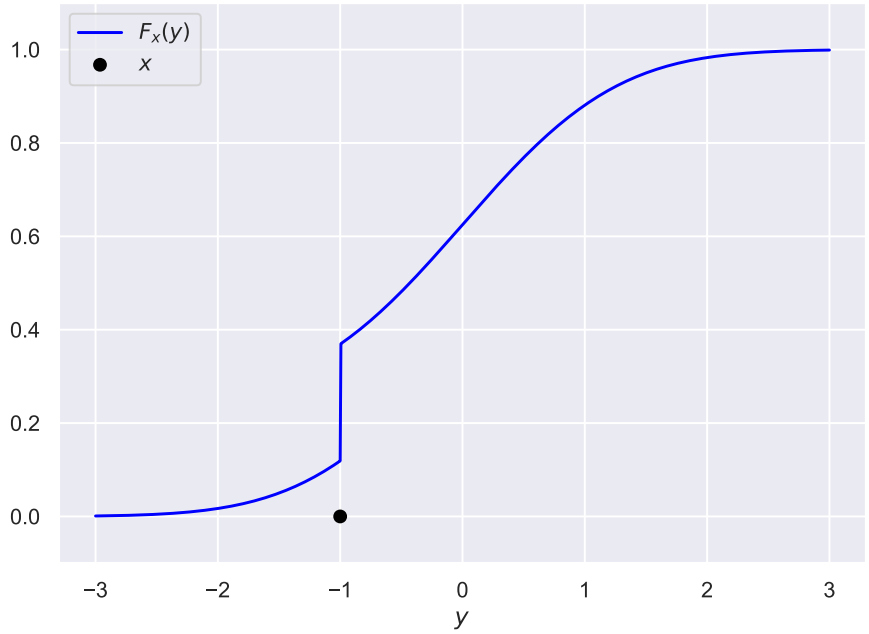

FIGURE 2.1 – Fonction de répartition $F_x(y)$ du noyau $P(x, dy) = p\delta_x(dy) + (1-p)f(y)dy$.

Définition 2.2 (Chaîne de Markov)
Soit μ_0 une mesure de probabilité sur $(E, \mathcal{E})$ et P un noyau de transition sur cet espace. On dit que la suite de variables aléatoires (X_n) est une chaîne de Markov (sous-entendu : homogène) de loi initiale μ_0 et de noyau P si $X_0 \sim \mu$ et si pour tout entier entier naturel n et tout B de $\mathcal{E}$:

$$\mathbb{P}(X_{n+1} \in B | X_n, \ldots, X_0) = \mathbb{P}(X_{n+1} \in B | X_n) = P(X_n, B),$$

c'est-à-dire

$$\text{Loi}(X_{n+1} | X_n, \ldots, X_0) = \text{Loi}(X_{n+1} | X_n) = P(X_n, \cdot).$$

Pour toute fonction test φ, on a donc

$$\mathbb{E}\left[\varphi(X_{n+1}) | X_n, \ldots, X_0\right] = \mathbb{E}\left[\varphi(X_{n+1}) | X_n\right] = \int P(X_n, dy)\varphi(y).$$

L'homogénéité de la chaîne signifie que le noyau P ne dépend pas de n, ce qui se traduit encore par

$$\text{Loi}(X_{n+1} | X_n = x) = \text{Loi}(X_1 | X_0 = x) = P(x, \cdot).$$

Bien entendu, dans le cas où E est dénombrable, nous retrouvons la notion usuelle de chaîne de Markov de matrice de transition $P = [p(x, y)]$: pour toute séquence $(x_0, \ldots, x_{n+1})$ telle que $\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n) > 0$, on a

$$\mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n, \ldots, X_0 = x_0) = \mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n) = p(x_n, x_{n+1}).$$

Rappelons qu'une interprétation de la propriété de Markov est que le futur ne dépend du passé que par l'état présent, ou encore : sachant l'état présent, passé et futur sont indépendants.

Remarque : Récurrence aléatoire. Soit μ_0 une mesure de probabilité sur $(E, \mathcal{E})$ et P un noyau de transition sur cet espace. Il existe une suite de variables aléatoires $(W_n)_{n\geq 1}$ i.i.d. à valeurs dans un espace mesurable $(F, \mathcal{F})$, indépendante de $X_0 \sim \mu_0$, et une fonction $h : E \times F \to E$ mesurable telles que la suite de variables (X_n) définie par

$$\forall n \geq 0, \qquad X_{n+1} = h(X_n, W_{n+1}) \tag{2.2}$$

soit une chaîne de Markov de loi initiale μ_0 et de noyau de transition P. Au vu du Lemme 1.3 d'inversion, on peut même choisir les $(W_n)_{n\geq 1}$ de loi uniforme sur $[0, 1]$. L'homogénéité de la chaîne se lit sur le fait que la fonction h ne dépend pas de n.

D'un point de vue numérique, cette représentation est très concrète : pour obtenir une réalisation de la chaîne, il suffit de savoir simuler selon μ_0, selon la loi des aléas W_n et, pour tout couple (x, w), de savoir calculer $h(x, w)$.

Exemples :

1. Cas discret : considérons $E = \mathbb{N}$ et P une matrice de transition P. Sachant $X_n = x$, nous savons que la loi de X_{n+1} est $P(x, \cdot)$. Si l'on note F_x la fonction de répartition associée, F_x^{-1} son inverse généralisée et W_{n+1} de loi uniforme sur $[0, 1]$, nous avons vu que $X_{n+1} \overset{\text{loi}}{=} F_x^{-1}(W_{n+1})$. On peut donc considérer la suite $(W_n)_{n\geq 1}$ i.i.d. de loi uniforme, indépendante de X_0, et $h(x, w) := F_x^{-1}(w)$.
2. Noyau gaussien : sachant $X_n = x$, on peut écrire $X_{n+1} = x + W_{n+1}$, i.e. considérer la suite $(W_n)_{n\geq 1}$ i.i.d. de loi gaussienne standard et $h(x, w) := x + w$.
3. Noyau gaussien normalisé : on peut considérer la suite $(W_n)_{n\geq 1}$ i.i.d. de loi gaussienne standard, et $h(x, w) := rx + \sqrt{1 - r^2}w$.
4. Mélange : on peut considérer la suite $(W_n)_{n\geq 1}$ i.i.d. avec $W_n = (B_n, G_n) \sim \mathcal{B}(p) \otimes \mathcal{N}(0, 1)$ et $h(x, w) := h(x, (b, g)) = bx + (1 - b)g$.

Soient μ une mesure de probabilité, P et Q des noyaux de transition et φ une fonction test. Si E est un espace d'états dénombrable, l'usage est de considérer que μ est un vecteur ligne, P et Q des matrices et φ un vecteur colonne. Les actions respectives d'un noyau sur une mesure (à gauche) et une fonction (à droite) correspondent tout simplement à des multiplications matricielles. Les notations suivantes en sont une généralisation naturelle : il suffit de remplacer les sommes par des intégrales.

Notations :

— $\mu\varphi = \mu(\varphi) = \int \mu(dx)\varphi(x) = \mathbb{E}[\varphi(X)]$ lorsque $X \sim \mu$.

— μP est la mesure de probabilité définie par

$$(\mu P)(dy) = \int \mu(dx) P(x, dy),$$

c'est-à-dire

$$(\mu P)(\varphi) = \iint \mu(dx) P(x, dy) \varphi(y).$$

Si (X_n) est une chaîne de Markov de loi initiale μ et de noyau P, ceci se traduit par

$$\text{Loi}(X_1) = \mu P \qquad \text{et} \qquad \mathbb{E}[\varphi(X_1)] = (\mu P)(\varphi).$$

— $P\varphi$ est la fonction définie par

$$(P\varphi)(x) = \int P(x, dy)\varphi(y) = (\delta_x P)(\varphi) = \mathbb{E}_x[\varphi(X_1)].$$

Par conséquent, nous pouvons écrire

$$(\mu P)(\varphi) = \mu(P\varphi) = \mu P\varphi.$$

et en particulier, pour tout couple $(x, B) \in E \times \mathcal{E}$,

$$P(x, B) = (P\mathbf{1}_B)(x) = \delta_x P\mathbf{1}_B.$$

— PQ est le noyau de transition défini par

$$(PQ)(x, dy) = \int P(x, dx')Q(x', dy),$$

c'est-à-dire

$$\mu(PQ)(\varphi) = \iiint \mu(dx)P(x, dx')Q(x', dy)\varphi(y).$$

Par récurrence, ceci permet de définir le noyau P^n pour tout $n \in \mathbb{N}$, avec la convention selon laquelle P^0 est le noyau identité, i.e. $P^0(x, dy) = \delta_x(dy)$.

Remarques :

1. Dans la suite, afin d'alléger les notations, nous n'écrirons pas systématiquement des intégrales multiples lorsque l'intégration se fait par rapport à plusieurs variables.
2. Si $X \sim \mu$, la loi μP est celle de la variable $Y = h_P(X, W)$ obtenue par application de la dynamique aléatoire (2.2) associée à P. Par conséquent, μPQ est la loi de la variable Z obtenue après avoir appliqué la dynamique associée à P **puis** celle associée à Q.
3. **Chapman-Kolmogorov.** Si nous revenons au contexte d'une chaîne de Markov (X_n) de loi initiale μ_0 et de noyau P, on en déduit que si l'on note μ_n la loi de X_n, alors $\mu_n = \mu_0 P^n$. Ainsi, pour toute fonction test φ, on a

$$\mathbb{E}[\varphi(X_n)] = \mu_0 P^n \varphi.$$

Exemples :

1. Cas discret : μ_0 est un vecteur ligne, P^n est la puissance n-ème de P et le vecteur ligne μ_n n'est rien d'autre que le produit matriciel $\mu_n = \mu_0 P^n$.
2. Noyau gaussien : Si $X_0 = x$, alors μ_n est la loi normale $\mathcal{N}(x, n)$. Plus généralement, si X_0 a pour loi μ_0, alors $\mu_n = \mu_0 * \mathcal{N}(0, n)$ où $*$ désigne le produit de convolution.
3. Noyau gaussien normalisé : Si $X_0 = x$, alors μ_n est la loi normale $\mathcal{N}(r^n x, 1 - r^{2n})$. Plus généralement, si X_0 a pour loi μ_0, alors $\mu_n = r^n \mu_0 * \mathcal{N}(0, 1 - r^{2n})$.
4. Mélange : Si $X_0 = x$, alors

$$\mu_n = p^n \delta_x + (1 - p^n)\mathcal{N}(0, 1).$$

Plus généralement, si X_0 a pour loi μ_0, alors μ_n est un mélange entre la loi initiale μ_0 et la gaussienne standard :

$$\mu_n = p^n \mu_0 + (1 - p^n)\mathcal{N}(0, 1).$$

2.1.2 Invariance et réversibilité

La notion de loi invariante pour une chaîne de Markov à espace d'états général est la même que dans le cas dénombrable.

Définition 2.3 (Loi invariante)
La mesure de probabilité π est invariante, ou stationnaire, ou d'équilibre pour le noyau de transition P si $\pi P = \pi$. En particulier, si (X_n) est une chaîne de Markov de noyau P et de loi initiale $\mu_0 = \pi$, alors $\mu_n = \pi$ pour tout n.

Plus généralement, nous dirons qu'une mesure λ est invariante si $\lambda P = \lambda$. Un exemple est donné ci-dessous pour le noyau gaussien et la mesure de Lebesgue.

Exemples :

1. Cas discret : Si P est irréductible, c'est-à-dire que pour tout couple de points (x, y) on peut aller de x à y sur le graphe de transition, ou plus formellement :

$$\forall (x,y) \in E \times E,\ \exists n = n(x,y) \text{ tel que } P^n(x,y) > 0, \tag{2.3}$$

   et récurrente positive, c'est-à-dire que $\mathbb{E}[T_x^*] < \infty$ pour tout x en notant

$$T_x^* := \min\{n \geq 1,\ X_n = x\} \in \mathbb{N} \cup \{\infty\}, \tag{2.4}$$

   alors il existe une unique loi invariante π et celle-ci s'écrit $\pi(x) = 1/\mathbb{E}[T_x^*]$. Rappelons aussi que si E est fini, il existe toujours au moins une loi invariante et qu'elle est unique si la chaîne est irréductible.

2. Noyau gaussien : Nous avons vu que $X_1 = X_0 + W_1$ avec W_1 gaussienne standard indépendante de X_0. Supposons π invariante, ce qui signifie que si $X_0 \sim \pi$ alors $X_1 \sim \pi$. Notons alors $\Phi_\pi(t) := \mathbb{E}[e^{itX_0}]$ la fonction caractéristique de la loi π. Dire que $X_1 \sim \pi$ revient à dire que pour tout réel t, on a $\Phi_\pi(t) = e^{-t^2/2}\Phi_\pi(t)$ donc $\Phi_\pi(t) = 0$ pour tout $t \neq 0$. Or, en tant que fonction caractéristique, Φ_π est continue sur $\mathbb{R}$ et vérifie $\Phi_\pi(0) = 1$, d'où une contradiction. Le noyau gaussien n'admet donc pas de loi invariante. En revanche, la mesure de Lebesgue $\lambda(dx) = dx$ est invariante puisque, pour tout x,

$$(\lambda P)(dx) = \int_y \lambda(dy)P(y,dx) = dx \int_y f(x-y)dy = dx \int_y \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-y)^2}{2}} dy = dx = \lambda(dx).$$

   C'est un peu l'équivalent en espace continu de la marche aléatoire sur $\mathbb{Z}$: il n'y a pas de loi invariante, mais la mesure de comptage sur $\mathbb{Z}$ est invariante.

3. Noyau gaussien normalisé : Cette fois $X_1 = rX_0 + \sqrt{1-r^2}W$ donc π est invariante si et seulement si $\Phi_\pi(t) = e^{-(1-r^2)t^2/2}\Phi_\pi(rt)$ pour tout réel t donc la fonction $\psi(t) := \Phi_\pi(t)e^{t^2/2}$ est continue sur $\mathbb{R}$, vérifie $\psi(0) = 1$ et $\psi(rt) = \psi(t)$ pour tout réel t. On en déduit que $\psi(r^n t) = \psi(t)$ pour tout n et par continuité en 0 il s'ensuit que $\psi(t) = 1$ pour tout réel t, ce qui signifie que la loi normale standard est la seule loi invariante.

4. Mélange : Nous avons $X_1 = B_1X_0 + (1-B_1)W_1$ avec les variables B_1, X_0 et W_1 indépendantes. En conditionnant par B_1, la fonction caractéristique d'une loi stationnaire π doit donc vérifier, pour tout réel t, $\Phi_\pi(t) = p\Phi_\pi(t) + (1-p)e^{-t^2/2}$, et la loi normale standard est à nouveau la seule loi invariante.

La notion de réversibilité est classique en physique statistique et interviendra lors de l'étude de l'algorithme de Metropolis. Elle se définit comme suit.

Définition 2.4 (Réversibilité)
Soit P un noyau de transition et π une mesure de probabilité. On dit que P est π-réversible, ou que π est réversible pour P, si

$$\forall (x,y) \in E \times E, \qquad \pi(dx)P(x,dy) = \pi(dy)P(y,dx). \tag{2.5}$$

Une autre façon de le dire : les mesures de probabilité $\mu(dx,dy) := \pi(dx)P(x,dy)$ et $\nu(dx,dy) := \pi(dy)P(y,dx)$ sur $E \times E$ sont les mêmes, i.e. pour toute fonction test $\varphi : E \times E \to \mathbb{R}$, on doit avoir

$$\iint \pi(dx)P(x,dy)\varphi(x,y) = \iint \pi(dy)P(y,dx)\varphi(x,y). \tag{2.6}$$

Plus généralement, une mesure λ est dite réversible pour P si (2.5) est vérifié avec λ en lieu et place de π.

Remarques :

1. Bien entendu, il suffit de vérifier (2.5) pour $x \neq y$.
2. S'il existe une mesure μ sur E par rapport à laquelle π et P sont absolument continus, nous adoptons l'abus de notation $\pi(dx) = \pi(x)\mu(dx)$, c'est-à-dire que l'on confond la loi π et sa densité. Quant au noyau, nous avons déjà convenu de noter $P(x,dy) = p(x,y)\mu(dy)$. Dans ce cas, P est π-réversible si et seulement si, pour $\mu \otimes \mu$ presque tout couple (x,y),

   $$\pi(x)p(x,y) = \pi(y)p(y,x), \tag{2.7}$$

   appelées **équations d'équilibre détaillé**.

La réversibilité assure l'invariance.

Lemme 2.5 (Réversibilité $\Longrightarrow$ Invariance)
Si π est réversible pour P, alors π est invariante pour P.

