Algorithmes stochastiques

ISAE DSYS 315

Benjamin Bobbia et Xavier Gendre

Institut Supérieur de l'Aéronautique et de l'Espace

SUPAERO

Algorithmes stochastiques

ISAE DSYS 315

Benjamin Bobbia et Xavier Gendre

Table des matières

1 — Motivations

1.1 Introduction

De façons générale, un *algorithme* est une séquence finie d'opérations qui permet d'obtenir un résultat. Il s'agit donc d'une méthode à suivre étape par étape afin de cheminer d'un état initial à un état final dans lequel un problème donné est résolu. Un exemple classique d'algorithme est celui du *tri rapide* qui permet de trier un ensemble d'entiers distincts par ordre croissant : la séquence d'opérations consiste à définir un élément dans la liste, appelé le *pivot*, et à trier séparément la sous-liste des éléments inférieurs au pivot et celle des éléments supérieurs au pivot.

ALGORITHME 1.1 – Tri rapide

Initialisation : une liste $x = \{x_1, \ldots, x_n\}$ d'entiers distincts
Si $n \leqslant 1$:
    Retourner x
Sinon :
    Définir le pivot $p = x_1$
    Extraire la liste x' des éléments inférieurs à p
    Extraire la liste x'' des éléments supérieurs à p
    Retourner $\{\text{TriRapide}(x'),\ p,\ \text{TriRapide}(x'')\}$

Cette formulation de l'algorithme du tri rapide est dite *récursive* car elle fait appel à elle-même pour se définir. Il existe également des versions non récursives du tri rapide que nous ne considérerons pas ici. L'étude générale des algorithmes est l'objet de l'*algorithmique* et il est possible de prouver formellement différentes propriétés telles que la *correction* ou la *finitude* d'un algorithme. Nous ne développerons pas cette théorie dans ce cours mais nous attirons l'attention du lecteur sur la notion de *complexité en temps* qui permet de donner l'ordre de grandeur du temps nécessaire à un algorithme pour obtenir son résultat en fonction de ses paramètres d'entrée. Dans le cas du tri rapide, cette complexité se mesure par rapport à la taille n de la liste à trier. En notant $T(n)$ le temps nécessaire pour trier n entiers distincts, nous obtenons la relation de récurrence

$$T(n) = T(n') + T(n'') + an$$

où n' (resp. n'') est le nombre d'entiers inférieurs (resp. supérieurs) au pivot et an correspond au temps (linéaire) de parcours de la liste pour en extraire les deux sous-listes. Au mieux, le

pivot sépare la liste à trier en deux sous-listes de même taille et la complexité devient

$$T(n) = 2T(n/2) + an.$$

Par itérations successives, il vient que la complexité en temps $T(n)$ du tri rapide dans ce cas favorable est de l'ordre de $n\ln(n)$. Cependant, il est également possible que les deux sous-listes soient de tailles très déséquilibrées, par exemple si la liste initiale est déjà triée. Le pire cas correspond à une relation de la forme

$$T(n) = T(1) + T(n-1) + an$$

qui donne lieu à une complexité en temps bien moins bonne de l'ordre de n^2.

Cette brève étude de la complexité en temps montre que les performances du tri rapide implanté selon l'algorithme 1.1 dépendent fortement de la liste à trier et peuvent être très mauvaises si il est appliqué systématiquement à des listes qui ne lui sont pas favorables. Idéalement, nous aimerions pouvoir dire que cet algorithme se comporte « généralement bien » quelle que soit la liste à trier. Pour cela, l'idée va consister à introduire un aléa dans l'algorithme de façon à assurer de bonnes performances en moyenne. En effet, par un raisonnement similaire à ce que nous avons fait ci-dessus, il est possible de montrer que $T(n)$ reste de l'ordre de grandeur idéal de $n\ln(n)$ si les tailles n' et n'' sont toutes les deux supérieures à $n/4$. Autrement dit, en tirant un pivot au hasard uniformément dans la liste, nous avons plus d'une chance sur deux de tomber sur un découpage favorable. Pour cette façon de faire, appelée *tri rapide aléatoire*, la quantité $T(n)$ est une variable aléatoire et nous pouvons montrer que son espérance $\mathbb{E}[T(n)]$ est bien de l'ordre de $n\ln(n)$ quelle que soit la liste de taille n à trier (voir la figure 1.1 dans le cas défavorable d'une liste déjà triée).

ALGORITHME 1.2 – Tri rapide aléatoire

Initialisation : une liste $x = \{x_1, \ldots, x_n\}$ d'entiers distincts
Si $n \leqslant 1$:
  Retourner x
Sinon :
  Tirer le pivot p uniformément dans x
  Extraire la liste x' des éléments inférieurs à p
  Extraire la liste x'' des éléments supérieurs à p
  Retourner $\{$TriRapideAleatoire(x'), p, TriRapideAleatoire(x'')$\}$

Un algorithme qui utilise une source de hasard est appelé *algorithme probabiliste* ou *algorithme stochastique*. Dans le cas du tri rapide aléatoire, l'aléa est utilisé pour sélectionner le pivot mais le résultat est déterministe et correspond toujours à la liste triée attendue. Ce cas particulier d'algorithme stochastique donnant toujours un résultat exact est appelé *algorithme de Las Vegas*. Plus généralement, l'utilisation du hasard n'est pas limitée à des considérations sur la complexité en temps et le résultat d'un algorithme stochastique est souvent une variable aléatoire. L'objet de ce cours est de proposer une étude de tels algorithmes stochastiques utiles pour des problèmes de *simulation*, d'*estimation* et d'*optimisation*.

1.2 Simulation de variables aléatoires

Pour définir l'algorithme du tri rapide aléatoire, nous avons invoqué l'opération « Tirer le pivot p uniformément dans x » sans aucun commentaire et sans prendre en compte sa