Preuve : Il suffit de montrer que $(\pi P)(\varphi) = \pi(\varphi)$ pour toute fonction test φ, or le Théorème de Fubini permet d'écrire

$$(\pi P)(\varphi) = \iint \pi(dx)P(x,dy)\varphi(y) = \int \left(\int P(y,dx) \right) \pi(dy)\varphi(y) = \int \pi(dy)\varphi(y) = \pi(\varphi).$$

■

Interprétation : Plaçons-nous pour simplifier sur un espace d'états dénombrable et considérons une chaîne (X_n) de noyau P supposé π-invariant, donc pour tout état x

$$\pi(x) = \sum_y \pi(y)p(y,x),$$

ce que l'on peut encore écrire

$$\sum_y \pi(x)p(x,y) = \sum_y \pi(y)p(y,x).$$

Si l'on interprète $\pi(x)$ comme la masse en x, l'équation précédente signifie que lorsque la chaîne est à l'équilibre, tout ce qui part de x est égal à tout ce qui arrive en x. Si l'on suppose que P est π-réversible, c'est-à-dire

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

alors, à l'équilibre, tout ce qui part de x pour aller en y est égal à tout ce qui part de y pour aller en x, d'où le nom d'équilibre détaillé.

Le terme réversible est à comprendre au sens de réversibilité par rapport au temps et est explicité par le résultat suivant.

Lemme 2.6 (Réversibilité temporelle)
Si (X_n) est une chaîne de Markov de loi initiale π et de noyau P supposé π-réversible, alors

$$\text{Loi}(X_0, \ldots, X_n) = \text{Loi}(X_n, \ldots, X_0).$$

Preuve : De façon générale, si (X_n) est une chaîne de Markov de loi initiale μ_0 et de noyau P, la loi du $(n+1)$-uplet $(X_0, \ldots, X_n)$ est

$$\mu_0(dx_0)P(x_0, dx_1) \ldots P(x_{n-1}, dx_n).$$

Si l'on suppose de plus que P est π-réversible et que la chaîne part sous la loi π, alors

$$\pi(dx_0)P(x_0, dx_1) \ldots P(x_{n-1}, dx_n) = \pi(dx_n)P(x_n, dx_{n-1}) \ldots P(x_1, dx_0).$$

Par conséquent, pour toute fonction test φ,

$$\mathbb{E}\left[\varphi(X_0, \ldots, X_n)\right] = \int \varphi(x_0, \ldots, x_n)\pi(dx_0)P(x_0, dx_1) \ldots P(x_{n-1}, dx_n)$$

devient

$$\mathbb{E}\left[\varphi(X_0, \ldots, X_n)\right] = \int \varphi(x_0, \ldots, x_n)\pi(dx_n)P(x_n, dx_{n-1}) \ldots P(x_1, dx_0),$$

ou encore, puisque les variables d'intégration sont muettes,

$$\mathbb{E}\left[\varphi(X_0, \ldots, X_n)\right] = \int \varphi(x_n, \ldots, x_0)\pi(dx_0)P(x_0, dx_1) \ldots P(x_{n-1}, dx_n) = \mathbb{E}\left[\varphi(X_n, \ldots, X_0)\right],$$

ce qui est exactement dire que

$$\text{Loi}(X_0, \ldots, X_n) = \text{Loi}(X_n, \ldots, X_0),$$

d'où le nom de réversibilité : à l'équilibre, la loi jointe reste la même par renversement du temps. ■

Exemples :

1. **Critère de Kolmogorov.** Soit P une matrice de transition irréductible sur E fini : elle admet donc une unique loi invariante π et celle-ci charge tous les états de E. L'existence d'une loi réversible peut en fait se lire directement sur le graphe de transition et ne nécessite pas la connaissance de celle-ci (voir par exemple [4]) : P admet une mesure de probabilité réversible si et seulement si pour tout $n \geq 0$ et toute séquence $(x_0, \ldots, x_n)$ on a, en notant $x_{n+1} = x_0$,
$$\prod_{i=0}^{n} p(x_i, x_{i+1}) = \prod_{i=0}^{n} p(x_{i+1}, x_i),$$
c'est-à-dire que la probabilité de transition sur un cycle ne dépend pas du sens de parcours de celui-ci. Si en général cette propriété n'est pas pratique pour prouver la réversibilité d'une chaîne, elle peut l'être en revanche pour prouver sa négation.
2. Marche aléatoire sur un graphe. Pour un graphe non orienté $(E, \mathcal{A})$ où E désigne l'ensemble des sommets et $\mathcal{A}$ l'ensemble des arêtes [4], on suppose que tout sommet x a un ensemble de voisins $\mathcal{V}(x) = \{y \in E, (x, y) \in \mathcal{A}\}$ non vide et fini. La marche aléatoire est la chaîne de Markov qui consiste, à chaque pas de temps, à passer de x à y en choisissant y au hasard équiprobable parmi les voisins de x, i.e. $p(x, y) = |\mathcal{V}(x)|^{-1}\mathbf{1}_{\mathcal{V}(x)}(y)$. On vérifie facilement que la mesure $\mu = \sum_{x \in E} |\mathcal{V}(x)|\delta_x$ est réversible. On peut la normaliser si et seulement si E est fini, auquel cas on obtient une loi réversible. Celle-ci est l'unique loi invariante si et seulement si le graphe est connexe.

4. Une arête peut relier un sommet à lui-même.

3. Noyau gaussien : Nous avons vu qu'il n'y a pas de loi invariante, mais aussi que la mesure de Lebesgue λ est invariante. Elle est en fait réversible puisque

$$\lambda(dx)P(x,dy) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-x)^2}{2}} dxdy = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-y)^2}{2}} dxdy = \lambda(dy)P(y,dx).$$

4. Noyau gaussien normalisé : Pour prouver que ce noyau est réversible par rapport à la loi normale standard, puisque tout est absolument continu par rapport à la mesure de Lebesgue, il suffit de vérifier (2.7) avec $\pi(x) = f(x)$ et

$$p(x,y) = \frac{1}{\sqrt{2\pi(1-r^2)}} e^{-\frac{(y-rx)^2}{2(1-r^2)}},$$

ce qui ne pose aucun problème.

5. Mélange : Cette fois le noyau P, donné en (2.1), n'est pas à densité puisqu'il s'écrit

$$P(x,dy) = p\delta_x(dy) + (1-p)f(y)dy.$$

Pour établir la réversibilité, il suffit de vérifier (2.6) :

$$\iint \pi(dx)P(x,dy)\varphi(x,y) = \iint \pi(dy)P(y,dx)\varphi(x,y).$$

Or, puisque $\pi(dx) = f(x)dx$,

$$\iint \pi(dx)P(x,dy)\varphi(x,y) = p\int \varphi(x,x)f(x)dx + (1-p)\iint \varphi(x,y)f(x)f(y)dxdy,$$

et l'égalité est claire. On peut le faire sans les fonctions tests : il suffit d'établir

$$\forall x \neq y, \qquad \pi(dx)P(x,dy) = \pi(dy)P(y,dx).$$

Or $\pi(dx) = f(x)dx$ est absolument continue par rapport à la mesure de Lebesgue et, pour $x \neq y$, le noyau P l'est aussi : $P(x,dy) = (1-p)f(y)dy$. La réversibilité est alors claire.

2.1.3 Irréductibilité, apériodicité

Si la notion d'invariance se définit de la même façon en espace d'états général qu'en espace d'états dénombrable, tout se complique pour la notion d'irréductibilité. Rappelons que si E est dénombrable, la chaîne est irréductible si, quel que soit le point de départ x, il est possible d'aller en un point quelconque y en un certain nombre d'étapes, c'est-à-dire qu'en notant

$$T_y^* := \min\{n \geq 1,\ X_n = y\} \in \mathbb{N}^* \cup \{\infty\},$$

on a

$$\forall (x,y) \in E \times E, \qquad \mathbb{P}_x\left(T_y^* < \infty\right) > 0.$$

Supposons maintenant $E = \mathbb{R}$ et un noyau absolument continu par rapport à la mesure de Lebesgue (par exemple le noyau gaussien normalisé), alors il est clair que

$$\forall (x,y) \in E \times E, \qquad \mathbb{P}_x\left(T_y^* < \infty\right) = 0.$$

Bien sûr, le problème vient de ce que tout point $\{y\}$ est de mesure de Lebesgue nulle. L'idée est donc de remplacer le singleton $\{y\}$ par un ensemble B de mesure positive. Encore faut-il préciser de quelle mesure on parle.

Notation : Pour tout ensemble B de $\mathcal{E}$, on note

$$T_B^* := \min\{n \geq 1,\ X_n \in B\} \in \mathbb{N}^* \cup \{\infty\}.$$

Ainsi

$$\mathbb{P}_x\left(T_B^* < \infty\right) > 0 \iff \mathbb{P}_x\left(\exists n \geq 1,\ X_n \in B\right) > 0 \iff \sum_{n=1}^{\infty} \mathbb{P}_x\left(X_n \in B\right) > 0,$$

ou encore

$$\mathbb{P}_x\left(T_B^* < \infty\right) > 0 \iff \exists n = n(x, B) \geq 1,\ \mathbb{P}_x\left(X_n \in B\right) > 0. \tag{2.8}$$

Définition 2.7 (Irréductibilité)
Soit ν une mesure sur $(E, \mathcal{E})$. La chaîne de Markov (X_n) est dite ν-irréductible si pour tout $B \in \mathcal{E}$ tel que $\nu(B) > 0$, on a

$$\forall x \in E, \qquad \mathbb{P}_x\left(T_B^* < \infty\right) > 0.$$

Si P désigne le noyau de la chaîne (X_n), on dira de façon équivalente que P est ν-irréductible, c'est-à-dire

$$\forall x \in E,\ \exists n \geq 1, \qquad P^n(x, B) > 0.$$

On dit parfois qu'une chaîne (X_n), ou son noyau P, est irréductible s'il existe une mesure ν pour laquelle (X_n) est ν-irréductible.

Exemples :

1. Cas discret : Par conséquent, si E est dénombrable, ce que l'on appelle usuellement une chaîne irréductible est, avec la définition précédente, une chaîne ν-irréductible lorsque ν est la mesure de comptage sur E.
2. Noyau gaussien : Prenons $\nu = \lambda$ mesure de Lebesgue sur $\mathbb{R}$, alors pour tout borélien B tel que $\lambda(B) > 0$, on a

   $$P(x, B) = \mathbb{P}(\mathcal{N}(x, 1) \in B) = \int_B f(y - x)dx > 0,$$

   donc la chaîne est ν-irréductible. Quel que soit le point de départ x, la probabilité que la chaîne atteigne un borélien non négligable en un seul coup est strictement positive.
3. Noyau gaussien normalisé : La conclusion est la même car, pour tout borélien non négligeable pour la mesure de Lebesgue,

   $$P(x, B) = \mathbb{P}(\mathcal{N}(rx, 1 - r^2) \in B) > 0.$$

4. Mélange : Même conclusion puisque

   $$P(x, B) = p\mathbf{1}_{x \in B} + (1 - p)\mathbb{P}(\mathcal{N}(0, 1) \in B) \geq (1 - p)\mathbb{P}(\mathcal{N}(0, 1) \in B) > 0.$$

Remarques :

1. Revenons sur le cas discret et considérons sur $E = \mathbb{N}^*$ la matrice de transition P définie par $p(1, 1) = 1$ et, pour tout $x > 1$, $p(x, x + 1) = 1 - p(x, 1) = 1 - \frac{1}{x^2}$ (voir Figure 2.2). L'état $\mathbf{1}$ étant absorbant, il est clair qu'une chaîne (X_n) de matrice de transition P n'est pas irréductible au sens usuel du terme, c'est-à-dire qu'elle n'est pas ν-irréductible pour ν la mesure de comptage sur $\mathbb{N}^*$. En revanche, elle est ν-irréductible pour $\nu = \delta_1$. C'est du reste la seule mesure pour laquelle elle l'est.

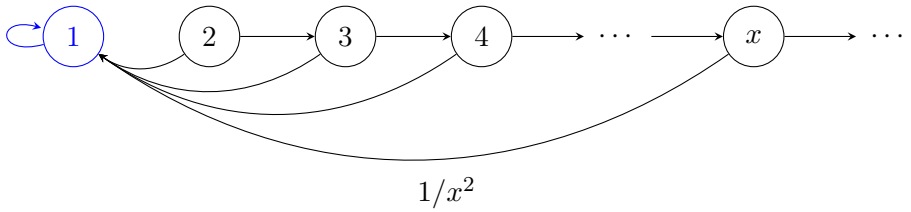

FIGURE 2.2 – Graphe de transition d'une chaîne non irréductible au sens usuel mais δ_1-irréductible.

2. Si une chaîne est ν-irréductible, elle est μ-irréductible pour toute mesure μ absolument continue par rapport à ν.

Le résultat suivant donne un lien entre les mesures intervenant dans l'irréductibilité et dans la stationnarité.

Lemme 2.8
Si la chaîne de Markov (X_n) est ν-irréductible et admet π pour loi stationnaire alors ν est absolument continue par rapport à π.

Preuve : Soit B tel que $\nu(B) > 0$, alors par (2.8) il existe un couple (m, n) tel que $P^n(x, B) \geq \frac{1}{m}$. Par conséquent

$$E = \left\{ x \in E,\ \exists (m,n),\ P^n(x,B) \geq \frac{1}{m} \right\} = \bigcup_{m,n} \left\{ x \in E,\ P^n(x,B) \geq \frac{1}{m} \right\}.$$

Or, par sous-σ-additivité, on sait que

$$1 = \pi(E) \leq \sum_{m,n} \pi\left(x \in E,\ P^n(x,B) \geq \frac{1}{m} \right).$$

Il existe donc un couple (m, n) et un ensemble mesurable $C = C(m, n)$ vérifiant $\pi(C) > 0$ tel que $P^n(x, B) \geq 1/m$ pour tout x de C. Ceci implique, par stationnarité de π,

$$\pi(B) = (\pi P^n)(B) = \int_E \pi(dx) P^n(x,B) \geq \int_C \pi(dx) P^n(x,B) \geq \frac{\pi(C)}{m} > 0.$$

Pour tout ensemble mesurable B, nous avons ainsi établi que $\nu(B) > 0$ implique $\pi(B) > 0$, ce qui est exactement dire que ν est absolument continue par rapport à π.

■

Le concept d'apériodicité sur lequel nous nous baserons existe déjà en espace d'états dénombrable (voir par exemple [18], Théorème 1.8.4, ou [4], Chapitre 5). La plupart des chaînes que nous construirons par méthodes MCMC auront cette propriété. Rappelons que le phénomène de périodicité nuit à la convergence en loi, mais pas à l'équivalent de la loi des grands nombres (appelé Théorème ergodique pour les chaînes de Markov).

Définition 2.9 (Apériodicité)
La chaîne de Markov (X_n) de noyau de transition P est dite apériodique s'il n'existe pas d'entier $d \geq 2$ et de partition de $E = \cup_{j=1}^{d} E_j$ tels que

$$\begin{cases} P(x, E_{j+1}) = 1 & \forall x \in E_j,\ \forall j \in \{1, \ldots, d-1\} \\ P(x, E_1) = 1 & \forall x \in E_d. \end{cases}$$

Bien entendu, dans la partition ci-dessus, aucun E_j n'est vide.

Exemples :

1. Cas discret : si E est dénombrable et la chaîne irréductible (pour la mesure de comptage), on peut montrer que la définition précédente revient à dire que pour tout couple (x, y) il existe $N = N(x, y)$ tel que pour tout $n \geq N$, on a $P^n(x, y) > 0$. Si par exemple $E = \{0, 1\}$ et $p(0, 1) = p(1, 0) = 1$, la chaîne n'est pas apériodique.
2. Noyau alterné : Soit $(W_n)_{n\geq 1}$ une suite de variables gaussiennes standards et la chaîne de Markov (X_n) définie sur $E = \mathbb{R}$ par $X_0 = x$ et la récurrence aléatoire

$$X_{n+1} = h(X_n, W_{n+1}) := (-1)^{\mathbf{1}_{X_n \geq 0}} |X_n + W_{n+1}|.$$

   La partition $\mathbb{R} = \mathbb{R}_+ \cup \mathbb{R}_-^*$ montre que la chaîne n'est pas apériodique.

Un critère suffisant d'irréductibilité et d'apériodicité est donné par le résultat suivant, que l'on appliquera en général avec $\mu = \pi$ la loi stationnaire de la chaîne.

Lemme 2.10 (Condition suffisante d'irréductibilité et d'apériodicité)
Soit P un noyau de transition et μ une mesure de probabilité vérifiant : pour tout x de E et tout $B \in \mathcal{E}$ tel que $\mu(B) > 0$, on a $P(x, B) > 0$. Alors la chaîne de noyau P est μ-irréductible et apériodique.

Preuve : Au vu de la Définition 2.7, l'irréductibilité est claire puisque, pour tout $B \in \mathcal{E}$ tel que $\mu(B) > 0$, on a

$$\forall x \in E, \qquad \mathbb{P}_x \left(T_B^* < \infty\right) \geq \mathbb{P}_x \left(T_B^* = 1\right) = P(x, B) > 0.$$

Ceci permet de montrer facilement l'apériodicité en raisonnant par l'absurde. Supposons en effet qu'il existe $d \geq 2$ et une partition $E = \cup_{j=1}^d E_j$ comme en Définition 2.9. Pour tout x de E_1 (qui n'est pas vide par hypothèse), on doit avoir $P(x, E_2) = 1$. Puisque $P(x, \cdot)$ est une mesure de probabilité, ceci implique

$$P(x, E_1) = P(x, E_3) = \cdots = P(x, E_d) = 0.$$

Puisque $\mu(B) > 0$ implique $P(x, B) > 0$, la contraposée donne

$$\mu(E_1) = \mu(E_3) = \cdots = \mu(E_d) = 0.$$

Puisque μ est une probabilité sur E, il en découle que $\mu(E_2) = 1 > 0$. Mais alors, si x appartient à E_2 (qui n'est pas vide non plus par hypothèse), on aura nécessairement $P(x, E_2) > 0$, ce qui contredit $P(x, E_3) = 1$ car $P(x, \cdot)$ est une mesure de probabilité et les ensembles E_2 et E_3 sont disjoints.

■

Exemple : Pour le noyau gaussien, le noyau gaussien normalisé et le noyau mélange, il suffit de considérer $\mu = \mathcal{N}(0, 1)$. Les calculs ci-dessus montrent que pour tout borélien B tel $\mu(B) > 0$, on a bien $P(x, B) > 0$, donc ces noyaux sont apériodiques.

Exercice : Montrer que le résultat du Lemme 2.10 se généralise comme suit : s'il existe $n_0 \geq 1$ et une mesure de probabilité μ tels que pour tout x de E et tout $B \in \mathcal{E}$ tel que $\mu(B) > 0$, on ait $P^{n_0}(x, B) > 0$, alors la chaîne de noyau P est μ-irréductible et apériodique.

2.1.4 Résultats asymptotiques

Nous nous intéressons à la convergence en temps long de la loi de la chaîne (X_n). Si $X_0 = x$, cette loi est $\delta_x P^n = P^n(x, \cdot)$. Dans le cas d'un espace d'états dénombrable, on sait que si la chaîne est irréductible, apériodique et récurrente positive de loi stationnaire π, alors pour tout x de E on a

$$\sum_{y\in E} |P^n(x,y) - \pi(y)| \xrightarrow[n\to\infty]{} 0.$$

Dans ce cas, pour tout x de E et tout B de $\mathcal{E} = \mathcal{P}(E)$, il vient

$$|P^n(x,B) - \pi(B)| = \left|\sum_{y\in B} (P^n(x,y) - \pi(y))\right| \leq \sum_{y\in E} |P^n(x,y) - \pi(y)|,$$

donc

$$\sup_{B\in\mathcal{E}} |P^n(x,B) - \pi(B)| \leq \sum_{y\in E} |P^n(x,y) - \pi(y)| \xrightarrow[n\to\infty]{} 0.$$

Réciproquement, si l'on suppose que, pour tout x de E,

$$\sup_{B\in\mathcal{E}} |P^n(x,B) - \pi(B)| \xrightarrow[n\to\infty]{} 0,$$

alors en considérant $B_n^- := \{y \in E,\ P^n(x,y) \leq \pi(y)\}$ et $B_n^+ := \{y \in E,\ P^n(x,y) > \pi(y)\}$, on a

$$\sum_{y\in E} |P^n(x,y) - \pi(y)| = \left|P^n(x,B_n^+) - \pi(B_n^+)\right| + \left|P^n(x,B_n^-) - \pi(B_n^-)\right|,$$

donc

$$\sum_{y\in E} |P^n(x,y) - \pi(y)| \leq 2 \sup_{B\in\mathcal{E}} |P^n(x,B) - \pi(B)| \xrightarrow[n\to\infty]{} 0.$$

Sur un espace dénombrable, nous avons ainsi établi l'équivalence suivante :

$$\sum_{y\in E} |P^n(x,y) - \pi(y)| \xrightarrow[n\to\infty]{} 0 \iff \sup_{B\in\mathcal{E}} |P^n(x,B) - \pi(B)| \xrightarrow[n\to\infty]{} 0. \tag{2.9}$$

Contrairement à celle de gauche, la convergence de droite se généralise naturellement à tout espace d'états et est liée à la distance en variation totale.

Définition 2.11 (Distance en variation totale)
Si μ et ν sont deux mesures de probabilités sur $(E, \mathcal{E})$, on appelle distance en variation totale entre μ et ν la quantité

$$\|\mu - \nu\| = \sup_{B\in\mathcal{E}} |\mu(B) - \nu(B)|.$$

Remarques :

1. Puisque μ et ν sont des mesures de probabilité, $\mu(E \setminus B) - \nu(E \setminus B) = \nu(B) - \mu(B)$, donc on peut se passer de la valeur absolue dans la définition précédente :

$$\|\mu - \nu\| = \sup_{B\in\mathcal{E}} (\mu(B) - \nu(B)).$$

2. Les 3 axiomes de définition d'une distance sont clairement vérifiés, à savoir symétrie, séparation [5] et inégalité triangulaire. De plus, on a toujours $\|\mu - \nu\| \leq 1$.

5. Au sens où $\mu = \nu$ sur $(E, \mathcal{E})$ ssi $\mu(B) = \nu(B)$ pour tout $B \in \mathcal{E}$.

3. En analyse, l'usage est de considérer la distance en variation totale entre deux mesures signées finies comme le double de celle de la Définition 2.11.

4. Supposons $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ et considérons une suite (μ_n) de lois convergeant en variation totale vers la loi μ, c'est-à-dire $\|\mu_n - \mu\| \to 0$ quand $n \to \infty$. Rappelons que si (X_n) est une suite de variables aléatoires réelles avec $X_n \sim \mu_n$ pour tout n, (X_n) converge en loi vers $X \sim \mu$ ssi en tout point de continuité x de la fonction de répartition F_X, on a $F_{X_n}(x) \to F_X(x)$ lorsque $n \to \infty$. Puisque, pour tout réel x,

$$|F_{X_n}(x) - F_X(x)| = |\mu_n(]-\infty, x]) - \mu(]-\infty, x])| \leq \|\mu_n - \mu\|,$$

la convergence en variation totale implique la convergence en loi. La réciproque est fausse : il suffit de prendre $\mu_n = \delta_{1/n}$, $\mu = \delta_0$ et $B = \{0\}$ pour s'en convaincre. Ainsi, pour des suites de variables aléatoires réelles, la convergence en variation totale est plus forte que la convergence en loi.

5. Pour une "bonne" chaîne sur un espace d'états dénombrable (i.e. irréductible au sens usuel, apériodique et récurrente positive), nous pouvons ainsi affirmer que pour toute condition initiale $X_0 = x$, la loi de X_n tend vers la loi stationnaire π en variation totale :

$$\|P^n(x, \cdot) - \pi\| \xrightarrow[n\to\infty]{} 0.$$

Noter que ces conditions sont suffisantes mais pas nécessaires, comme le montre la matrice de transition P sur $E = \{0, 1\}$ définie par $p(0, 0) = p(1, 0) = 1$. Elle n'est pas irréductible, mais indécomposable.

La distance en variation totale possède de nombreuses propriétés, nous nous contentons d'en énoncer quelques-unes.

Propriétés 2.12
Dans ce qui suit, μ, ν et π sont des mesures de probabilité sur un espace $(E, \mathcal{E})$, P un noyau de transition sur cet espace et $\varphi : (E, \mathcal{E}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ une fonction mesurable.

1. *Fonctions tests : $\|\mu - \nu\| = \sup_{\varphi: E\to[0,1]} |\mu(\varphi) - \nu(\varphi)| = \frac{1}{2} \sup_{\varphi: E\to[-1,1]} |\mu(\varphi) - \nu(\varphi)|$.*
2. *Décroissance : $\|\mu P - \nu P\| \leq \|\mu - \nu\|$. En particulier, si π est stationnaire pour le noyau P, alors pour tout x de E, la suite $(\|P^n(x, \cdot) - \pi\|)_{n\geq 0}$ est décroissante.*
3. *Densité : Soit λ une mesure telle que $\mu = f \cdot \lambda$ et $\nu = g \cdot \lambda$, alors*

$$\|\mu - \nu\| = \frac{1}{2} \|f - g\|_{L^1(\lambda)} := \frac{1}{2} \int_E |f(x) - g(x)| \lambda(dx). \tag{2.10}$$

4. *Inégalité de couplage : Pour tout couple de variables aléatoires (X, Y) vérifiant $X \sim \mu$ et $Y \sim \nu$, on a*

$$\|\mu - \nu\| \leq \mathbb{P}(X \neq Y).$$

Preuve : On rappelle que toutes les mesures sont supposées σ-finies, ce qui permet d'appliquer le Théorème de Radon-Nikodym.

1. Soit λ une mesure dominant à la fois μ et ν, par exemple $\lambda = \mu + \nu$, notons $f = \frac{d\mu}{d\lambda}$, $g = \frac{d\nu}{d\lambda}$ et $A = \{x \in E,\ f(x) \geq g(x)\}$, alors

$$|\mu(\varphi) - \nu(\varphi)| = \left| \int_E \varphi(x) f(x) \lambda(dx) - \int_E \varphi(x) g(x) \lambda(dx) \right| = \left| \int_E \varphi(x)(f(x) - g(x)) \lambda(dx) \right|.$$

Puisque $0 \leq \varphi \leq 1$, cette quantité est maximale si et seulement si $\varphi = \mathbf{1}_A$ ou $\varphi = 1 - \mathbf{1}_A$, ce qui donne dans les deux cas

$$|\mu(\varphi) - \nu(\varphi)| = |\mu(A) - \nu(A)|\,,$$

et prouve la première égalité. Pour la seconde, il faut prendre $\varphi = \mathbf{1}_A - \mathbf{1}_{E\setminus A}$ ou son opposée, ce qui quoi qu'il en soit donne

$$|\mu(\varphi) - \nu(\varphi)| = 2\,|\mu(A) - \nu(A)|\,.$$

2. Pour $(x, A) \in E \times \mathcal{E}$, la fonction $\varphi(x) := P(x, A)$ est mesurable de E dans $[0, 1]$ donc le point précédent entraîne

$$|(\mu P)(A) - (\nu P)(A)| = |\mu(\varphi) - \nu(\varphi)| \leq \|\mu - \nu\|.$$

Ceci étant vrai pour tout A, il s'ensuit que $\|\mu P - \nu P\| \leq \|\mu - \nu\|$. La décroissance de la suite $(\|P^n(x, \cdot) - \pi\|)$ lorsque π est stationnaire découle alors de la formulation

$$\|P^n(x, \cdot) - \pi\| = \|\delta_x P^n - \pi P^n\|.$$

3. On applique le même raisonnement que ci-dessus, ce qui donne

$$\|\mu - \nu\| = \frac{1}{2}\left|\int_E (\mathbf{1}_A - \mathbf{1}_{E\setminus A})(x)(f(x) - g(x))\lambda(dx)\right| = \frac{1}{2}\int_E |f(x) - g(x)|\lambda(dx).$$

4. Si $X \sim \mu$ et $Y \sim \nu$ alors, pour tout $A \in \mathcal{E}$,

$$|\mu(A) - \nu(A)| = |\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)|\,,$$

or

$$\mathbb{P}(X \in A) = \mathbb{P}(X \in A, X = Y) + \mathbb{P}(X \in A, X \neq Y),$$

et la même décomposition appliquée à $\mathbb{P}(Y \in A)$ mène à

$$|\mu(A) - \nu(A)| = |\mathbb{P}(X \in A, X \neq Y) - \mathbb{P}(Y \in A, X \neq Y)|\,.$$

Les deux termes dans la valeur absolue étant entre 0 et $\mathbb{P}(X \neq Y)$, il s'ensuit que

$$|\mathbb{P}(X \in A, X \neq Y) - \mathbb{P}(Y \in A, X \neq Y)| \leq \mathbb{P}(X \neq Y),$$

et il ne reste plus qu'à prendre le supremum sur $A \in \mathcal{E}$ pour conclure.

■

Remarques :

1. Concernant le dernier point, on peut en fait prouver qu'il existe $X \sim \mu$ et $Y \sim \nu$ telles que $\|\mu - \nu\| = \mathbb{P}(X \neq Y)$.
2. Sur un espace d'états dénombrable, l'équation (2.10) appliquée à la mesure de comptage pour λ donne ainsi

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

ce qui permet de retrouver l'équivalence (2.9).

3. Lorsque $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, le premier point des Propriétés 2.12 reflète le fait que la convergence en variation totale est plus forte que la convergence en loi puisque dans la définition de la convergence en loi les fonctions tests sont supposées continues bornées, et non simplement mesurables bornées.

Exemples :

1. Espace d'états fini : Si la matrice P est irréductible et apériodique, alors 1 est valeur propre simple et c'est la seule valeur propre de module 1. Si l'on note r le maximum en module des autres valeurs propres, on sait que $0 \leq r < 1$ et que pour tout $r < \rho \leq 1$, il existe une constante $C = C(\rho)$ telle que
$$\|P^n(x, \cdot) - \pi\| \leq C\rho^n.$$
La quantité $(1 - r)$ est appelée le trou spectral.

2. Noyau gaussien normalisé : De façon générale, dans le cas à densité, on définit la distance de Hellinger entre μ et ν par
$$H(\mu, \nu) := \left\| \sqrt{f} - \sqrt{g} \right\|_{L^2(\lambda)} := \left(\int_E \left(\sqrt{f(x)} - \sqrt{g(x)} \right)^2 \lambda(dx) \right)^{\frac{1}{2}},$$
et on voit facilement que $\|\mu - \nu\| \leq H(\mu, \nu)$. De plus, si $\mu = \mathcal{N}(m_1, \sigma_1^2)$ et $\nu = \mathcal{N}(m_2, \sigma_2^2)$, alors on peut montrer que
$$H(\mu, \nu)^2 = 2 \left\{ 1 - \sqrt{2 \frac{\sigma_1 \sigma_2}{\sigma_1^2 + \sigma_2^2}} \exp\left(-\frac{(m_1 - m_2)^2}{4(\sigma_1^2 + \sigma_2^2)} \right) \right\}.$$
Dans le cas gaussien normalisé, ceci permet de voir que
$$\|P^n(x, \cdot) - \pi\| = \left\| \mathcal{N}\left(r^n x, 1 - r^{2n}\right) - \mathcal{N}(0, 1) \right\| \leq H\left(\mathcal{N}\left(r^n x, 1 - r^{2n}\right), \mathcal{N}(0, 1) \right),$$
et un calcul d'équivalent aboutit à
$$H\left(\mathcal{N}\left(r^n x, 1 - r^{2n}\right), \mathcal{N}(0, 1) \right) \sim \frac{|x| r^n}{2}. \tag{2.11}$$
Puisqu'on vérifie aussi que $H(\mu, \nu)^2 \leq 2\|\mu - \nu\|$, on conclut qu'il y a convergence vers l'équilibre à vitesse géométrique. Sans surprise, celle-ci est d'autant plus rapide que r est proche de 0 et le point de départ x proche de l'origine.

3. Mélange : Nous avons vu que $P^n(x, \cdot) = p^n \delta_x + (1 - p^n)\mathcal{N}(0, 1)$ et que $\pi = \mathcal{N}(0, 1)$ est invariante. Si l'on note λ_1 la mesure de Lebesgue sur la droite réelle, une mesure dominant à la fois $P^n(x, \cdot)$ et π est $\lambda := \delta_x + \lambda_1$. En notant toujours f la densité de la gaussienne standard, les densités associées sont les fonctions
$$\frac{dP^n(x, \cdot)}{d\lambda}(y) = p^n \mathbf{1}_{y=x} + (1 - p^n) f(y) \mathbf{1}_{y \neq x} \qquad \text{et} \qquad \frac{d\pi}{d\lambda}(y) = f(y) \mathbf{1}_{y \neq x}.$$
Il s'ensuit que
$$\|P^n(x, \cdot) - \pi\| = \frac{1}{2} \int_{\mathbb{R}} \left| \frac{dP^n(x, \cdot)}{d\lambda}(y) - \frac{d\pi}{d\lambda}(y) \right| \lambda(dy) = p^n, \tag{2.12}$$
et on retrouve une convergence à vitesse géométrique. Bien entendu, celle-ci est d'autant plus rapide que la probabilité de rester sur place est faible.