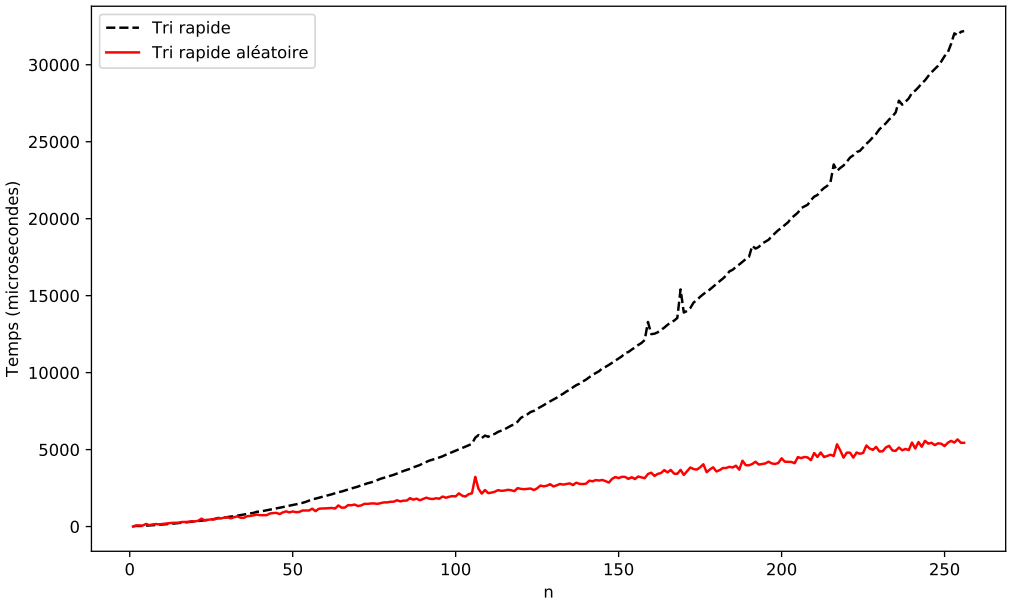

FIGURE 1.1 – Estimation de la complexité en temps pour le tri de la suite ordonnée $\{1,\dots,n\}$ avec les algorithmes de tri rapide et de tri rapide aléatoire.

complexité. Cependant, la capacité à produire du hasard ne va pas de soi et il est raisonnable de se demander si cette étape de l'algorithme est effectivement négligeable.

Étant donné une loi de probabilité μ, le *problème de la simulation* consiste à trouver une méthode pour produire des réalisations indépendantes et de même loi μ. Cette problématique est centrale en ce qui concerne l'objet de ce cours puisque tous les algorithmes stochastiques nécessiteront la génération d'aléas. Plus généralement, ce problème se retrouve dans d'autres cadres mathématiques tels que la statistique bayésienne où il peut être nécessaire de produire des échantillons selon une loi a posteriori complexe, par exemple.

Le lecteur intéressé par une introduction plus poussée au problème de la simulation pourra se référer à l'excellent ouvrage [**?**].

1.2.1 Loi uniforme

Les algorithmes considérés dans ce cours ont vocation à être mis en œuvre à l'aide de moyens informatiques. L'unité élémentaire manipulée par un ordinateur est le *bit* que nous représentons communément comme un entier 0 ou 1. De fait, la première loi de probabilité qu'il convient de savoir simuler est la loi μ uniforme sur $\{0,1\}$, appelée loi de Bernoulli $\mathscr{B}(1/2)$,

$$\mu(0) = \mu(1) = \frac{1}{2}.$$

Comme il n'existe pas de véritable générateur de hasard, les méthodes utilisées par les ordinateurs pour simuler μ sont imparfaites et se basent sur des algorithmes déterministes qui produisent récursivement des suites de 0 et de 1 à partir d'un état initial appelé la *graine* (ou *seed* en anglais). Connaissant la graine utilisée, la suite générée est reproductible et prédictible mais elle passe une série de tests statistiques, appelés *tests de Diehard* et proposés par

Marsaglia en 1995, qui valident certaines propriétés qui seraient naturelles pour de véritables suites de variables aléatoires indépendantes. Le fait qu'une même graine induise toujours le même aléa peut être utile en pratique pour rejouer certains événements perçus comme aléatoires lorsque nous souhaitons tester des procédures statistiques. L'étude de ces méthodes prédictibles sort du cadre de ce cours mais nous en connaissons de très performantes qui nous permettent de considérer que nous disposons effectivement de très grandes séquences de réalisations indépendantes de la loi de Bernoulli $\mathscr{B}(1/2)$ sur un ordinateur. En pratique, l'algorithme *Mersenne Twister*, introduit en 1997 par Matsumoto et Nishimura, est très utilisé.

En disposant de réalisations d'une loi de Bernoulli de paramètre $1/2$, il est possible de résoudre une approximation du problème de la simulation pour une loi uniforme sur $[0,1]$. En effet, pour un entier $K > 0$ donné, si $B_1, \ldots, B_K$ sont des variables indépendantes et de même loi $\mathscr{B}(1/2)$, alors la variable aléatoire

$$U_K = \sum_{k=1}^{K} B_k 2^{-k}$$

suit une loi uniforme sur l'ensemble dyadique $\{0, 1/2^K, \ldots, (2^K-1)/2^K\}$. La loi de la variable U_K approche la loi uniforme sur $[0,1]$ au sens où, si $X \sim \mathscr{U}([0,1])$, tout événement défini sur cette grille dyadique aura la même probabilité pour U_K et pour X. Cette propriété découle du calcul élémentaire suivant

$$\forall \ell \in \{0, \ldots, 2^K - 1\},\ \mathbb{P}(U_K < \ell/2^K) = \sum_{k=0}^{\ell-1} \mathbb{P}(U_K = k/2^K) = \frac{\ell}{2^K} = \mathbb{P}(X < \ell/2^K).$$

Pour un ordinateur, la représentation des nombres à virgule flottante en binaire est définie par la norme *IEEE 754* et cela induit l'existence d'une plus petite valeur représentable avec un nombre de bits fixés (*e.g.* avec 32 bits, cette valeur vaut $1,4 \times 10^{-45}$). En considérant une valeur de K suffisamment grande pour avoir une résolution dyadique inférieure à cette valeur, les réalisations de la variables U_K deviennent indiscernables de réalisations d'une véritable loi $\mathscr{U}([0,1])$ sur un ordinateur. Bien que cette construction ne corresponde pas exactement à ce qui est réellement fait dans les logiciels que nous utilisons, le principe demeure et nous pouvons donc également considérer que nous disposons d'un très grand nombre de réalisations indépendantes de la loi uniforme $\mathscr{U}([0,1])$.

Une fois le problème de la simulation résolu pour la loi uniforme sur $[0,1]$, il en va de même pour celui de la loi $\mathscr{U}([a,b])$, pour tout $a < b$. En effet, si $U \sim \mathscr{U}([0,1])$, alors

$$V = a + (b-a)U \sim \mathscr{U}([a,b])$$

et nous pouvons obtenir des réalisations indépendantes de V à partir de celles de U. De même, la loi uniforme sur un pavé $[a_1,b_1] \times \cdots \times [a_d,b_d] \subset \mathbb{R}^d$ avec $a_i < b_i$, pour tout $i \in \{1, \ldots, d\}$, s'obtient en simulant indépendamment chaque coordonnée $V_i \sim \mathscr{U}([a_i,b_i])$,

$$(V_1, \ldots, V_d) \sim \mathscr{U}([a_1,b_1] \times \cdots \times [a_d,b_d]).$$

1.2.2 Méthode d'inversion

Considérons une loi μ sur $\mathbb{R}$ et sa fonction de répartition F définie par

$$\forall x \in \mathbb{R},\ F(x) = \mu(]-\infty, x]).$$

La fonction $F : \mathbb{R} \to [0,1]$ est croissante et continue à droite avec limite à gauche.