Dans l'étude des chaînes de Markov sur un espace d'états dénombrable, l'usage est d'exhiber des conditions suffisantes sous lesquelles il existe une unique probabilité invariante π vers laquelle il y a convergence en variation totale de la loi de X_n, et ce pour toute condition initiale $X_0 = x$. Comme nous l'avons dit précédemment, ceci est par exemple vrai si la chaîne est irréductible, apériodique et récurrente positive, auquel cas $\pi(x) = 1/\mathbb{E}_x[T_x^*]$.

Dans ce qui suit, le contexte est un peu différent, puisque nous supposons d'emblée que la chaîne admet une loi invariante π. Cette approche est motivée par le cadre spécifique des méthodes MCMC que nous présenterons ensuite. Nous admettrons le résultat suivant.

Théorème 2.13 (Convergence en variation totale)
S'il existe une mesure ν et une loi de probabilité π telles que la chaîne (X_n) soit ν-irréductible, apériodique et π-invariante, alors :

- *la chaîne est π-irréductible ;*
- *π est son unique loi invariante ;*
- *pour π presque toute condition initiale x, la loi de X_n tend vers π en variation totale :*

$$\|P^n(x,\cdot) - \pi\| \xrightarrow[n\to\infty]{} 0.$$

Remarques :

1. Il peut sembler étonnant qu'il n'y ait aucune hypothèse spécifique sur la mesure ν intervenant dans l'irréductibilité et qu'elle n'apparaisse pas dans le résultat. On a vu en Lemme 2.8 que si une chaîne (X_n) est ν-irréductible et π-invariante, alors $\nu \ll \pi$. Le résultat ci-dessus montre que la chaîne est en fait nécessairement π-irréductible. Ainsi, dans l'énoncé du Théorème 2.13, on considère souvent $\nu = \pi$.

2. Le fait que le résultat ne soit vrai que pour π-presque toute condition initiale peut se comprendre sur un cas discret, en l'occurrence celui déjà mentionné où $E = \mathbb{N}^*$, $p(1,1) = 1$ et, pour tout $x > 1$, $p(x, x+1) = 1 - p(x,1) = 1 - \frac{1}{x^2}$ (cf. Figure 2.3). La chaîne (X_n) est ν-irréductible pour $\nu = \delta_1$ et admet pour loi invariante $\pi = \delta_1$, ce qui peut par exemple se voir de façon élémentaire en résolvant le système linéaire $\pi P = \pi$. Elle est de plus apériodique : en effet, supposons qu'il existe $d \geq 2$ et une partition de $E = \cup_{j=1}^d E_j$. Sans perte de généralité, on peut supposer que l'état 1 appartient à E_1, mais alors le fait que $P(1, E_2) = 1$ implique que 1 appartient aussi à E_2, ce qui est absurde. Les conditions du Théorème 2.13 sont donc vérifiées. Or, pour toute condition initiale $X_0 = x \geq 2$, il est clair que $P^n(x,\cdot) = (1 - p_x(n))\delta_1 + p_x(n)\delta_{x+n}$ avec

$$p_x(n) := \mathbb{P}_x\left(\forall 1 \leq k \leq n,\ X_k = x + k\right) = \prod_{j=x}^{x+n-1}\left(1 - \frac{1}{j^2}\right) \xrightarrow[n\to\infty]{} p_x > 0,$$

   c'est-à-dire qu'avec probabilité p_x la chaîne part à l'infini tandis qu'avec probabilité $(1-p_x)$ elle finit en l'état 1. D'après la relation (2.10) des Propriétés 2.12, il vient

$$\|P^n(x,\cdot) - \pi\| = p_x(n) \xrightarrow[n\to\infty]{} p_x > 0.$$

   Il n'y a cependant aucune contradiction avec le résultat du Théorème 2.13 puisque si $X_0 = 1$ alors $P^n(x,\cdot) = \pi$ pour tout n.

3. Le phénomène précédent n'est pas propre aux espaces d'états dénombrables, comme le montre l'exercice suivant.

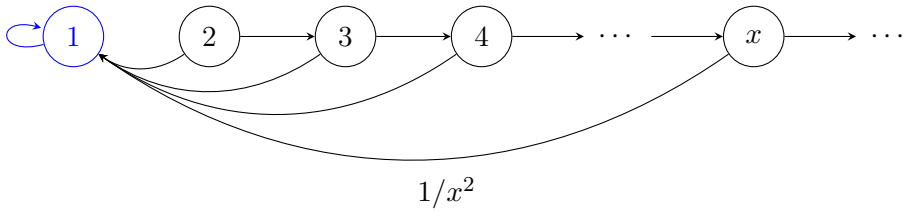

FIGURE 2.3 – Chaîne telle que $\|P^n(x,\cdot) - \pi\| \xrightarrow[n\to\infty]{} 0$ n'est vrai que pour $x = 1$.

Exercice : Soit $E = [0,1]$ muni de sa tribu borélienne et le noyau P défini comme suit : s'il existe $m \in \mathbb{N}^*$ tel que $x = 1/m$, alors

$$P(x,\cdot) = x^2 \mathrm{Unif}([0,1]) + (1-x^2)\delta_{1/(m+1)},$$

sinon $P(x,\cdot) = \mathrm{Unif}([0,1])$. Soit (X_n) une chaîne de noyau P. Etablir qu'elle admet $\pi = \mathrm{Unif}([0,1])$ comme loi invariante, qu'elle est π-irréductible et apériodique. Montrer que si $X_0 = 1/m$ avec $m \geq 2$, alors

$$\mathbb{P}\left(\forall n \in \mathbb{N},\ X_n = \frac{1}{m+n}\right) > 0,$$

et conclure.

Exemples :

1. Cas discret : on retrouve le fait que pour une chaîne irréductible (au sens usuel), la récurrence positive est équivalente à l'existence d'une loi stationnaire, auquel cas celle-ci est unique. Si de plus la chaîne est apériodique, alors il y a convergence en variation totale pour toute condition initiale.
2. Noyau gaussien : nous avons vu qu'il n'y a pas de loi invariante donc le Théorème 2.13 ne s'applique pas.
3. Noyau gaussien normalisé : ce noyau est irréductible pour la mesure de Lebesgue, apériodique et la loi gaussienne standard $\pi = \mathcal{N}(0,1)$ est invariante, donc pour π-presque tout x, c'est-à-dire pour presque tout x (puisque les mesures de Lebesgue et gaussienne standard sont équivalentes [6])
$$\|P^n(x,\cdot) - \pi\| \xrightarrow[n\to\infty]{} 0.$$
Comme nous l'avons vu en (2.11), ceci est en fait vrai pour toute condition initiale x et la vitesse de convergence est géométrique en $\mathcal{O}(|r|^n)$.
4. Noyau mélange : la conclusion est la même et nous avons même établi en (2.12) que, pour toute condition initiale, la vitesse de convergence est géométrique et égale à p^n.

Nous énonçons maintenant l'équivalent de la loi des grands nombres, appelé Théorème ergodique. Il n'y a aucune hypothèse d'apériodicité : ceci n'a rien de surprenant puisque tout éventuel phénomène de périodicité est tué par moyennisation trajectorielle.

Exemple : Reprenons l'exemple $E = \{0,1\}$ et $p(0,1) = p(1,0) = 1$. Clairement cette chaîne n'est pas apériodique. Elle est néanmoins irréductible (pour la mesure de comptage), d'unique loi stationnaire la loi uniforme π sur E. Pour toute condition initiale $x \in E$, on a $\mu_{2n} = \delta_x$ et

6. Deux mesures μ et ν sont dites équivalentes si $\mu \ll \nu$ et $\nu \ll \mu$, i.e. elles ont les mêmes ensembles négligeables.

$\mu_{2n+1} = \delta_{1-x}$ et il n'y a pas de convergence en loi. Cependant, il est évident que pour toute fonction $\varphi : E \to \mathbb{R}$ et pour toute condition initiale

$$\frac{1}{n}\sum_{k=1}^{n}\varphi(X_k) \xrightarrow[n\to\infty]{\text{p.s.}} \frac{1}{2}\left(\varphi(0)+\varphi(1)\right) = \pi(\varphi).$$

De façon générale, pour une chaîne récurrente positive sur E dénombrable, si l'on note π son unique loi invariante, alors pour toute condition initiale x et pour toute fonction φ telle que $\pi(|\varphi|) < \infty$, on a

$$\frac{1}{n}\sum_{i=1}^{n}\varphi(X_i) \xrightarrow[n\to\infty]{\text{p.s.}} \pi(\varphi) = \sum_{x\in E}\pi(x)\varphi(x).$$

Modulo la même restriction sur la condition initiale que pour la convergence en variation totale, on retrouve cette ergodicité dans le cas général. Ce résultat est lui aussi admis.

Théorème 2.14 (Théorème ergodique)
S'il existe une mesure ν et une loi de probabilité π telles que la chaîne (X_n) soit ν-irréductible et π-invariante, alors pour toute fonction $\varphi : E \to \mathbb{R}$ telle que $\pi(|\varphi|) < \infty$ et pour π-presque toute condition initiale x, on a

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

Exemple : Notons $Y \sim \pi = \mathcal{N}(0,1)$ et $\varphi : \mathbb{R} \to \mathbb{R}$ telle que $\mathbb{E}[|\varphi(Y)|] < \infty$ alors, que ce soit pour le noyau gaussien ou le noyau mélange, on en déduit que pour presque toute condition initiale x

$$\frac{1}{n}\sum_{k=1}^{n}\varphi(X_k) \xrightarrow[n\to\infty]{\text{p.s.}} \pi(\varphi) = \mathbb{E}[\varphi(Y)] = \int_{\mathbb{R}}\varphi(y)\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy.$$

Remarque : Sous une hypothèse plus forte sur le noyau de transition P, il y a des résultats vrais pour **toute** condition initiale. En fait, si le noyau P est ν-irréductible, admet π comme loi invariante et si $P(x,\cdot) \ll \pi$ pour tout x, alors :

— Pour toute fonction $\varphi : E \to \mathbb{R}$ telle que $\pi(|\varphi|) < \infty$ et pour toute condition initiale x :

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

— Si de plus P est apériodique, alors pour toute condition initiale x, la loi de X_n tend vers π en variation totale :

$$\|P^n(x,\cdot) - \pi\| \xrightarrow[n\to\infty]{} 0.$$

De façon plus générale, ces résultats pour toute condition initiale sont vrais si la chaîne est Harris récurrente, c'est-à-dire si pour tout B tel que $\pi(B) > 0$ et toute condition initiale x, la chaîne partant de x atteint B en temps fini presque sûrement, ce qui s'écrit

$$\forall B \text{ tel que } \pi(B) > 0, \forall x \in E, \qquad \mathbb{P}_x(T_B^* < \infty) = 1.$$

Cette condition est plus forte que la π-irréductibilité, qui impose uniquement que cette probabilité soit strictement positive.

2.2 Méthodes MCMC

Etant donné une loi de probabilité π sur un espace d'états général, le but est de simuler une chaîne (X_n) de loi invariante π. Nous nous limitons dans ces notes à deux méthodes classiques : l'algorithme de Metropolis et l'échantillonneur de Gibbs.

2.2.1 Algorithme de Metropolis

Pour la loi cible π sur l'espace $(E, \mathcal{E})$, nous considérons un noyau de transition $Q(x, dy)$, appelé noyau de proposition ou noyau instrumental, et une mesure μ tels que π et Q soient tous deux absolument continus par rapport à μ. Nous convenons de noter $\pi(dx) = \pi(x)\mu(dx)$ et $Q(x, dy) = q(x, y)\mu(dy)$. Pour tout couple (x, y), le rapport de Metropolis est alors défini par

$$\rho(x, y) := \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}.$$

Nous posons également

$$\alpha(x, y) := \min(1, \rho(x, y)),$$

avec la convention [7] $\alpha(x, y) = 1$ si $\pi(x)q(x, y) = 0$, et enfin

$$r(x) := 1 - \int_E \alpha(x, y)Q(x, dy) = 1 - \int_E \alpha(x, y)q(x, y)\mu(dy).$$

Définition 2.15 (Noyau de Metropolis)
Le noyau de Metropolis associé à la loi cible π et au noyau de proposition Q est défini par

$$P(x, dy) := \alpha(x, y)Q(x, dy) + r(x)\delta_x(dy) = \alpha(x, y)q(x, y)\mu(dy) + r(x)\delta_x(dy).$$

Avec des mots (cf. plus bas pour la justification formelle) : partant de x, si le noyau Q propose une transition vers y, alors celle-ci est acceptée avec probabilité $\alpha(x, y)$, et refusée avec probabilité $1 - \alpha(x, y)$, auquel cas on reste sur place, c'est-à-dire en x.

Pour toute fonction test $\varphi : E \to \mathbb{R}$, on a donc

$$(P\varphi)(x) = \int_E P(x, dy)\varphi(y) = r(x)\varphi(x) + \int_E \alpha(x, y)Q(x, dy)\varphi(y). \tag{2.13}$$

Ceci définit bien un noyau de transition sur E car en prenant $\varphi(x) = 1$ pour tout x, on obtient par définition de $r(x)$

$$\int_E P(x, dy) = r(x) + \int_E \alpha(x, y)Q(x, dy) = 1.$$

Remarque : Si $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ et si la mesure dominante μ est la mesure de Lebesgue, le noyau de Metropolis n'est pas absolument continu par rapport à celle-ci : la probabilité $r(x)$ de rester sur place est en général strictement positive ! On retrouve donc un noyau de type mélange avec une partie Dirac et une partie à densité par rapport à la mesure de Lebesgue.

Lemme 2.16 (Noyau de Metropolis et réversibilité)
La loi cible π est réversible pour le noyau P. En particulier, π est invariante par P.

Preuve : Il suffit de montrer que, pour tout $x \neq y$,

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

Comme expliqué auparavant, puisque pour $x \neq y$ le noyau $P(x, dy) = \alpha(x, y)q(x, y)\mu(dy)$ est, tout comme π, à densité par rapport à μ, il suffit de vérifier que

$$\pi(x)\alpha(x, y)q(x, y) = \pi(y)\alpha(y, x)q(y, x).$$

Or, en vertu de la convention $\alpha(x, y) = 1$ si $\pi(x)q(x, y) = 0$, la formule suivante est toujours valide :

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

Cette dernière quantité étant symétrique en x et y, le résultat est établi.

7. Ceci n'a aucune espèce d'importance car en pratique cette situation ne se présentera pas.

■

Hypothèse 2.17
Les conditions d'application de l'algorithme de Metropolis sont les suivantes :
(a) Pour tout x, savoir simuler selon le noyau $Q(x,\cdot)$.
(b) Pour tout tirage $Y = y$ de l'étape précédente, savoir calculer le rapport $\rho(x,y)$.

Si ces conditions sont satisfaites, l'implémentation est élémentaire.

Définition 2.18 (Algorithme de Metropolis)
Considérons la suite de variables aléatoires (X_n) définie par une condition initiale X_0 et séquentiellement comme suit : sachant X_n,

1. *Simuler $Y_{n+1} \sim Q(X_n,\cdot)$*
2. *Simuler $U_{n+1} \sim \text{Unif}([0,1])$ indépendamment de toutes les variables précédentes, puis*
   — *si $U_{n+1} \leq \alpha(X_n, Y_{n+1})$, poser $X_{n+1} = Y_{n+1}$;*
   — *sinon, poser $X_{n+1} = X_n$.*

Alors (X_n) est une chaîne de Markov de noyau de transition le noyau P de la Définition 2.15. En particulier, (X_n) admet π comme loi réversible, donc invariante.

Preuve : Pour vérifier que (X_n) est bien une chaîne de Markov, nous montrons qu'elle peut s'écrire sous forme de récurrence aléatoire. Commençons par noter que

$$X_{n+1} = Y_{n+1}\mathbf{1}_{U_{n+1}\leq\alpha(X_n,Y_{n+1})} + X_n\mathbf{1}_{U_{n+1}>\alpha(X_n,Y_{n+1})} = h(X_n,(U_{n+1},Y_{n+1})). \tag{2.14}$$

Or $Y_{n+1} \sim Q(X_n,\cdot)$ donc on peut l'écrire sous la forme $Y_{n+1} = g(X_n, V_{n+1})$, avec la suite $(V_n)_{n\geq 1}$ i.i.d. et indépendante de X_0, ainsi bien entendu que de la suite $(U_n)_{n\geq 1}$. Au total, on a

$$X_{n+1} = \psi(X_n, W_{n+1}),$$

où la suite $(W_n)_{n\geq 1} = (U_n, V_n)_{n\geq 1}$ est i.i.d. et indépendante de X_0, ce qui assure que (X_n) est bien une chaîne de Markov. L'étape suivante consiste à établir que son noyau de transition est bien P, c'est-à-dire que pour toute fonction test $\varphi : E \to \mathbb{R}$, on a, d'après (2.13),

$$\mathbb{E}[\varphi(X_{n+1})|X_n] = (P\varphi)(X_n) = r(X_n)\varphi(X_n) + \int_E \alpha(X_n,y)Q(X_n,dy)\varphi(y).$$

On applique d'abord la propriété classique de conditionnement :

$$\mathbb{E}[\varphi(X_{n+1})|X_n] = \mathbb{E}[\mathbb{E}[\varphi(X_{n+1})|X_n, Y_{n+1}]|X_n].$$

Via (2.14) et par indépendance de la variable uniforme U_{n+1} vis-à-vis du couple (X_n, Y_{n+1}), on a

$$\begin{aligned}\mathbb{E}[\varphi(X_{n+1})|X_n,Y_{n+1}] &= \varphi(Y_{n+1})\mathbb{P}(U_{n+1}\leq\alpha(X_n,Y_{n+1})|X_n,Y_{n+1})\\ &\quad +\varphi(X_n)\mathbb{P}(U_{n+1}>\alpha(X_n,Y_{n+1})|X_n,Y_{n+1})\\ &= \varphi(Y_{n+1})\alpha(X_n,Y_{n+1}) + \varphi(X_n)(1-\alpha(X_n,Y_{n+1})).\end{aligned}$$

Puisque $Y_{n+1} \sim Q(X_n,\cdot)$, on a de façon générale

$$\mathbb{E}[\psi(Y_{n+1})|X_n] = (Q\psi)(X_n) = \int_E Q(X_n,dy)\psi(y),$$

mais aussi

$$\mathbb{E}[\psi(X_n,Y_{n+1})|X_n] = \int_E Q(X_n,dy)\psi(X_n,y).$$

Ceci appliqué à la formule obtenue ci-dessus pour $\mathbb{E}[\varphi(X_{n+1})|X_n,Y_{n+1}]$ donne bien, par définition de $r(X_n)$,

$$\mathbb{E}[\varphi(X_{n+1})|X_n] = r(X_n)\varphi(X_n) + \int_E \alpha(X_n,y)Q(X_n,dy)\varphi(y) = (P\varphi)(X_n).$$

■

Dans cet algorithme, il faut bien comprendre que le noyau de proposition Q, ou noyau instrumental, est choisi par l'utilisateur.

Exemples :

1. **Metropolis symétrique.** Il correspond au cas particulier d'un noyau symétrique $q(x,y) = q(y,x)$. Historiquement, c'est en fait sous cette forme qu'il a été présenté pour la première fois en 1953 dans l'article de Metropolis *et al.* [15]. Dans ce cas, le rapport de Metropolis se simplifie en $\rho(x,y) = \pi(y)/\pi(x)$. La généralisation à un noyau instrumental non nécessairement symétrique est généralement attribuée à Hastings en 1970 [9], d'où le nom générique d'algorithme de Metropolis-Hastings.
2. **Metropolis par marche aléatoire.** Cette fois $q(x,y) = f(y-x)$ où f est une densité par rapport à la mesure μ. Dit autrement, sachant $X_n = x$, on a $Y_{n+1} = x + W_{n+1}$ avec W_{n+1} de densité f. Sur la droite réelle, on peut penser à $W_{n+1} \sim \mathcal{N}(0,\sigma^2)$ ou encore $W_{n+1} \sim \mathcal{U}_{[-\sigma,+\sigma]}$ avec σ un paramètre de réglage qui correspond à la distance typique entre le point de départ x et la proposition Y_{n+1}. Noter que ceci suppose un minimum de structure algébrique sur E (au minimum de groupe abélien) pour pouvoir donner un sens à l'addition $Y_{n+1} = x + W_{n+1}$.
3. **Metropolis indépendant.** Dans ce cas $q(x,y) = f(y)$ où f est une densité par rapport à la mesure μ. Le point proposé Y_{n+1} ne tient pas compte de la position actuelle $X_n = x$ de la chaîne.

Interprétation : Considérons le cas du Metropolis symétrique. Lorsque la chaîne est en $X_n = x$, une transition vers $Y_{n+1} = y$ est proposée. Si le rapport $\rho(x,y) = \pi(y)/\pi(x)$ est supérieur à 1, autrement dit si y est plus vraisemblable que x aux yeux de la loi cible π, cette proposition est automatiquement acceptée ; dans le cas contraire, on ne s'interdit pas systématiquement de transiter vers y, mais on le fera seulement avec probabilité $\pi(y)/\pi(x)$. Si le noyau $Q(x,dy)$ correspond à un déplacement de x vers un point "voisin" y[8], accepter d'aller vers un point moins vraisemblable pour π permet de ne pas rester coincé en un maximum local de π.

Remarques :

1. **Constante de normalisation.** Supposons que l'on ne connaisse la densité cible π qu'à une constance de normalisation près, c'est-à-dire que $\pi(dx) = \pi(x)\mu(dx) = c\pi_u(x)\mu(dx)$, avec[9] $\pi_u(x)$ évaluable en tout x mais où la constante de normalisation

$$c = \left(\int_E \pi_u(x)\mu(dx) \right)^{-1}$$

est hors d'atteinte. On voit que ceci n'empêche en rien d'appliquer l'algorithme de Metropolis puisque π n'intervient que dans le rapport

$$\rho(x,y) = \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} = \frac{\pi_u(y)q(y,x)}{\pi_u(x)q(x,y)}.$$

Il suffit donc de remplacer π par π_u. Ce point a priori anecdotique est en fait crucial et a grandement contribué à la popularité de cet algorithme. On le rencontre aussi bien en physique statistique (voir par exemple le modèle d'Ising) qu'en statistique bayésienne, comme nous le verrons plus loin.

8. C'est par exemple le cas sur $\mathbb{R}$ lorsque $Y_{n+1} = x + W_{n+1}$ avec $W_{n+1} \sim \mathcal{N}(0,\sigma^2)$.
9. Le u en indice de la notation π_u signifie *unnormalized*.

2. **Noyau instrumental paramétré.** Considérons par exemple une loi cible π unimodale sur $\mathbb{R}$ et le noyau de proposition gaussien ci-dessus, i.e. $Y_{n+1} = x+W_{n+1}$ avec $W_{n+1} \sim \mathcal{N}(0, \sigma^2)$. Si la chaîne se trouve en $X_n = x$ proche du mode de π, alors si σ est "grand" par rapport à l'écart-type de la loi π, la proposition Y_{n+1} a de grandes chances de se retrouver dans la queue de la distribution de π, auquel cas le rapport de Metropolis sera proche de 0, cette proposition sera très probablement refusée et la chaîne restera sur place, c'est-à-dire $X_{n+1} = X_n$; dans l'autre sens, si σ est "petit", le rapport de Metropolis sera proche de 1, donc cette proposition sera très probablement acceptée et la chaîne bougera, c'est-à-dire $X_{n+1} = Y_{n+1}$, mais très peu. Ceci permet d'appréhender un principe général de Metropolis : un paramètre σ trop grand entraînera de nombreux refus, donc une chaîne qui reste sur place la plupart du temps, tandis qu'un paramètre σ trop faible entraînera de nombreuses transitions très petites, donc une chaîne qui bouge beaucoup mais très peu à chaque fois.

En l'état, nous avons donc construit une chaîne (X_n) qui admet π comme loi stationnaire. Pour pouvoir appliquer les résultats asymptotiques des Théorèmes 2.13 et 2.14, encore faut-il vérifier les hypothèses d'irréductibilité, voire d'apériodicité. Ceci se fera souvent au cas par cas, mais nous pouvons donner un résultat assez général en ce sens.

Proposition 2.19 (Irréductibilité de Metropolis sur $\mathbb{R}^d$)
Supposons $E = \mathbb{R}^d$, π et Q absolument continus par rapport à la mesure de Lebesgue, et la fonction $(x, y) \mapsto q(x, y)$ continue et strictement positive sur $\mathbb{R}^d \times \mathbb{R}^d$, alors la chaîne (X_n) produite par l'algorithme de Metropolis est π-irréductible et apériodique. Par conséquent, les Théorèmes 2.13 et 2.14 s'appliquent.

Preuve : Pour l'irréductiblité, nous devons montrer que pour tout borélien A tel que $\pi(A) > 0$ et pour tout $x \in \mathbb{R}^d$, on a

$$\mathbb{P}_x \left(\exists n \geq 1,\ X_n \in A\right) > 0.$$

Nous allons prouver qu'il suffit de prendre $n = 1$, ce qui revient à dire qu'on peut atteindre A en un coup via le noyau P, i.e. $P(x, A) > 0$. Commençons par régler le cas sans intérêt où $\pi(x) = 0$: la convention $\alpha(x, y) = 1$ si $\pi(x)q(x, y) = 0$ implique

$$P(x, A) = \int_A q(x, y)dy > 0,$$

puisque q est strictement positive sur le borélien A de mesure de Lebesgue $\lambda(A)$ strictement positive car $\pi \ll \lambda$ et $\pi(A) > 0$ par hypothèse. Supposons donc désormais $\pi(x) > 0$: puisque q est strictement positive, le rapport de Metropolis est bien défini dans la formule

$$\rho(x, y) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}.$$

Maintenant, soit $A_k := A \cap B(0, k)$ intersection de A avec la boule de rayon k centrée en l'origine. Par hypothèse sur A et continuité monotone croissante,

$$0 < \pi(A) = \pi\left(\bigcup_{k=1}^{\infty} A_k\right) = \lim_{k\to\infty} \pi(A_k),$$

donc il existe k tel que $\pi(A_k) > 0$. Le point x étant fixé, la fonction $y \mapsto h(y) := \min(q(x, y), q(y, x))$ est continue et strictement positive sur A_k. On en déduit que $\varepsilon = \varepsilon(k) := \inf_{y\in A_k} h(y) > 0$. En effet, supposons que $\varepsilon = 0$, alors il existerait une suite (y_n) de A_k telle que $h(y_n) \to 0$, mais A_k étant borné la suite (y_n) admet une valeur d'adhérence, c'est-à-dire qu'il existe une sous-suite $(y_{\varphi(n)})$ et un point y^* dans l'adhérence de A_k tel que $(y_{\varphi(n)})$ tende vers y^*, ce qui par continuité de

h impliquerait $h(y^*) = \lim_{n\to\infty} h(y_{\varphi(n)}) = 0$, qui est exclu car $h(y^*) = \min(q(x,y^*), q(y^*,x)) > 0$. Par ailleurs,

$$P(x,A) \geq P(x,A_k) \geq \int_{A_k} q(x,y) \min\left(1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right) dy.$$

Soit la partition $A_k = B_k \cup C_k$ où $B_k := \{y \in A_k : \ \pi(y) \geq \pi(x)\}$ et $C_k := \{y \in A_k : \ \pi(y) < \pi(x)\}$. Sur B_k, on a directement

$$\int_{B_k} q(x,y) \min\left(1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right) dy \geq \int_{B_k} \min\left(q(x,y), q(y,x)\right) dy \geq \int_{B_k} \varepsilon dy = \varepsilon\lambda(B_k),$$

tandis que sur C_k il vient

$$\int_{C_k} q(x,y) \min\left(1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right) dy \geq \frac{1}{\pi(x)} \int_{C_k} \pi(y) \min\left(q(x,y), q(y,x)\right) dy \geq \frac{\varepsilon}{\pi(x)}\pi(C_k),$$

donc

$$P(x,A) \geq \varepsilon\left(\lambda(B_k) + \frac{1}{\pi(x)}\pi(C_k)\right) \geq \varepsilon \min\left(1, \frac{1}{\pi(x)}\right)\left(\lambda(B_k) + \pi(C_k)\right).$$

Or le dernier terme ne peut être nul car ceci signifierait que $\lambda(B_k) = \pi(C_k) = 0$, mais puisque $\pi \ll \lambda$ il s'ensuivrait que $\pi(B_k) = 0$ donc $\pi(A_k) = \pi(B_k) + \pi(C_k) = 0$, ce qui est exclu. Ainsi $P(x,A) > 0$ et l'irréductibilité est établie. Le fait que $P(x,A) > 0$ pour tout x de E et tout A tel que $\pi(A) > 0$ implique l'apériodicité comme vu au Lemme 2.10.

■

Exemple : Si la loi cible π a une densité strictement positive par rapport à la mesure de Lebesgue sur $\mathbb{R}$, celles-ci sont donc équivalentes. Si l'on considère par exemple le noyau de proposition gaussien $Y_{n+1} = x + W_{n+1}$ avec $W_{n+1} \sim \mathcal{N}(0, \sigma^2)$, ce qui correspond à

$$q(x,y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-x)^2}{2\sigma^2}},$$

les conditions de la Proposition 2.19 sont satisfaites, donc pour presque toute condition initiale, la convergence en variation totale et le Théorème ergodique sont vérifiés.

Dans le cas discret, c'est-à-dire en espace d'états dénombrable, il est naturel de considérer que la loi cible π met du poids en chaque point de E, quitte à restreindre celui-ci. Dans ce cas, π et la mesure de comptage μ sont équivalentes donc la π-irréductibilité d'une chaîne correspond à l'irréductibilité au sens classique. Le résultat suivant montre que tout se passe bien pour P dès lors que c'est le cas pour le noyau de proposition Q, lequel est, comme nous l'avons dit, choisi par l'utilisateur.

Proposition 2.20 (Metropolis dans le cas discret)
Sur E dénombrable, considérons une loi π telle que $\pi(x) > 0$ pour tout x et un noyau de proposition Q tel que $q(x,y) > 0$ ssi $q(y,x) > 0$. Si le noyau Q est irréductible et apériodique, alors le noyau de Metropolis P l'est aussi.

Preuve : Commençons par noter que, dans le cas discret, la matrice de transition $P = [p(x,y)]$ correspondant au noyau P s'écrit, si $x \neq y$,

$$p(x,y) = \alpha(x,y)q(x,y),$$

et, pour les termes diagonaux,

$$p(x,x) = 1 - \sum_{y\neq x} p(x,y).$$

Donc, si $x \neq y$,

$$p(x,y) = \min\left(1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right) q(x,y),$$

d'où nous déduisons deux choses : d'une part $q(x,y) > 0$ implique $p(x,y) > 0$, d'autre part $p(x,y) \leq q(x,y)$. Par conséquent, si $x = y$, on a

$$p(x,x) = 1 - \sum_{y \neq x} p(x,y) \geq 1 - \sum_{y \neq x} q(x,y) = q(x,x).$$

Au total, quel que soit le couple (x,y), $q(x,y) > 0$ implique $p(x,y) > 0$. Or, dire que Q est irréductible (au sens usuel) signifie que pour tout couple de points (x,y), il existe $n \geq 1$ et une séquence $x_0 = x \to x_1 \to \cdots \to x_n = y$ telle que $q(x_i, x_{i+1}) > 0$ pour tout i. Mais puisque $q(x_i, x_{i+1}) > 0$ implique $p(x_i, x_{i+1}) > 0$, il s'ensuit que le même chemin permet d'aller de x à y via le noyau P, qui est donc irréductible. Concernant l'apériodicité, rappelons qu'en espace discret, une condition suffisante pour qu'une chaîne irréductible soit apériodique est qu'il existe un état x sur lequel on puisse boucler. Supposons que P, dont nous venons d'établir l'irréductibilité, ne soit pas apériodique : ceci implique donc en particulier que $p(x,x) = 0$ pour tout x. Puisque $p(x,x) \geq q(x,x)$, on en déduit d'une part que $q(x,x) = 0$ et d'autre part que, pour tout x,

$$0 = p(x,x) = 1 - \sum_{y \neq x} \alpha(x,y) q(x,y).$$

Puisque $\sum_y q(x,y) = 1$ avec $q(x,x) = 0$, on peut donc écrire

$$0 = 1 - \sum_{y \neq x} \alpha(x,y) q(x,y) = \sum_{y \neq x} (1 - \alpha(x,y)) q(x,y),$$

ce qui impose $\alpha(x,y) = 1$ pour tout couple $x \neq y$ tel que $q(x,y) > 0$. Ainsi, pour $x \neq y$, le noyau P se simplifie en

$$p(x,y) = \alpha(x,y) q(x,y) = q(x,y),$$

c'est-à-dire que $P = Q$, mais Q est supposé apériodique, d'où la contradiction.

■

Exemple pathologique (et sans aucun intérêt) : 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.

Dans le contexte précédent, puisque π et la mesure de comptage μ sont équivalentes, le π-presque partout sur la condition initiale des Théorèmes 2.13 et 2.14 est même vrai partout.