Définition 1.3. Soit μ une loi sur $\mathbb{R}$ de fonction de répartition F, l'*inverse généralisée* de F est la fonction F^{-1} définie par

$$\forall 0 < u \leqslant 1,\ F^{-1}(u) = \inf\{x \in \mathbb{R} \text{ tel que } F(x) \geqslant u\}.$$

La fonction F^{-1} est également appelée la *fonction quantile* de la loi μ.

Lorsque la fonction de répartition F est inversible, son inverse généralisée F^{-1} coïncide avec l'inverse de F. Cette définition permet d'énoncer le résultat suivant qui illustre le rôle privilégié de la loi uniforme sur $[0,1]$ dans les problèmes de simulation de variables aléatoires.

Théorème 1.4 (Méthode d'inversion)**.** *Soit μ une loi sur $\mathbb{R}$ de fonction de répartition F. Si U est une variable uniforme sur $[0,1]$, alors la variable aléatoire $F^{-1}(U)$ suit la loi μ.*

Démonstration. Soient $x \in \mathbb{R}$ et $0 < u \leqslant 1$, montrons que

$$u \leqslant F(x) \iff F^{-1}(u) \leqslant x.$$

Si $u \leqslant F(x)$, alors $F^{-1}(u) \leqslant x$ par définition de F^{-1}. Inversement, si $F^{-1}(u) \leqslant x$, alors nous déduisons $F(F^{-1}(u)) \leqslant F(x)$ par croissance de F. Comme la fonction F est continue à droite, nous avons toujours $u \leqslant F(F^{-1}(u))$, ce qui implique $u \leqslant F(x)$ et prouve l'équivalence annoncée. Ainsi, puisque $F(x) \in [0,1]$, nous pouvons écrire

$$\mathbb{P}(F^{-1}(U) \leqslant x) = \mathbb{P}(U \leqslant F(x)) = F(x).$$

La fonction de répartition de la variable $F^{-1}(U)$ est égale à F, donc la loi de cette variable est μ. □

ALGORITHME 1.5 – Méthode d'inversion
Initialisation : la fonction quantile F^{-1} de la loi μ
À l'étape n :
  Tirer U_n selon la loi $\mathscr{U}([0,1])$
  Calculer $X_n = F^{-1}(U_n)$
Retourner les valeurs $X_1, X_2, \ldots$

EXEMPLE 1.6 (Simulation de la loi exponentielle). Soit $\lambda > 0$, la loi exponentielle de paramètre λ, notée $\mathscr{E}(\lambda)$, admet la densité suivante par rapport à la mesure de Lebesgue,

$$\forall x \in \mathbb{R},\ f(x) = \lambda e^{-\lambda x} \mathbf{1}_{x>0}$$

et sa fonction de répartition est donnée par

$$\forall x \in \mathbb{R},\ F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{si } x > 0, \\ 0 & \text{sinon.} \end{cases}$$

En inversant cette fonction sur $\mathbb{R}_+^*$, nous obtenons l'inverse généralisée

$$\forall 0 < u \leqslant 1,\ F^{-1}(u) = -\frac{1}{\lambda}\ln(1-u).$$

Soit $U \sim \mathscr{U}([0,1])$, comme $1-U$ suit également une loi uniforme sur $[0,1]$, nous déduisons par la méthode d'inversion que

$$-\frac{1}{\lambda}\ln(U) \sim \mathscr{E}(\lambda).$$

La figure 1.2 illustre ce résultat en simulant la loi exponentielle $\mathscr{E}(1)$. ■

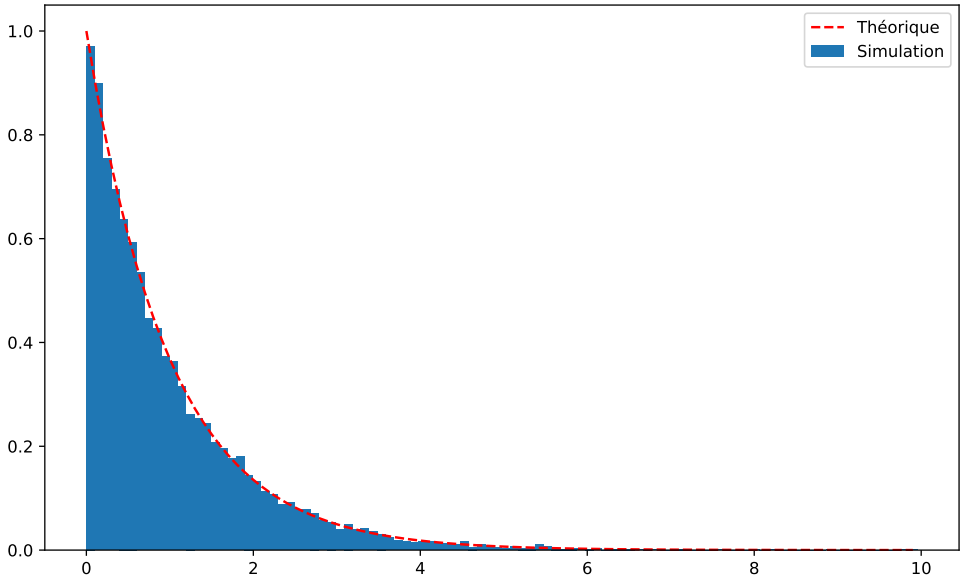

FIGURE 1.2 – Simulation de la loi $\mathscr{E}(1)$ par la méthode d'inversion.

D'un point de vue théorique, la méthode d'inversion est séduisante car elle est simple et fournit une simulation exacte. Cependant, en pratique, elle n'est pas toujours satisfaisante à cause de certains obstacles. Tout d'abord, l'inverse généralisée F^{-1} peut ne pas être explicite ou être trop coûteuse à évaluer. D'autre part, la loi uniforme sur $[0,1]$ étant généralement générée sur une grille comme nous l'avons vu précédemment, les parties plates de la fonction de répartition F sont plus difficilement atteinte, ce qui entraîne une raréfaction (ou parfois même une absence totale) de certaines valeurs possibles.