Corollaire 2.21 (Résultats asymptotiques dans le cas discret)
Sous les hypothèses de la Proposition 2.20, pour toute condition initiale x de l'espace dénombrable E, on a convergence en variation totale :

$$\|P^n(x,\cdot) - \pi\| \xrightarrow[n\to\infty]{} 0,$$

ainsi que le Théorème ergodique, c'est-à-dire que pour toute fonction $\varphi : E \to \mathbb{R}$ telle que $\pi(|\varphi|) < \infty$ et pour toute condition initiale x :

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

2.2.2 Echantillonneur de Gibbs

L'échantillonneur de Gibbs est une méthode MCMC mais n'entre pas stricto sensu dans le cadre de l'algorithme de Metropolis hormis, comme nous le verrons, dans sa version balayage aléatoire sur un espace discret. Cette méthode requiert par ailleurs davantage de connaissances sur la loi cible π. De plus, si son implémentation est élémentaire, les notations sont un peu lourdes.

Sur un espace produit $E = E_1 \times \cdots \times E_d$, nous considérons une loi cible π supposée avoir une densité par rapport à une mesure de référence $\mu = \mu_1 \otimes \cdots \otimes \mu_d$, c'est-à-dire

$$\pi(dx) = \pi(x)\mu(dx) = \pi(x_1, \ldots, x_d)\mu_1(dx_1) \ldots \mu_d(dx_d).$$

On a bien sûr en tête les deux situations suivantes :

— Les espaces E_ℓ sont dénombrables, munis de la mesure de comptage, donc E est dénombrable muni de la mesure de comptage. En notant $X = (X_1, \ldots, X_d)$ un vecteur aléatoire de loi π, on écrira ainsi

$$\mathbb{P}(X = x) = \pi(x) = \pi(x_1, \ldots, x_d) = \mathbb{P}(X_1 = x_1, \ldots, X_d = x_d).$$

— Les espaces E_ℓ sont des intervalles de $\mathbb{R}$ muni de la mesure de Lebesgue $\lambda(dx_\ell)$, et $\pi(x)$ est la densité par rapport à la mesure de Lebesgue λ_d sur $\mathbb{R}^d$. En notant $X = (X_1, \ldots, X_d)$ un vecteur aléatoire de loi π on a donc, pour toute fonction test $\varphi : \mathbb{R}^d \to \mathbb{R}$,

$$\mathbb{E}[\varphi(X)] = \int_{\mathbb{R}^d} \varphi(x)\pi(dx) = \int_{\mathbb{R}^d} \varphi(x)\pi(x)dx_1 \ldots dx_d.$$

Remarques :

1. Une fois la densité π choisie[10], nous appellerons support de la loi π l'ensemble

$$\mathcal{S} := \{x \in E,\ \pi(x) > 0\}.$$

   En toute rigueur, on devrait l'appeler support essentiel.

2. En toute généralité, on pourrait définir l'échantillonneur de Gibbs même dans le cas où π n'est pas à densité par rapport à une mesure produit $\mu = \mu_1 \otimes \cdots \otimes \mu_d$, mais notre cadre suffit pour toutes les situations que nous rencontrerons en pratique.

Notations : Pour tout $x = (x_1, \ldots, x_d)$ de E et tout $1 \leq \ell \leq d$, nous écrirons

$$x_{-\ell} := (x_1, \ldots, x_{\ell-1}, x_{\ell+1}, \ldots, x_d),$$

et la densité jointe sous la forme $\pi(x) = \pi(x_\ell, x_{-\ell})$. En notant

$$E_{-\ell} := E_1 \times \cdots \times E_{\ell-1} \times E_{\ell+1} \times \cdots \times E_d$$

l'espace de dimension $(d-1)$ associé, on peut donc définir la loi jointe associée à π et $E_{-\ell}$ par

$$\pi_{-\ell}(x_{-\ell}) := \int_{E_\ell} \pi(x_\ell, x_{-\ell})\mu_\ell(dx_\ell).$$

Si $\pi_{-\ell}(x_{-\ell}) > 0$, la densité conditionnelle est définie comme d'habitude par

$$\pi(x_\ell | x_{-\ell}) := \frac{\pi(x_\ell, x_{-\ell})}{\pi_{-\ell}(x_{-\ell})} = \frac{\pi(x)}{\pi_{-\ell}(x_{-\ell})}.$$

10. Celle-ci n'est définie qu'à un ensemble de μ-mesure nulle près.

Remarque : Comme nous le verrons, l'échantillonneur de Gibbs nécessite $\pi_{-\ell}(x_{-\ell}) > 0$. Ce n'est pas un problème car la loi π ne voit pas les points x pour lesquels il existe une coordonnée ℓ telle que $\pi_{-\ell}(x_{-\ell}) = 0$. En effet, soit

$$N_1 := E_1 \times F_{-1} = E_1 \times \{x_{-1} \in E_{-1},\ \pi_{-1}(x_{-1}) = 0\},$$

alors, en notant $\mu_{-\ell} := \mu_1 \otimes \dots \mu_{\ell-1} \otimes \mu_{\ell+1} \otimes \cdots \otimes \mu_d$,

$$\pi(N_1) = \int_{F_{-1}} \left(\int_{E_1} \pi(x_1, x_{-1}) \mu_1(dx_1) \right) \mu_{-1}(dx_{-1}) = \int_{F_{-1}} \pi_{-1}(x_{-1}) \mu_{-1}(dx_{-1}) = 0.$$

Si $N := N_1 \cup \cdots \cup N_d$, il s'ensuit que $\pi(N) = 0$. Dans la suite, lorsque nous écrirons $\pi(x_\ell | x_{-\ell})$, il sera implicitement supposé que $\pi_{-\ell}(x_{-\ell}) > 0$. Ainsi nous aurons toujours $\pi(x) = \pi(x_\ell | x_{-\ell}) \pi_{-\ell}(x_{-\ell})$.

Enfin, pour tout $\ell \in \{1, \dots, d\}$, nous noterons π_ℓ la densité marginale de rang ℓ c'est-à-dire, pour tout $x_\ell \in E_\ell$

$$\pi_\ell(x_\ell) = \int_{E_{-\ell}} \pi(x_\ell, x_{-\ell}) \mu_{-\ell}(dx_{-\ell}).$$

Nous appellerons support (essentiel) de la marginale π_ℓ l'ensemble

$$\mathcal{S}_\ell := \{x_\ell \in E_\ell,\ \pi_\ell(x_\ell) > 0\}.$$

Pour résumer, si $X = (X_1, \dots, X_d) \sim \pi$, alors $X_\ell \sim \pi_\ell$ pour tout ℓ et

$$X_{-\ell} := (X_1, \dots, X_{\ell-1}, X_{\ell+1}, \dots, X_d) \sim \pi_{-\ell}.$$

L'hypothèse fondamentale (et très forte) pour pouvoir appliquer l'échantillonneur de Gibbs est la suivante.

Hypothèse 2.22
Pour tous ℓ et $x_{-\ell}$ tel que $\pi_{-\ell}(x_{-\ell}) > 0$, on sait simuler selon la densité conditionnelle $\pi(\cdot | x_{-\ell})$.

Exemple : Modèle auto-exponentiel. Sur $E = \mathbb{R}_+^2$ muni de la tribu borélienne et de la mesure de Lebesgue $\mathbf{1}_E(x)dx$, considérons la densité

$$\pi(x) = \pi(x_1, x_2) \propto \exp\{-(x_1 + x_2 + x_1 x_2)\}.$$

On voit immédiatement que l'Hypothèse 2.22 est vérifiée puisque, pour tout $x_{-1} = x_2 \in \mathbb{R}_+$,

$$\text{Loi}(X_1 | x_{-1}) = \text{Loi}(X_1 | x_2) = \text{Exp}(1 + x_2),$$

et de même pour la loi de X_2 sachant x_1, par symétrie. Ces lois conditionnelles sont facilement simulables. Noter qu'on saurait simuler directement selon π si l'on savait aussi simuler facilement selon la loi marginale de X_2, mais ce n'est pas le cas puisque, pour tout $x_2 \geq 0$,

$$\pi_2(x_2) \propto \frac{e^{-x_2}}{1 + x_2},$$

qui n'est pas aisément simulable [11].

Revenons au cas général. Nous noterons Q_ℓ le noyau consistant à modifier uniquement la ℓ-ème coordonnée en effectuant un tirage suivant la densité conditionnelle $\pi(\cdot | x_{-\ell})$, c'est-à-dire

$$Q_\ell(x, dx') := \pi(x'_\ell | x_{-\ell}) \mu_\ell(dx'_\ell) \delta_{x_{-\ell}}(dx'_{-\ell}). \tag{2.15}$$

11. Noter qu'on pourrait appliquer une méthode de rejet avec la loi exponentielle comme loi de proposition, mais avec l'inconvénient inhérent à cette technique, à savoir qu'on rejette...

Noter qu'en général le noyau Q_ℓ n'est pas à densité par rapport à μ puisqu'il gèle $(d-1)$ coordonnées. C'est cependant le cas si E est dénombrable, muni de la tribu pleine $\mathcal{E} = \mathcal{P}(E)$, et que la mesure μ est la mesure de comptage sur celui-ci : cette dernière admet en effet pour seul ensemble négligeable l'ensemble vide, or $Q_\ell(x, \emptyset) = 0$ pour tout x par définition même d'une mesure.

Lemme 2.23 (Réversibilité)
Pour tout $\ell \in \{1, \ldots, d\}$, la loi π est réversible par rapport à Q_ℓ, donc invariante pour Q_ℓ.

Preuve : Par définition du noyau Q_ℓ, on a $Q_\ell(x, dx') = Q_\ell(x', dx) = 0$ si $x_{-\ell} \neq x'_{-\ell}$. Ainsi, seule est à considérer la situation où $x = (x_\ell, x_{-\ell})$ et $x' = (x'_\ell, x_{-\ell})$, auquel cas pour montrer

$$\pi(dx)Q_\ell(x, dx') = \pi(dx')Q_\ell(x', dx),$$

il suffit d'établir

$$\pi(x)\pi(x'_\ell|x_{-\ell}) = \pi(x')\pi(x_\ell|x'_{-\ell}).$$

Or, puisque $x_{-\ell} = x'_{-\ell}$,

$$\pi(x)\pi(x'_\ell|x_{-\ell}) = \pi(x_\ell, x_{-\ell})\pi(x'_\ell|x_{-\ell}) = \pi(x_\ell|x'_{-\ell})\pi(x'_\ell, x_{-\ell}) = \pi(x_\ell|x'_{-\ell})\pi(x').$$

■

Balayage aléatoire

Cet algorithme revient à considérer le noyau de transition obtenu par mélange uniforme des noyaux individuels Q_ℓ. Nous avons vu en Section 1.1.4 que si $\nu_1, \ldots, \nu_d$ sont des lois de probabilités simulables, alors la loi $\nu = \frac{1}{d}(\nu_1 + \cdots + \nu_d)$ peut se simuler comme suit : on tire un numéro $L = \ell$ au hasard entre 1 et d, puis on simule selon la loi ν_ℓ.

Définition 2.24 (Gibbs à balayage aléatoire)
Considérons la suite de variables aléatoires (X_n) définie par une condition initiale X_0 et séquentiellement comme suit : sachant $X_n = x$,

1. *Tirer une coordonnée $L = \ell$ de façon équiprobable dans $\{1, \ldots, d\}$;*
2. *Simuler $X'_\ell = x'_\ell$ suivant la densité conditionnelle $\pi(\cdot|x_{-\ell})$;*
3. *Poser $X_{n+1} = (x'_\ell, x_{-\ell})$.*

Alors (X_n) est une chaîne de Markov de noyau de transition

$$P := \frac{1}{d}\left(Q_1 + \cdots + Q_d\right).$$

La loi π est réversible, donc invariante, pour cette chaîne.

Preuve : La réversibilité de π pour P découle de sa réversibilité pour chaque Q_ℓ du Lemme 2.23 :

$$\pi(dx)P(x, dx') = \frac{1}{d}\left(\pi(dx)Q_1(x, dx') + \cdots + \pi(dx)Q_d(x, dx')\right) = \pi(dx')P(x', dx).$$

■

Remarques :

1. On peut à nouveau noter qu'en général le noyau P n'est pas à densité par rapport à μ puisque le nouveau point x' coïncide avec le point de départ x sur $(d-1)$ coordonnées. Si par exemple $E = \mathbb{R}^2$ muni de la mesure de Lebesgue, alors en notant

   $$B := (\{x_1\} \times \mathbb{R}) \cup (\mathbb{R} \times \{x_2\}),$$

   on voit que $P(x, B) = 1$ tandis que la mesure de Lebesgue de B est nulle (union de deux droites dans le plan).

2. Pour les mêmes raisons que ci-dessus, c'est néanmoins le cas si E est dénombrable et muni de la mesure de comptage. Dans ce cas, la densité du noyau Q_ℓ par rapport à la mesure de comptage est
$$q_\ell(x, x') = \pi(x'_\ell | x_{-\ell}) \mathbf{1}_{x'_{-\ell} = x_{-\ell}},$$
et celle du noyau P est donc
$$p(x, x') = \frac{1}{d} \sum_{\ell=1}^{d} \pi(x'_\ell | x_{-\ell}) \mathbf{1}_{x'_{-\ell} = x_{-\ell}}. \tag{2.16}$$

Le résultat qui suit montre que, dans le cas spécifique d'un espace discret, l'algorithme de Gibbs à balayage aléatoire correspond à un algorithme de Metropolis.

Lemme 2.25 (Lien entre Metropolis et Gibbs en espace discret)
Sur un espace discret E, l'échantillonneur de Gibbs à balayage aléatoire correspond à un algorithme de Metropolis où toutes les propositions sont acceptées.

Preuve : Si l'on reprend les notations de l'algorithme de Metropolis, l'échantillonneur de Gibbs revient, partant de $X_n = x$, à proposer un point $Y_{n+1} = x'$ selon $p(x, x')$ donnée en équation (2.16). De ce fait,
$$\pi(x) p(x, x') = \frac{1}{d} \sum_{\ell=1}^{d} \pi(x) \pi(x'_\ell | x_{-\ell}) \mathbf{1}_{x'_{-\ell} = x_{-\ell}} = \frac{1}{d} \sum_{\ell=1}^{d} \frac{\pi(x)\pi(x')}{\pi_{-\ell}(x_{-\ell})} \mathbf{1}_{x'_{-\ell} = x_{-\ell}},$$
d'où
$$\pi(x') p(x', x) = \frac{1}{d} \sum_{\ell=1}^{d} \frac{\pi(x)\pi(x')}{\pi_{-\ell}(x'_{-\ell})} \mathbf{1}_{x'_{-\ell} = x_{-\ell}} = \pi(x) p(x, x'),$$
et le rapport de Metropolis vaut donc 1.

■

En pratique et en particulier en petite dimension, ce n'est pas l'algorithme à balayage aléatoire qui est implémenté, et ce pour une raison très simple : supposons que $d = 2$ et que la chaîne parte de $X_0 = x = (x_1, x_2)$. Une coordonnée est tirée au hasard et modifiée, par exemple la seconde, ce qui donne $X_1 = (x_1, x'_2)$ avec x'_2 obtenu par tirage selon la densité conditionnelle $\pi(\cdot | x_1)$. Au coup suivant, si l'on tire à nouveau la seconde coordonnée, alors $X_2 = (x_1, x''_2)$ avec x''_2 obtenu par tirage selon la densité conditionnelle $\pi(\cdot | x_1)$, c'est-à-dire la même qu'au coup précédent : on a fait deux fois la même chose, c'est clairement une perte de temps. D'où le principe de l'échantillonneur de Gibbs à balayage séquentiel que nous présentons maintenant.

Balayage séquentiel (ou déterministe)

Cette méthode revient à composer successivement les noyaux Q_ℓ. Rappelons qu'un noyau de transition agit sur les mesures à gauche : si l'on part de $X_0 \sim \mu$ et que l'on applique à X_0 les transitions P puis Q, la loi de la variable X_2 obtenue après ces deux transitions est μPQ.

Définition 2.26 (Gibbs à balayage séquentiel)
Considérons la suite de variables aléatoires (X_n) définie par une condition initiale X_0 et séquentiellement comme suit : sachant $X_n = x$,

1. *Pour ℓ allant de 1 à d, simuler successivement $X'_\ell = x'_\ell$ selon la densité*
$$\pi(\cdot | x'_1, \ldots, x'_{\ell-1}, x_{\ell+1}, \ldots, x_d).$$

2. *Poser* $X_{n+1} = (x'_1, \ldots, x'_d)$.

Alors (X_n) *est une chaîne de Markov de noyau de transition*

$$P := Q_1 \ldots Q_d,$$

et la loi π *est invariante pour cette chaîne.*

Preuve : L'invariance de π pour P découle de son invariance pour chaque Q_ℓ vue en Lemme 2.23 :

$$\pi P = \pi Q_1 \ldots Q_d = \pi Q_2 \ldots Q_d = \pi.$$

■

Exemple : Modèle auto-exponentiel. Sachant $X_n = x = (x_1, x_2) \in \mathbb{R}_+^2$:

1. On simule $X'_1 = x'_1$ selon la loi $\mathrm{Exp}(1 + x_2)$, puis $X'_2 = x'_2$ selon la loi $\mathrm{Exp}(1 + x'_1)$.
2. On pose $X_{n+1} = (x'_1, x'_2)$.

Remarques :

1. Contrairement au noyau à balayage aléatoire, celui à balayage déterministe est absolument continu par rapport à μ, avec pour densité

$$p(x, x') = \pi(x'_1|x_2, \ldots, x_d)\pi(x'_2|x'_1, x_3, \ldots, x_d) \cdots \pi(x'_d|x'_1, \ldots, x'_{d-1}). \tag{2.17}$$

   En effet, en dimension $d = 2$, on peut passer par les fonctions tests ou le voir comme suit :

$$P(x, dx') = (Q_1 Q_2)(x, dx') = \int_E Q_1(x, dy) Q_2(y, dx'),$$

   ce qui donne, par l'expression (2.15) des noyaux Q_1 et Q_2,

$$P(x, dx') = \int_E \pi(y_1|x_2)\mu_1(dy_1)\delta_{x_2}(dy_2)\pi(x'_2|y_1)\mu_2(dx'_2)\delta_{y_1}(dx'_1),$$

   c'est-à-dire, par Fubini,

$$P(x, dx') = \left(\int_{E_1} \pi(y_1|x_2)\mu_1(dy_1)\pi(x'_2|y_1)\delta_{y_1}(dx'_1) \left(\int_{E_2} \delta_{x_2}(dy_2) \right) \right) \mu_2(dx'_2),$$

   avec $\int_{E_2} \delta_{x_2}(dy_2) = 1$ donc

$$P(x, dx') = \left(\int_{E_1} \pi(y_1|x_2)\mu_1(dy_1)\pi(x'_2|y_1)\delta_{y_1}(dx'_1) \right) \mu_2(dx'_2),$$

   et finalement

$$P(x, dx') = \pi(x'_1|x_2)\pi(x'_2|x'_1)\mu_1(dx'_1)\mu_2(dx'_2) = \pi(x'_1|x_2)\pi(x'_2|x'_1)\mu(dx'),$$

   ce qui prouve que, pour tout $x \in E$, on a $P(x, \cdot) \ll \mu$ avec la densité annoncée en (2.17).

2. En général, la loi cible π n'est pas réversible pour P. Par exemple, en dimension 2, il faudrait en effet qu'elle vérifie les équations d'équilibre détaillé, i.e. pour presque tous couples $x = (x_1, x_2)$ et $x' = (x'_1, x'_2)$:

$$\pi(x)p(x, x') = \pi(x')p(x', x),$$

   c'est-à-dire, vu l'expression ci-dessus,

$$\pi(x_1, x_2)\pi(x'_1|x_2)\pi(x'_2|x'_1) = \pi(x'_1, x'_2)\pi(x_1|x'_2)\pi(x_2|x_1),$$

   ou encore

$$\pi(x_1)\pi(x'_1|x_2) = \pi(x'_1)\pi(x_1|x'_2).$$

   Cette relation n'est pas vraie en général : il suffit par exemple de noter que le membre de gauche dépend de x_2, contrairement à celui de droite.

3. Néanmoins, lorsque les composantes de π sont indépendantes, il y a bien réversibilité puisqu'alors les densités conditionnelles ne dépendent pas du conditionnement donc, toujours d'après (2.17),
$$p(x, x') = \pi(x'_1)\pi(x'_2)\cdots\pi(x'_d) = \pi(x').$$
Cependant, dans ce cas, l'hypothèse fondamentale selon laquelle on sait simuler selon les lois conditionnelles implique que l'on sait simuler directement selon la loi jointe π car ces lois conditionnelles correspondent alors aux marginales d'une loi à composantes indépendantes.
4. En général, l'échantillonneur de Gibbs séquentiel ne correspond pas à un algorithme de Metropolis, attendu que pour ce dernier il y a toujours réversibilité.
5. Dans la suite, nous nous focalisons sur la version séquentielle de l'échantillonneur de Gibbs.

Concernant la convergence de (X_n), les Théorèmes 2.13 et 2.14 donnent le résultat suivant. Rappelons qu'un noyau est dit irréductible s'il existe une mesure ν pour lequel il est ν-irréductible.

Corollaire 2.27 (Résultats asymptotiques)
Si le noyau P de l'échantillonneur de Gibbs séquentiel est irréductible, alors :
— *Pour toute fonction $\varphi : E \to \mathbb{R}$ telle que $\pi(|\varphi|) < \infty$ et pour π presque toute condition initiale x :*
$$\frac{1}{n}\sum_{k=1}^{n} \varphi(X_k) \xrightarrow[n\to\infty]{\text{p.s.}} \pi(\varphi).$$
— *Si de plus P est apériodique, alors pour π presque toute condition initiale x, la loi de X_n tend vers π en variation totale :*
$$\|P^n(x,\cdot) - \pi\| \xrightarrow[n\to\infty]{} 0.$$

Donnons un exemple élémentaire où l'irréductibilité pose problème.

Exemple : Problème d'irréductibilité. Plaçons-nous sur $E := [-1,1]\times[-1,1]$ muni de la tribu borélienne et considérons la loi π uniforme sur $\mathcal{S} := \mathcal{C}^- \cup \mathcal{C}^+$ avec $\mathcal{C}^+ := [0,1]\times[0,1]$ et $\mathcal{C}^-$ son symétrique par rapport à l'origine (voir Figure 2.4). La mesure dominante est la mesure de Lebesgue λ_2 sur E. Il est clair que si $X_0 = x$ est intérieur à $\mathcal{C}^+$, alors la chaîne reste p.s. éternellement dans ce carré et la convergence vers π est impossible. Autrement dit, cette chaîne n'est pas π-irréductible puisque pour tout x intérieur à $\mathcal{C}^+$ et tout n, on a $P^n(x, \mathcal{C}^-) = 0$ alors que $\pi(\mathcal{C}^-) = 1/2$. L'origine du problème est claire : il y a un hiatus entre le support $\mathcal{S}$ de la loi π et les supports de ses marginales.

Proposition 2.28 (Condition suffisante d'irréductibilité)
Si le support de la loi π est le produit cartésien des supports des marginales π_ℓ, alors la chaîne (X_n) est π-irréductible et apériodique.

Preuve : Afin d'alléger les notations, nous considérons le cas $d = 2$. Soit donc B tel que $\pi(B) > 0$: nous voulons montrer que $P(x, B) > 0$ pour tout x. Commençons par noter que
$$\pi(B) = \int_{E_1}\left(\int_{E_2} \mathbf{1}_B(y_1,y_2)\pi(y_1,y_2)\mu_2(dy_2)\right)\mu_1(dy_1) > 0.$$
Ceci implique qu'il existe un ensemble mesurable $B_1 \subseteq E_1$ tel que $\mu_1(B_1) > 0$ et pour tout $y_1 \in B_1$
$$0 < \int_{E_2} \mathbf{1}_B(y_1,y_2)\pi(y_1,y_2)\mu_2(dy_2) \le \int_{E_2} \pi(y_1,y_2)\mu_2(dy_2) = \pi_1(y_1),$$

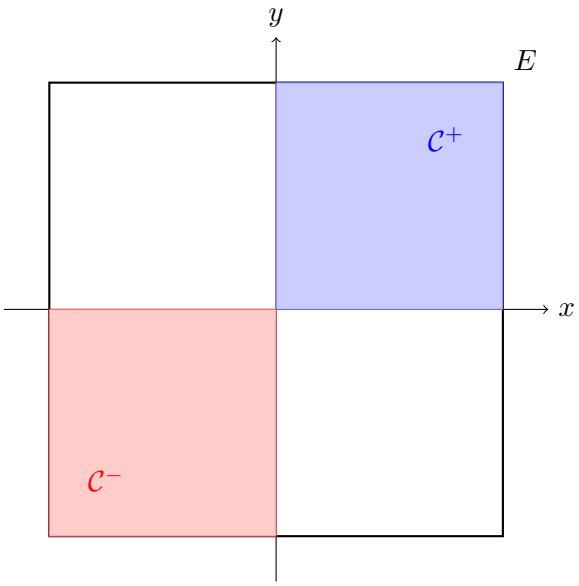

FIGURE 2.4 – Problème d'irréductibilité pour l'échantillonneur de Gibbs lorsque $\pi = \text{Unif}(\mathcal{C}^- \cup \mathcal{C}^+)$.

c'est-à-dire que y_1 est dans le support de π_1. Pour tout y_1 dans B_1, soit le sous-ensemble mesurable de E_2 défini par $C(y_1) := \{y_2 \in E_2 : \ (y_1, y_2) \in B\}$, on a donc par la ligne précédente

$$\int_{C(y_1)} \pi(y_1, y_2)\mu_2(dy_2) = \int_{E_2} \mathbf{1}_B(y_1, y_2)\pi(y_1, y_2)\mu_2(dy_2) > 0.$$

Il en résulte qu'il existe un sous-ensemble mesurable $D(y_1) \subseteq C(y_1)$ tel que $\mu_2(D(y_1)) > 0$ et pour tout $y_2 \in D(y_1)$:

$$\pi(y_1, y_2) > 0.$$

Considérons alors un point de départ $x = (x_1, x_2)$, et prouvons l'irréductibilité en un coup, i.e. $P(x, B) > 0$. Nous supposons bien sûr que $\pi_2(x_2) > 0$, sans quoi la loi conditionnelle $\pi(\cdot|x_2)$ n'est pas définie. Autrement dit, x_2 est dans le support de π_2. Puisque tout y_1 de B_1 est dans le support de π_1, l'hypothèse de la Proposition 2.28 implique que (y_1, x_2) est dans le support de π, i.e. $\pi(y_1, x_2) > 0$, donc $\pi(y_1|x_2) > 0$. De plus, pour tout y_2 dans $D(y_1)$, nous avons aussi $\pi(y_2|y_1) > 0$. Or, d'après la formule (2.17) pour le noyau de transition,

$$P(x, B) = \int_{E_1 \times E_2} \mathbf{1}_B(y_1, y_2)\pi(y_1|x_2)\pi(y_2|y_1)\mu_1(dy_1)\mu_2(dy_2),$$

d'où

$$P(x, B) \geq \int_{B_1} \pi(y_1|x_2) \left(\int_{D(y_1)} \pi(y_2|y_1)\mu_2(dy_2) \right) \mu_1(dy_1) > 0.$$

Puisque nous avons $P(x, B) > 0$ pour tout B tel que $\pi(B) > 0$, le Lemme 2.10 assure l'apériodicité. ■

Exemples :

1. Modèle auto-exponentiel : Rappelons que nous nous sommes placés d'emblée sur $E = \mathbb{R}_+^2$ avec pour mesure dominante la mesure de Lebesgue $\mu(dx) = \mathbf{1}_E(x)dx$. La densité étant

$$\pi(x) = \pi(x_1, x_2) \propto \exp\{-(x_1 + x_2 + x_1x_2)\},$$

le support de π est bien le produit cartésien du support de ses marginales.

2. Cette condition de support est suffisante mais pas nécessaire. Modifions l'exemple problématique ci-dessus en considérant toujours $E := [-1,1] \times [-1,1]$ et considérons cette fois la loi π uniforme sur $\mathcal{S} := \mathcal{C}^- \cup \mathcal{C} \cup \mathcal{C}^+$ avec $\mathcal{C} := [-1/2, 1/2] \times [-1/2, 1/2]$ (voir Figure 2.5). Le support de π n'est toujours pas le produit cartésien des supports des marginales, mais l'échantillonneur de Gibbs est tout de même irréductible : on se convainc en effet facilement que de tout point de départ on peut atteindre tout borélien non négligeable en au plus deux coups, i.e. pour tout x dans le support de π et tout B tel que $\pi(B) > 0$, on a $P^2(x, B) > 0$.

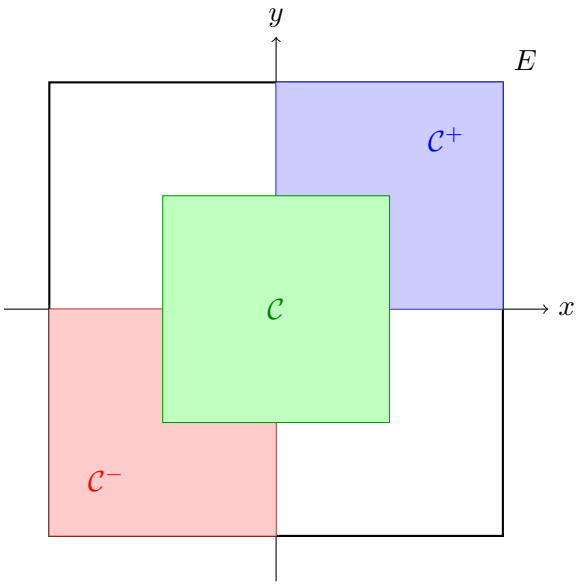

FIGURE 2.5 – Echantillonneur de Gibbs irréductible pour $\pi = \text{Unif}(\mathcal{C}^- \cup \mathcal{C} \cup \mathcal{C}^+)$ bien que le support de la loi jointe ne soit pas le produit cartésien des supports des marginales.

2.3 Illustration en statistique bayésienne

Rappelons le cadre bayésien :

— Le paramètre aléatoire $\boldsymbol{\theta}$ vit dans l'espace mesurable $(\Theta, \mathcal{F})$ et a pour loi a priori Π supposée absolument continue par rapport à une mesure μ :

$$\Pi(d\theta) = \pi(\theta)\mu(d\theta).$$

— Sachant $\boldsymbol{\theta} = \theta$, l'observation aléatoire X, vivant dans l'espace $(E, \mathcal{E})$, a pour loi P_θ et le modèle $(P_\theta)_{\theta \in \Theta}$ est dominé par une mesure ν :

$$P_\theta(dx) = p_\theta(x)\nu(dx).$$

La fonction p_θ est appelée la vraisemblance.

— Il en découle que la loi de X (ou marginale) est absolument continue par rapport à ν et a pour densité

$$f(x) = \int_\Theta p_\theta(x)\Pi(d\theta) = \int_\Theta p_\theta(x)\pi(\theta)\mu(d\theta).$$

— Dans ce cadre, on appelle loi a posteriori la loi de $\boldsymbol{\theta}$ sachant X, notée $\Pi(\cdot|X)$. Celle-ci est absolument continue par rapport à μ et a pour densité

$$\forall \theta \in \Theta, \qquad \pi(\theta|X) = \frac{p_\theta(X)\pi(\theta)}{f(X)}.$$

On souhaite simuler suivant cette loi a posteriori. En Section 1.3.1, nous avons mentionné deux situations favorables : d'une part si la loi a posteriori est une loi classique (cf. modèle fondamental gaussien) ; d'autre part si l'on sait simuler selon la loi a priori, calculer la vraisemblance en tout point ainsi que l'estimateur du maximum de vraisemblance (cf. modèle Cauchy-Gauss), auquel cas on peut appliquer une méthode de rejet.

Si nous ne sommes pas dans l'une de ces situations, que faire ? C'est ici qu'interviennent les méthodes MCMC. Sachant X, le principe est de construire une chaîne de Markov $(\boldsymbol{\theta}_n)$ à la fois irréductible, apériodique et invariante pour la loi a posteriori $\Pi(\cdot|X)$. Dans ce cas, la convergence en variation totale et le théorème ergodique s'appliqueront. L'idée de cette section est donc de préciser dans quelle mesure les méthodes MCMC vues précédemment s'adaptent à ce contexte bayésien.

2.3.1 Application de Metropolis

L'objectif est de simuler selon la loi a posteriori $\Pi(\cdot|X)$, dont la densité par rapport à la mesure μ est

$$\forall \theta \in \Theta, \qquad \pi(\theta|X) = \frac{p_\theta(X)\pi(\theta)}{f(X)} \propto p_\theta(X)\pi(\theta).$$

Donnons-nous un noyau de proposition $Q(\theta, d\theta')$ sur $(\Theta, \mathcal{F})$ absolument continu par rapport à μ :

$$Q(\theta, d\theta') = q(\theta, \theta')\mu(d\theta').$$

Il suffit alors de traduire l'Hypothèse 2.17 dans le contexte bayésien.

Hypothèse 2.29
Les conditions d'application sont les suivantes :

(a) Pour tout θ, savoir simuler selon le noyau $Q(\theta, \cdot)$.
(b) Pour tout couple (θ, θ'), savoir calculer $q(\theta, \theta')$.
(c) Pour tout θ, savoir calculer $p_\theta(X)\pi(\theta)$.