Il est intéressant de remarquer comment s'applique la méthode d'inversion pour une loi μ sur $\mathbb{R}$ à support discret, *i.e.* la simulation d'une variable X à valeurs dans un ensemble $\{x_k,\ k \in \mathbb{N}\}$ au plus dénombrable avec $x_k < x_{k+1}$, pour tout $k \in \mathbb{N}$. Dans ce cadre, la fonction de répartition F est une fonction constante par morceaux dont les paliers sont donnés par la suite $(p_k)_{k\in\mathbb{N}}$ définie par $p_0 = 0$ et

$$\forall k > 0,\ p_k = \mathbb{P}(X \leqslant x_{k-1}) = \sum_{\ell=0}^{k-1} \mu(x_\ell).$$

Nous en déduisons que l'inverse généralisée est également une fonction constante par morceaux,

$$\forall 0 < u \leqslant 1,\ F^{-1}(u) = x_k \iff p_k < u \leqslant p_{k+1}.$$

Afin de résoudre le problème de la simulation pour μ, il suffit donc de considérer une variable $U \sim \mathscr{U}([0,1])$ et de définir $F^{-1}(U)$ comme la valeur x_k correspondant à la tranche $]p_k, p_{k+1}]$ dans laquelle se trouve U. Bien que l'inverse généralisée soit ici explicite, la méthode souffrira encore du manque de précision avec lequel U est simulée et cela aura un impact important sur les valeurs générées car certaines seront manquantes. Pour contourner ce problème en pratique, d'autres méthodes de simulation sont souvent envisagées.

EXEMPLE 1.7 (Loi de Bernoulli). Une application immédiate de ce cas particulier est la simulation d'une loi de Bernoulli de paramètre $p \in [0,1]$. Ainsi, pour $U \sim \mathscr{U}([0,1])$, la variable X définie par

$$X = \begin{cases} 0 & \text{si } U \leqslant 1-p, \\ 1 & \text{sinon.} \end{cases}$$

suite une loi $\mathscr{B}(p)$. ■

EXEMPLE 1.8 (Loi uniforme discrète). Soit un entier $n > 0$, le cas de la loi μ uniforme sur $\{1,\dots,n\}$ conduit à considérer $p_0 = 0$ et

$$\forall k \in \{1,\dots,n\},\ p_k = \sum_{\ell=1}^{k} \mu(\ell) = \frac{k}{n}.$$

À partir de $U \sim \mathscr{U}([0,1])$, la méthode d'inversion permet de construire X de loi μ par

$$\forall k \in \{1,\dots,n\},\ X = k \iff k-1 < nU \leqslant k.$$

Autrement dit, nous obtenons $X = \lfloor nU \rfloor + 1$ où $\lfloor \cdot \rfloor$ désigne la partie entière. ■

EXEMPLE 1.9 (Loi géométrique). La loi géométrique de paramètre $p \in [0,1]$, notée $\mathscr{G}(p)$, est définie sur $\mathbb{N} \setminus \{0\}$ par

$$\forall k > 0,\ \mu(k) = (1-p)^{k-1} p.$$

En considérant $U \sim \mathscr{U}([0,1])$, la méthode d'inversion conduit à définir une variable X par

$$\forall k > 0,\ X = k \iff 1-(1-p)^{k-1} < U \leqslant 1-(1-p)^k.$$

En pratique, quitte à faire appel à la fonction transcendante ln, il peut être utile d'écrire la variable X sous une forme différente,

$$X = \left\lfloor \frac{\ln(U)}{\ln(1-p)} \right\rfloor + 1.$$

où nous avons utilisé le fait que $1-U \sim \mathscr{U}([0,1])$. ■

1.2.3 Méthode de Box-Muller

Au vu de la prépondérance de la loi normale en statistiques il semble important de savoir la simuler facilement. Étant donnée une variable aléatoire $X \sim \mathcal{N}(0,1)$, et $(\mu,\sigma) \in \mathbb{R} \times \mathbb{R}_+$ on sait que $\mu + \sigma X$ suit une loi $\mathcal{N}(\mu,\sigma^2)$. Ainsi pour simuler n'importe qu'elle loi normale il suffit de savoir simuler la loi normale centrée et réduite. La méthode méthode d'inversion est difficilement applicable car la fonction de répartition de la loi normale n'est pas inversible explicitement. Il n'en existe même pas de version analytique.

Néanmoins la méthode du rejet permet de fournir une intuition si on cherche à générer une variable aléatoire normale en dimension 2. Considérons (X,Y) un vecteur gaussien centré et de matrice de covariance

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

La loi d'un tel vecteur gaussien est invariante par rotation. Notons pour $\theta \in [0,2\pi]$,

$$R_\theta = \begin{pmatrix} cos(\theta) & -sin(\theta) \\ sin(\theta) & cos(\theta) \end{pmatrix}$$

la matrice de rotation d'angle θ. On sait que que les vecteurs gaussien sont stable par transformation linéaire, i.e $R_\theta(X,Y)^T$ est aussi un vecteur gaussien. Son espérance est alors $R_\theta(0,0)^T$ et sa matrice de covariance

$$R_\theta \Sigma R_\theta^T = \Sigma.$$

Ainsi pour simuler la loi de (X,Y), il suffit de tirer un angle uniformément sur $[0,2\pi]$ puis de tirer un rayon. Ce dernier peut être tiré en utilisant la méthode du rejet. En effet, la distance entre (X,Y) et $(0,0)$ est $R = \sqrt{X^2+Y^2}$. Il ne reste plus qu'a trouver la fonction de répartition de R.

Soit h une fonction continue bornée, on a alors

$$\begin{aligned}
\mathbb{E}(h(R)) =& \mathbb{E}(h(\sqrt{X^2+Y^2})) \\
=& \int_{\mathbb{R}^2} h(\sqrt{x^2+y^2}) \frac{1}{2\pi} e^{\frac{-(x^2+y^2)}{2}} dxdy \\
=& \int_{\mathbb{R}^2} h(r) \frac{r}{2\pi} e^{\frac{-r^2}{2}} drd\theta. \text{ Changement de variable en polaire.}
\end{aligned}$$

Ainsi la densité de R est donnée par la fonction

$$r \mapsto r e^{\frac{-r^2}{2}}.$$

Ce qui conduit à la fonction de répartition

$$F_R(r) = 1 - e^{\frac{-r^2}{2}}.$$

En résumé, pour généré une réalisation du couple (X,Y), on tire deux variables aléatoires U et V uniforme sur $[0,1]$. L'angle est alors $\Theta = 2\pi V$ et le rayon $R = \sqrt{2ln(U)}$. On obtient alors (X,Y) en prenant les coordonnées cartésiennes de (R,Θ). Plus formellement, on obtient le lemme suivant.

Lemme 1.10. *Soient U et V deux variables aléatoires de même loi $\mathscr{U}([0,1])$ et indépendantes. Alors les variables aléatoires (X,Y) définies par*

$$X = \sqrt{-2ln(U)}cos(2\pi V) \quad et \quad Y = \sqrt{-2ln(U)}sin(2\pi V),$$

suivent des lois normales $\mathscr{N}(0,1)$ indépendantes.