Si ces modalités sont satisfaites, alors le rapport de Metropolis est calculable :

$$\rho(\theta, \theta') = \frac{\pi(\theta'|X)q(\theta', \theta)}{\pi(\theta|X)q(\theta, \theta')} = \frac{p_{\theta'}(X)\pi(\theta')q(\theta', \theta)}{p_\theta(X)\pi(\theta)q(\theta, \theta')}.$$

Remarques :

1. Si le noyau de proposition Q est symétrique, le rapport de Metropolis se simplifie en
$$\rho(\theta, \theta') = \frac{p_{\theta'}(X)\pi(\theta')}{p_\theta(X)\pi(\theta)}.$$
2. On voit ici le rôle essentiel joué par le fait que l'algorithme de Metropolis ne requiert de connaître la loi cible qu'à une constante multiplicative près : la quantité compliquée dans la loi a posteriori est la marginale $f(X)$, or ici on n'en a nullement besoin !
3. Non seulement la marginale $f(X)$ disparaît dans le rapport de Metropolis, mais également toute constante multiplicative ne faisant pas intervenir θ (voir plus bas).

Rappelons la notation

$$\alpha(\theta, \theta') = \min\big(1, \rho(\theta, \theta')\big)$$

apparaissant dans l'algorithme de Metropolis de la Définition 2.18, que nous explicitons maintenant dans le cadre bayésien.

Définition 2.30 (Metropolis bayésien)
Considérons la suite de variables aléatoires $(\boldsymbol{\theta}_n)$ définie par une condition initiale $\boldsymbol{\theta}_0$ et séquentiellement comme suit : sachant $\boldsymbol{\theta}_n$,

1. *Simuler $\boldsymbol{\theta}'_{n+1} \sim Q(\boldsymbol{\theta}_n, \cdot)$*
2. *Simuler $U_{n+1} \sim \mathrm{Unif}([0,1])$ indépendamment de toutes les variables précédentes, puis*
   — *si $U_{n+1} \leq \alpha(\boldsymbol{\theta}_n, \boldsymbol{\theta}'_{n+1})$, poser $\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}'_{n+1}$;*
   — *sinon, poser $\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n$.*

Alors $(\boldsymbol{\theta}_n)$ est une chaîne de Markov admettant la loi a posteriori $\Pi(\cdot|X)$ comme loi réversible, donc invariante.

Exemple : Modèle Cauchy-Gauss. Revenons à l'exemple de la Section 1.3.1 où

$$\boldsymbol{\theta} \sim \text{Cauchy} \qquad \text{et} \qquad X|\boldsymbol{\theta} = (X_1, \ldots, X_N)|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, 1)^{\otimes N}.$$

Par rapport aux mesures dominantes de Lebesgue $\mu = \lambda$ et $\nu = \lambda_N$, les densités sont

$$\pi(\theta) = \frac{1}{\pi(1+\theta^2)} \qquad \text{et} \qquad p_\theta(X) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_j-\theta)^2}{2}} = \frac{e^{-\frac{N}{2}\left\{(\theta-\bar{X}_N)^2+V_X\right\}}}{(2\pi)^{\frac{N}{2}}},$$

d'où la densité a posteriori

$$\pi(\theta|X) \propto \frac{e^{-\frac{N}{2}(\theta-\bar{X}_N)^2}}{1+\theta^2}.$$

Si l'on prend comme noyau instrumental un noyau gaussien (symétrique) de la forme

$$q(\theta, \theta') = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\theta'-\theta)^2}{2\sigma^2}},$$

où σ est un paramètre de réglage, alors le rapport de Metropolis s'écrit

$$\rho(\theta, \theta') = \frac{1+\theta^2}{1+\theta'^2} \exp\left(\frac{N}{2}\left\{\left(\theta - \bar{X}_N\right)^2 - \left(\theta' - \bar{X}_N\right)^2\right\}\right) \tag{2.18}$$

Si la chaîne est en $\boldsymbol{\theta}_n = \theta$, l'algorithme commence par proposer $\boldsymbol{\theta}'_{n+1} = \theta'$ selon la loi $\mathcal{N}(\theta, \sigma^2)$, c'est-à-dire $\boldsymbol{\theta}'_{n+1} = \theta + \sigma W_{n+1}$ avec W_{n+1} gaussienne standard indépendante de toutes les variables précédentes. Il suffit ensuite de calculer le rapport de Metropolis et, en fonction de celui-ci, de voir si l'on accepte ou non cette transition.

Remarque : La forme produit en (2.18) montre que l'algorithme favorise les transitions réalisant un compromis entre la loi a priori et la vraisemblance, cette dernière étant incarnée ici par le ventre mou des données, c'est-à-dire leur moyenne empirique $\bar{X}_N$. Cependant, plus il y a de données et plus le second terme prend l'ascendant sur le premier. Par ailleurs, on retrouve un principe général de Metropolis déjà évoqué : un paramètre σ trop grand entraînera de nombreux refus, tandis qu'un paramètre σ trop faible entraînera de nombreuses transitions très petites.

On peut bien entendu adapter la Proposition 2.19 au contexte bayésien.

Corollaire 2.31 (Convergence du Metropolis bayésien)
Supposons $\Theta = \mathbb{R}^d$, Π et Q absolument continues par rapport à la mesure de Lebesgue, et la fonction $(\theta, \theta') \mapsto q(\theta, \theta')$ continue et strictement positive sur $\mathbb{R}^d \times \mathbb{R}^d$, alors la chaîne $(\boldsymbol{\theta}_n)$ est apériodique et $\Pi(\cdot|X)$-irréductible. Par conséquent, les Théorèmes 2.13 et 2.14 s'appliquent :

— *Pour toute fonction* $\varphi : E \to \mathbb{R}$ *telle que* $\Pi(|\varphi||X) < \infty$ *et pour* $\Pi(\cdot|X)$*-presque toute condition initiale* θ *:*

$$\frac{1}{n}\sum_{k=1}^{n}\varphi(\boldsymbol{\theta}_k) \xrightarrow[n\to\infty]{\text{p.s.}} \Pi(\varphi|X) = \mathbb{E}\left[\varphi(\boldsymbol{\theta})|X\right].$$

— *Pour* $\Pi(\cdot|X)$*-presque toute condition initiale* θ*, la loi de* $\boldsymbol{\theta}_n$ *tend vers* $\Pi(\cdot|X)$ *en variation totale :*

$$\|\text{Loi}(\boldsymbol{\theta}_n|X) - \Pi(\cdot|X)\| \xrightarrow[n\to\infty]{\text{p.s.}} 0.$$

Remarque : Le "p.s." dans la convergence en variation totale vient du fait que tout est conditionné par X, donc ces variations totales sont une suite de variables aléatoires à valeurs dans $[0, 1]$.

Exemple : Modèle Cauchy-Gauss. Nous sommes exactement dans le cadre d'application de ce résultat puisque la loi a priori est absolument continue sur $\mathbb{R}$ et

$$(\theta, \theta') \mapsto q(\theta, \theta') = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\theta'-\theta)^2}{2\sigma^2}}$$

est continue et strictement positive sur $\mathbb{R}^2$. Ainsi, la convergence en variation totale et le Théorème ergodique sont vérifiés pour $\Pi(\cdot|X)$-presque toute condition initiale. De plus, puisque

$$\forall \theta \in \mathbb{R} \qquad \pi(\theta|X) \propto p_\theta(X)\pi(\theta) > 0,$$

la loi a posteriori et la mesure de Lebesgue sont équivalentes, donc les résultats de convergence sont vrais pour presque toute condition initiale θ au sens usuel de la mesure de Lebesgue.

2.3.2 Application de Gibbs

En statistique bayésienne, supposons le paramètre aléatoire $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)$ multidimensionnel de loi a priori Π à densité par rapport à une mesure $\mu = \mu_1 \otimes \cdots \otimes \mu_d$, c'est-à-dire

$$\pi(d\theta) = \pi(\theta)\mu(d\theta) = \pi(\theta_1, \ldots, \theta_d)\mu_1(d\theta_1)\ldots\mu_d(d\theta_d).$$

La loi a posteriori $\Pi(\cdot|X)$ est alors elle-même absolument continue par rapport à μ, de densité

$$\pi(\theta|X) \propto p_\theta(X)\pi(\theta).$$

Pour construire une chaîne de Markov $(\boldsymbol{\theta}_n)$ de loi stationnaire $\Pi(\cdot|X)$, l'échantillonneur de Gibbs est tout désigné lorsque l'on sait simuler facilement suivant les densités conditionnelles $\pi(\cdot|\theta_{-\ell}, X)$.

Exemple jouet : Considérons le modèle suivant

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \sim \mathcal{N}(0,1) \otimes \text{Exp}(1) \qquad \text{et} \qquad X|\boldsymbol{\theta} = (X_1, \ldots, X_N)|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2^{-1})^{\otimes N}.$$

La densité a posteriori par rapport à la mesure de Lebesgue sur $E := \mathbb{R}\times]0, \infty[$ est donc

$$\pi(\theta|X) = \pi(\theta_1, \theta_2|X) \propto \theta_2^{\frac{N}{2}} \exp\left(-\frac{1}{2}\left\{\theta_1^2 + \theta_2\left(2 + \sum_{j=1}^{N}(\theta_1 - X_j)^2\right)\right\}\right), \tag{2.19}$$

où il n'y a pas besoin de l'indicatrice $\mathbf{1}_{\theta_2>0}$ puisque l'espace d'états est $E := \mathbb{R}\times]0, \infty[$. Ce n'est pas une loi classique. Pour la loi a priori, les variables $\boldsymbol{\theta}_1$ et $\boldsymbol{\theta}_2$ sont indépendantes, mais bien entendu ce n'est plus le cas pour la loi a posteriori. Cependant, puisque

$$\pi(\theta_1|\theta_2, X) \propto \pi(\theta_1, \theta_2|X) \quad \text{et} \quad \pi(\theta_2|\theta_1, X) \propto \pi(\theta_1, \theta_2|X)$$

on voit facilement que

$$\text{Loi}(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, X) = \mathcal{N}\left(\frac{N\bar{X}_N\boldsymbol{\theta}_2}{1+N\boldsymbol{\theta}_2}, \frac{1}{1+N\boldsymbol{\theta}_2}\right),$$

et

$$\text{Loi}(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, X) = \text{Gamma}\left(1+\frac{N}{2}, 1+\frac{1}{2}\sum_{j=1}^{N}(\boldsymbol{\theta}_1 - X_j)^2\right).$$

Ces deux lois conditionnelles étant aisément simulables, l'échantillonneur de Gibbs s'implémente sans peine : il suffit d'alterner des étapes de simulations de lois normales et de lois Gamma.

Nous pouvons maintenant traduire l'Hypothèse 2.22, en gardant en tête que l'observation aléatoire X est fixée et que tout se fait sachant X. Rappelons au préalable que

$$\pi_{-\ell}(\theta_{-\ell}|X) = \int_{E_\ell} \pi(\theta_\ell, \theta_{-\ell}|X)\mu_\ell(d\theta_\ell).$$

Hypothèse 2.32
Pour tout ℓ et tout $\theta_{-\ell}$ tel que $\pi_{-\ell}(\theta_{-\ell}|X) > 0$, nous supposons savoir simuler suivant la densité conditionnelle $\pi(\cdot|\theta_{-\ell}, X)$.

Dans le cadre bayésien, l'échantillonneur de Gibbs à balayage séquentiel s'écrit alors comme suit.

Définition 2.33 (Gibbs bayésien)
Considérons la suite de variables aléatoires $(\boldsymbol{\theta}_n)$ définie par une condition initiale $\boldsymbol{\theta}_0$ et séquentiellement comme suit : sachant $\boldsymbol{\theta}_n = \theta$,

1. *Pour ℓ allant de 1 à d, simuler successivement $\boldsymbol{\theta}'_\ell = \theta'_\ell$ selon la densité*

$$\pi(\cdot|\theta'_1, \ldots, \theta'_{\ell-1}, \theta_{\ell+1}, \ldots, \theta_d, X).$$

2. *Poser $\boldsymbol{\theta}_{n+1} = (\theta'_1, \ldots, \theta'_d)$.*

Alors $(\boldsymbol{\theta}_n)$ est une chaîne de Markov admettant la loi a posteriori $\Pi(\cdot|X)$ comme loi invariante.

Exemple jouet : Sachant $\boldsymbol{\theta}_n = (\theta_1, \theta_2)$,

1. Simuler $\boldsymbol{\theta}'_1 = \theta'_1$ selon la loi

$$\mathcal{N}\left(\frac{N\bar{X}_N\theta_2}{1+N\theta_2}, \frac{1}{1+N\theta_2}\right)$$

puis $\boldsymbol{\theta}'_2 = \theta'_2$ selon la loi

$$\text{Gamma}\left(1+\frac{N}{2}, 1+\frac{1}{2}\sum_{j=1}^{N}(\theta'_1 - X_j)^2\right).$$

2. Poser $\boldsymbol{\theta}_{n+1} = (\theta'_1, \theta'_2)$.

Concernant les résultats de convergence, nous pouvons en particulier rappeler la Proposition 2.28.

Proposition 2.34 (Convergence du Gibbs Bayésien)
Si le support de la loi a posteriori est le produit cartésien des supports des lois marginales a posteriori, alors la chaîne $(\boldsymbol{\theta}_n)$ est $\Pi(\cdot|X)$-irréductible et apériodique. De plus :

— *Pour toute fonction* $\varphi : E \to \mathbb{R}$ *telle que* $\Pi(|\varphi||X) < \infty$ *et pour* $\Pi(\cdot|X)$*-presque toute condition initiale* θ *:*

$$\frac{1}{n}\sum_{k=1}^{n}\varphi(\boldsymbol{\theta}_k) \xrightarrow[n\to\infty]{\text{p.s.}} \Pi(\varphi|X) = \mathbb{E}\left[\varphi(\boldsymbol{\theta})|X\right].$$

— *Pour* $\Pi(\cdot|X)$*-presque toute condition initiale* θ*, la loi de* $\boldsymbol{\theta}_n$ *tend vers* $\Pi(\cdot|X)$ *en variation totale :*

$$\|\text{Loi}(\boldsymbol{\theta}_n|X) - \Pi(\cdot|X)\| \xrightarrow[n\to\infty]{\text{p.s.}} 0.$$

Exemple jouet : La formule (2.19) pour la densité a posteriori $\pi(\theta|X)$ montre que $\pi(\theta|X) > 0$ pour tout $\theta \in E = \mathbb{R}\times]0,\infty[$, qui est donc le support de la loi a posteriori. Par ailleurs, pour tout $\theta_1 \in \mathbb{R}$ et tout $\theta_2 \in \times]0,\infty[$, on a bien sûr

$$\pi(\theta_1|X) = \int_{]0,\infty[} \pi(\theta|X)d\theta_2 > 0 \qquad \text{et} \qquad \pi(\theta_2|X) = \int_{\mathbb{R}} \pi(\theta|X)d\theta_1 > 0.$$

La condition de la Proposition 2.34 est donc satisfaite. De plus, le "$\Pi(\cdot|X)$-presque toute condition initiale" se traduit en "presque toute condition initiale θ dans $E = \mathbb{R}\times]0,\infty[$" au sens de Lebesgue.

Bibliographie

[1] A. Ben-Hamou. *Statistiques bayésiennes*. Sorbonne Université, 2025.

[2] B. Bercu et D. Chafaï. *Modélisation stochastique et simulation*. Dunod, 2007.

[3] T. Bodineau. *Modélisation de phénomènes aléatoires*. Ecole Polytechnique, 2022.

[4] D. Chafaï. *Processus stochastiques*. Ecole Normale Supérieure, 2026.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.

[6] B. Delyon. *Simulation et modélisation*. Université de Rennes, 2017.

[7] L. Devroye. *Nonuniform random variate generation*. Springer, 1986.

[8] R. Douc, E. Moulines, P. Priouret, and P. Soulier. *Markov Chains*. Springer, 2018.

[9] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1) :97–109, April 1970.

[10] W. Hoermann. The transformed rejection method for generating Poisson random variables. *Insurance : Mathematics and Economics*, 12, 39-45, 1993.

[11] P. L'Ecuyer. Random number generation. *Handbook of computational statistics*, 35–71 Springer, 2012.

[12] T. Lévy. *Probabilités Approfondies*. Sorbonne Université, 2025.

[13] G. Marsaglia and W. W. Tsang. The Ziggurat Method for Generating Random Variables. *Journal of Statistical Software*, 5(8) :1–7, 2000.

[14] M. Matsumoto and T. Nishimura. Mersenne twister : A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Model. Comput. Simul.*, 8(1) :3–30, January 1998.

[15] 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.

[16] S. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer, 2009.

[17] H. Niederreiter. *Random number generation and quasi-Monte Carlo methods*. SIAM, 1992.

[18] J. R. Norris. *Markov Chains*. Cambridge University Press, 1997.

[19] C. P. Robert. *The Bayesian Choice*. Springer, 2001.

[20] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

[21] G. O. Roberts and J. S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1 :20–71, 2004.

[22] B. Tuffin. *La simulation de Monte Carlo*. Hermes Science Publications, 2010.