Démonstration. Soit h une fonction continue bornée, on a

$$\begin{aligned}\mathbb{E}(h(X,Y)) =& \mathbb{E}(h(\sqrt{-2ln(U)}cos(2\pi V), \sqrt{-2ln(U)}sin(2\pi V))) \\ =& \int_0^1 \int_0^1 h(\sqrt{-2ln(u)}cos(2\pi v), \sqrt{-2ln(u)}sin(2\pi v))dudv.\end{aligned}$$

On applique le changement de variable

$$\begin{cases} x &= \sqrt{-2ln(u)}cos(2\pi v) \\ y &= \sqrt{-2ln(u)}sin(2\pi v) \end{cases}$$

La fonction $\varphi : (u,v) \mapsto (x,y)$ est un difféomorphisme C^∞ de $[0,1]^2$ dans $\mathbb{R}^2 \backslash \mathbb{R}_-$ d'inverse

$$\varphi^{-1} : (x,y) \mapsto \left(e^{\frac{-x^2-y^2}{2}}, \frac{arctan\left(\frac{y}{x}\right)}{2\pi} \right).$$

Ce qui permet de conclure. □

ALGORITHME 1.11 – Algorithme de Box-Muller

Initialisation :
- k le nombre souhaité de réalisations

Pour i de 1 à $\lfloor k/2 \rfloor$:

Tirer (U,V) deux variables aléatoires uniformes indépendantes
- $X_{2i} = \sqrt{2ln(U)}cos(2\pi V)$
- $X_{2i+1} = \sqrt{2ln(U)}sin(2\pi V)$

Retourner les valeurs $X_1, \ldots, X_k$

La figure 1.3 illustre un histogramme d'un échantillon d'une loi $\mathscr{N}(0,1)$ généré par cet algorithme.

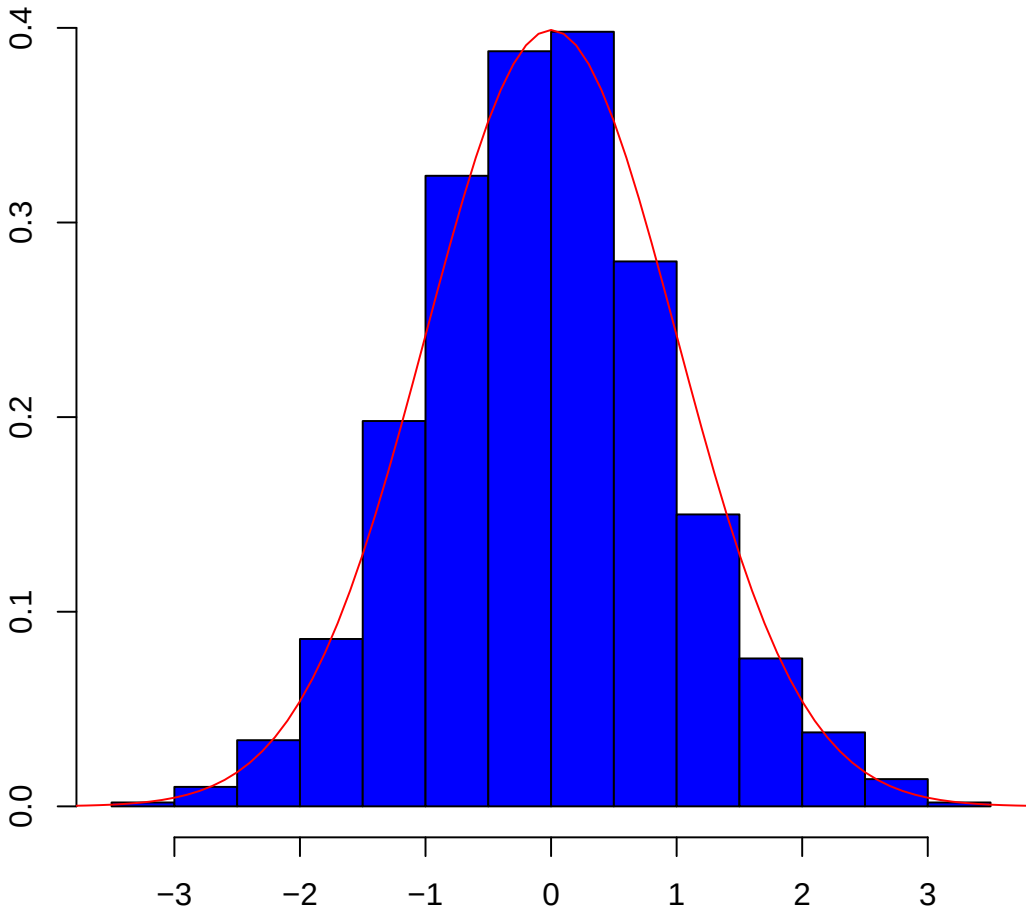

FIGURE 1.3 – Simulation d'une loi normale par l'algorithme de Box-Muller.

1.2.4 Mélanges de lois

Un *modèle de mélange* correspond au cas fréquemment rencontré en pratique où la loi d'un variable aléatoire X peut se comprendre comme une somme de différentes lois. Pour les *mélanges finis ou discrets*, nous pouvons considérer une variable aléatoire Y à valeurs dans $\mathbb{N}$ et une collection de lois de probabilité $\{\mu_k,\ k \in \mathbb{N}\}$. En notant,

$$\forall k \in \mathbb{N},\ \mathbb{P}(Y = k) = p_k,$$

nous définissons la loi de X comme le *mélange* des lois μ_k par la loi de Y en posant, pour tout ensemble mesurable A,

$$\mathbb{P}(X \in A) = \sum_{k \in \mathbb{N}} p_k \mu_k(A).$$

Pour tout $k \in \mathbb{N}$, la loi μ_k est donc la loi de X conditionnellement à l'événement $\{Y = k\}$. Si X est une variable aléatoire réelle, cette décomposition en combinaison linéaire se retrouve dans

sa fonction de répartition,

$$\forall x \in \mathbb{R},\ F(x) = \mathbb{P}(X \leqslant x) = \sum_{k \in \mathbb{N}} p_k F_k(x)$$

où, pour tout $k \in \mathbb{N}$, $F_k(x) = \mu_k(]-\infty, x])$ est la fonction de répartition de la loi μ_k. Les mélanges discrets sont très utilisés en modélisation et ce sont des structures importantes à déceler pour simplifier la simulation de la loi de X. En effet, pour résoudre le problème de la simulation pour la loi mélangée, il suffit de simuler une réalisation k de Y et de générer une réalisation de X grâce à la loi μ_k.

EXEMPLE 1.12 (Modèle de mélange gaussien). Un mélange de n lois gaussiennes sur $\mathbb{R}$ est une loi μ donnée par, pour tout ensemble mesurable A,

$$\mu(A) = \sum_{k=1}^{n} p_k \mu_k(A)$$

où $p_1, \ldots, p_n$ sont des réels positifs tels que $p_1 + \cdots + p_n = 1$ et les lois $\mu_1, \ldots, \mu_n$ sont gaussiennes, *i.e.* pour tout $k \in \{1, \ldots, n\}$, la loi μ_k, notée $\mathcal{N}(m_k, \sigma_k^2)$, admet pour densité par rapport à la mesure de Lebesgue

$$\forall x \in \mathbb{R},\ f_k(x) = (2\pi\sigma_k^2)^{-1/2} \exp\left(-\frac{(x-m_k)^2}{2\sigma_k^2}\right)$$

avec $m_k \in \mathbb{R}$ et $\sigma_k^2 > 0$. L'algorithme suivant permet de simuler la loi μ et la figure 1.4 illustre un échantillon pour un mélange de deux gaussiennes.

ALGORITHME 1.13 – Mélange gaussien

Initialisation :
- le vecteur des probabilités $p = (p_1, \ldots, p_n)'$
- les réels $m_1, \ldots, m_n \in \mathbb{R}$ et $\sigma_1^2, \ldots, \sigma_n^2 > 0$

À l'étape i :
- Tirer $k \in \{1, \ldots, n\}$ selon les probabilités p
- Tirer X_i selon une loi $\mathcal{N}(m_k, \sigma_k^2)$

Retourner les valeurs $X_1, X_2, \ldots$

■

Un *modèle de mélange continu* généralise le modèle de mélange discret au cas où la variable mélangeante Y est à valeurs dans $\mathbb{R}$. En notant f_Y la densité de la loi de Y par rapport à la mesure de Lebesgue, si $\{\mu_y,\ y \in \mathbb{R}\}$ est une collection de lois de probabilités, alors nous pouvons considérer une variable réelle X dont la loi est donnée par,

$$\mathbb{P}(X \in A) = \int_{\mathbb{R}} f_Y(y) \mu_y(A) dy$$

pour tout ensemble mesurable A. Comme dans le cas discret, pour résoudre le problème de la simulation pour la loi de X, il faut simuler une réalisation y de la loi de Y et générer une réalisation de X avec la loi μ_y.

EXEMPLE 1.14 (Loi de Student). La loi de Student de paramètre $\nu > 0$, notée $\mathcal{T}(\nu)$, apparaît naturellement dans divers problèmes statistiques et correspond à une variable aléatoire T définie par

$$T = \frac{Z}{\sqrt{X/\nu}}$$

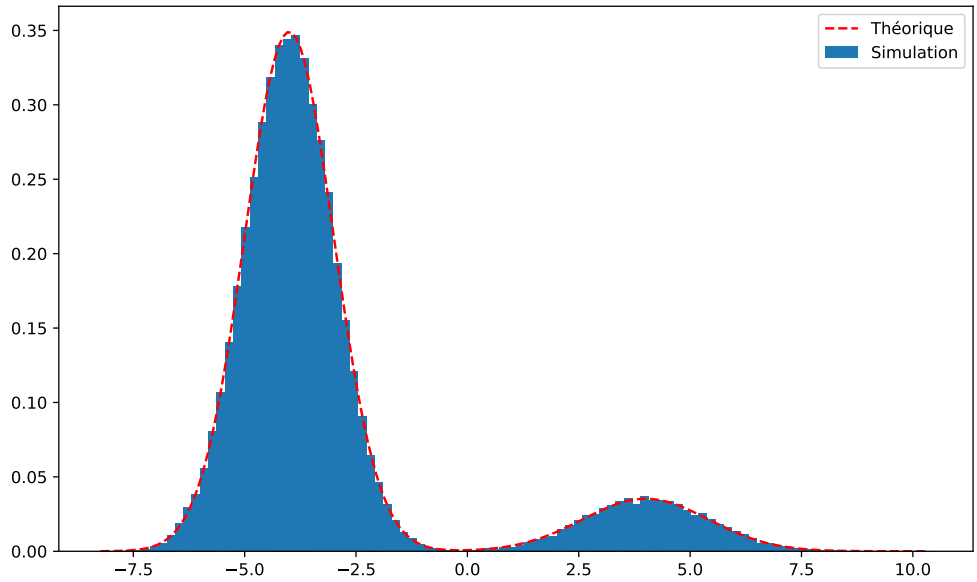

FIGURE 1.4 – Simulation de la loi mélange de deux gaussiennes.

où $Z \sim \mathcal{N}(0,1)$ est une variable indépendante de $X \sim \chi^2(\nu)$. Pour simuler des réalisations de T, nous pouvons voir la loi de Student comme un mélange gaussien continu,

$$\forall x > 0,\ T \mid X = x \sim \mathcal{N}(0, \nu/x) \quad \text{avec} \quad X \sim \chi^2(\nu).$$

Ainsi, pour simuler une réalisation t de la loi $\mathcal{T}(\nu)$, nous générons une réalisation x suivant la loi $\chi^2(\nu)$ et nous tirons ensuite t selon une loi gaussienne centrée de variance ν/x. La figure 1.5 donne un échantillon de loi de Student de paramètre 5 généré selon cette méthode. ■

Le principe utilisé pour résoudre le problème de la simulation pour les mélanges de lois se généralise sans difficulté. Cela s'appelle la *méthode du conditionnement* et prend la forme du théorème suivant.

Théorème 1.15 (Méthode du conditionnement)**.** *Soit (X,Y) un couple de variables aléatoires à valeurs dans $\mathcal{X} \times \mathcal{Y}$. Si le problème de la simulation est résolu pour la loi μ_Y de Y ainsi que pour toutes les lois conditionnelles $\mu_X(\cdot \mid Y = y)$ de X où $y \in \mathcal{Y}$, alors il est résolu pour la loi μ_X de X et pour la loi jointe $\mu_{X,Y}$ du couple (X,Y).*

Démonstration. Ce résultat est évident une fois établi que la loi de X est le mélange des lois $\mu_X(\cdot \mid Y = y)$ par la loi μ_Y car, pour tout ensemble mesurable A,

$$\mu_X(A) = \mathbb{P}(X \in A) = \mathbb{E}\left[\mathbb{P}(X \in A \mid Y)\right] = \int_{\mathcal{Y}} \mu_X(A \mid Y = y) d\mu_Y(y).$$

Pour la loi jointe, si y est une réalisation de la loi μ_Y et si x est une réalisation de la loi conditionnelle $\mu_X(\cdot \mid Y = y)$, alors (x,y) est une réalisation de la loi $\mu_{X,Y}$. □

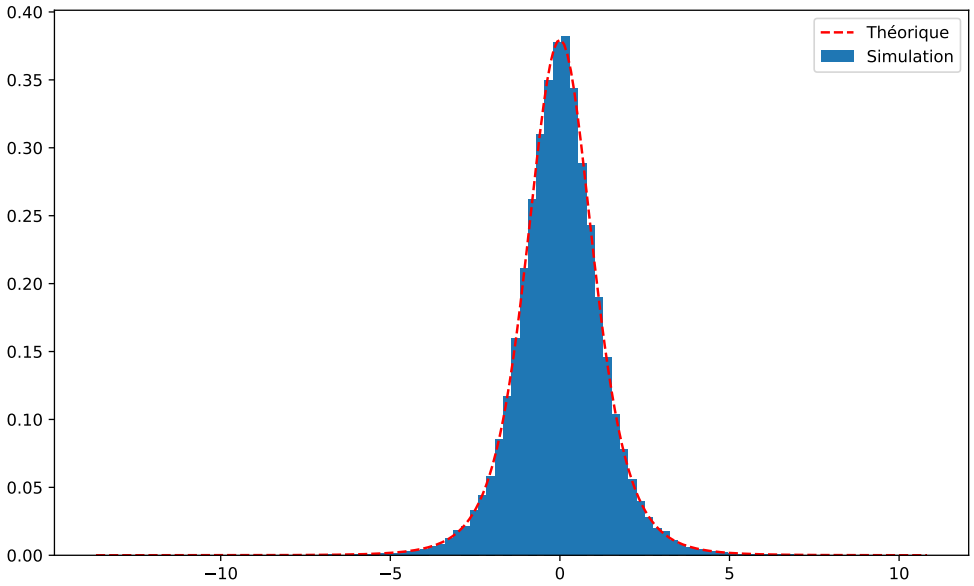

FIGURE 1.5 – Simulation de la loi $\mathscr{T}(5)$ par un mélange gaussien continu.

1.2.5 Méthode du rejet

Un des inconvénients de la méthode d'inversion évoquée précédemment était la nécessité de connaître explicitement (ou du moins de pouvoir évaluer) la fonction de répartition F de la loi μ à simuler. Or, comme nous le verrons à plusieurs reprises dans la suite de ce cours, cette loi μ n'est souvent connue qu'à une constante multiplicative près qui est bien trop coûteuse à évaluer, ce qui rend la méthode d'inversion inutilisable dans ces cas. Dans les années 1950, le mathématicien et physicien américano-hongrois John von Neumann introduit la *méthode du rejet* basée sur la relation entre une probabilité et la surface sous une courbe pour contourner cette difficulté.

La *méthode du rejet* est un premier exemple de *méthodes de simulation indirectes* au sens où, pour générer une réalisation d'une loi μ, elle génère une réalisation d'une *loi candidate* ν, plus simple à simuler que μ, qui devra passer un test pour être acceptée comme une réalisation de μ. Ce type d'approche est très souple et peut être utilisé pour simuler un grand nombre de lois de probabilité. Les algorithmes stochastiques étudiés dans le chapitre 2 sont essentiellement basés sur le même principe.

Le théorème suivant donne une première version simple de la méthode du rejet pour résoudre le problème de la simulation de la loi uniforme sur un ensemble mesurable $B \subset \mathbb{R}^2$ de mesure de Lebesgue finie.

FIGURE 1.6 – John von Neumann (1903-1957)

Théorème 1.16 (Méthode du rejet)**.** *Soient $B \subset A$ deux sous-ensembles de $\mathbb{R}^2$ de mesures de Lebesgue respectives $\lambda(A)$ et $\lambda(B)$ telles que $0 < \lambda(B) \leqslant \lambda(A) < +\infty$. Considérons une suite $(X_n)_{n\geqslant 1}$ de variables aléatoires indépendantes de loi uniforme sur A et posons*

$$T = \inf\{n \geqslant 1 \text{ tel que } X_n \in B\}.$$

Alors,

- *T suit la loi géométrique $\mathcal{G}(p)$ avec $p = \lambda(B)/\lambda(A)$,*
- *T et X_T sont indépendantes,*
- *X_T suit la loi uniforme sur B.*

Démonstration. Par définition de la loi uniforme sur A, nous savons que

$$\forall n \geqslant 1,\ \mathbb{P}(X_n \in B) = p.$$

La loi de T se déduit par indépendance,

$$\begin{aligned}\forall k > 0,\ \mathbb{P}(T = k) &= \mathbb{P}(X_1 \notin B, \dots, X_{k-1} \notin B, X_k \in B)\\ &= \mathbb{P}(X_1 \notin B) \times \cdots \times \mathbb{P}(X_{k-1} \notin B) \times \mathbb{P}(X_k \in B)\\ &= (1-p)^{k-1} p.\end{aligned}$$

Soit $C \subset B$ un sous-ensemble mesurable, les mêmes arguments donnent

$$\forall k > 0,\ \mathbb{P}(X_T \in C \text{ et } T = k) = (1-p)^{k-1} \frac{\lambda(C)}{\lambda(A)}.$$

L'indépendance entre T et X_T découle de la forme produit de ces probabilités. Enfin, la loi uniforme sur B de la variable X_T s'obtient en sommant les probabilités,

$$\mathbb{P}(X_T \in C) = \sum_{k>0} \mathbb{P}(X_T \in C \text{ et } T = k) = \sum_{k>0} (1-p)^{k-1} \frac{\lambda(C)}{\lambda(A)} = \frac{\lambda(C)}{\lambda(B)}.$$

□

Il est intéressant de remarquer que la méthode décrite dans le théorème 1.16 ne nécessite pas de connaître explicitement $\lambda(B)$, *i.e.* la loi uniforme sur B à simuler n'est connue que à une constante multiplicative $1/\lambda(B)$ près. Cependant, il est toujours nécessaire de pouvoir tester si un point appartient à B. La loi candidate est ici la loi uniforme sur A qui peut être facile à simuler si A est un simple pavé de $\mathbb{R}^2$, par exemple. Le nombre moyen de rejets pour obtenir une réalisation de la loi uniforme sur B vaut $p^{-1} = \lambda(B)/\lambda(A) \geqslant 1$. Comme chaque valeur candidate rejetée représente du temps de calcul gâché, il est crucial de minimiser la surface de $A \setminus B$ pour assurer de bonnes performances. Notons également que le rapport entre le nombre de valeurs proposées et le nombre de valeurs acceptées par cette méthode donne une estimation de la surface de B (il s'agit d'une méthode de Monte Carlo comme nous le verrons dans la section 1.3). L'algorithme ci-dessous implante cette première version de la méthode du rejet tandis que la figure 1.7 montre une simulation de la loi uniforme sur le quart de cercle et donne une estimation de π grâce à elle.

ALGORITHME 1.17 – Méthode du rejet

Initialisation : $B \subset A$ deux sous-ensembles de $\mathbb{R}^2$ de surfaces finies
À l'étape n :
  Répéter
    Tirer X_n selon la loi uniforme sur A
  Tant que $X_n \notin B$
Retourner les valeurs $X_1, X_2, \ldots$

À partir de la méthode du rejet, il est relativement simple de mettre en place une procédure de simulation pour une loi continue sur $\mathbb{R}$ en se basant sur le lemme suivant.

Lemme 1.18. *Soit μ une loi sur $\mathbb{R}$ de densité f_μ par rapport à la mesure de Lebesgue. Si le vecteur aléatoire (X,Y) suit une loi uniforme sur*

$$B_\mu = \left\{(x,y) \in \mathbb{R}^2 \text{ tel que } 0 < y < f_\mu(x)\right\}$$

alors l'abscisse X suit la loi μ.

Démonstration. Soit un intervalle $I \subset \mathbb{R}$, nous définissons l'ensemble

$$C = \left\{(x,y) \in B_\mu \text{ tel que } x \in I\right\}.$$

Comme f_μ est une densité de probabilité, la mesure de Lebesgue de B_μ vaut $\lambda(B_\mu) = 1$ et celle de C vaut $\lambda(C) = \mu(I)$. Par définition de la loi uniforme sur B_μ, nous avons donc

$$\mathbb{P}((X,Y) \in C) = \frac{\lambda(C)}{\lambda(B_\mu)} = \mu(I). \tag{1.1}$$

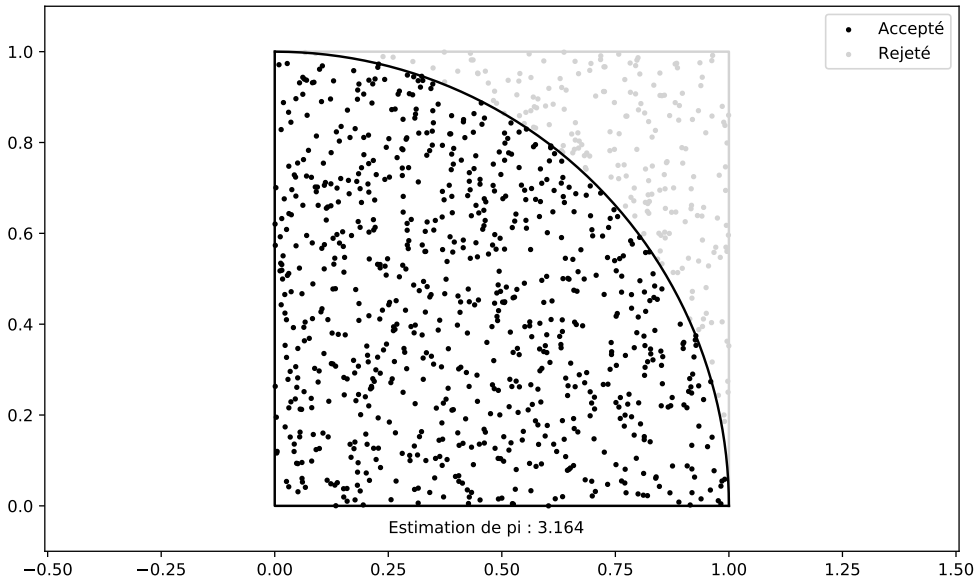

FIGURE 1.7 – Simulation de la loi uniforme sur le quart de cercle par la méthode du rejet avec une loi candidate uniforme sur $[0,1]^2$ (les points noirs sont uniformément distribués sur le quart de cercle, les points gris sont rejetés) et estimation de π par 4 fois la proportion de points acceptés.

D'autre part, nous savons que $Y < f_\mu(X)$ presque sûrement par définition, donc

$$\mathbb{P}((X,Y) \in C) = \mathbb{P}\left(X \in I \text{ et } Y < f_\mu(X)\right) = \mathbb{P}(X \in I).$$

Il s'ensuit $\mathbb{P}(X \in I) = \mu(I)$, X suit donc la loi μ. □

En pratique, ce lemme et la méthode du rejet permettent de réduire le problème de la simulation pour la loi μ à celui de la loi uniforme sur un sous-ensemble $A \subset \mathbb{R}^2$ qui contient B_μ. Le cas le plus simple est celui d'une densité f_μ bornée par $M > 0$ et de support inclus dans un intervalle I borné. Il suffit alors de considérer le pavé $A = I \times [0,M]$. Le choix de la borne M peut s'avérer difficile dans des cas concrets et parfois nécessiter une évaluation coûteuse. Il est bien sûr toujours possible de prendre une borne plus grande que nécessaire mais cela aura pour conséquence d'augmenter le nombre de candidats rejetés et donc de gâcher plus de temps de calcul. Par contre, la borne M ne doit pas être sous-estimée sous peine de tronquer la densité. Il convient également de remarquer qu'il suffit de connaître la densité f_μ à une constante multiplicative près pour obtenir le même résultat. En effet, si $\tilde{f}_\mu$ est une fonction telle qu'il existe $K > 0$ et $f_\mu(x) = K\tilde{f}_\mu(x)$, pour tout $x \in \mathbb{R}$, alors l'ensemble B_μ peut être remplacé par

$$\tilde{B}_\mu = \left\{(x,y) \in \mathbb{R}^2 \text{ tel que } 0 < y < \tilde{f}_\mu(x)\right\}.$$

Le lemme reste vrai et l'abcisse X suit encore la loi μ car cette constante K se simplifie dans le rapport des mesures de Lebesgue (1.1). L'algorithme suivant détaille les étapes de cette méthode pour sa mise en pratique.

ALGORITHME 1.19 – Méthode du rejet pour une densité bornée à support borné

Initialisation :
- une fonction $\tilde{f}_\mu$ proportionnelle à la densité de la loi à simuler
- un pavé $A = [a,b] \times [0,M]$ contenant l'ensemble $\tilde{B}_\mu$

À l'étape n :

Répéter

Tirer (X_n, Y_n) selon la loi uniforme sur A

Tant que $\tilde{f}_\mu(X_n) \leqslant Y_n$

Retourner les valeurs $X_1, X_2, \ldots$

EXEMPLE 1.20 (Loi bêta 1). Soit $\alpha, \beta > 0$, la loi bêta $\mathscr{B}e(\alpha,\beta)$ est la loi sur $[0,1]$ dont la densité par rapport à la mesure de Lebesgue est donnée par

$$\forall x \in [0,1],\ f_{\alpha,\beta}(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

où $\Gamma(z) = \displaystyle\int_0^{+\infty} t^{z-1}e^{-t}dt$, $z > 0$. Nous considérons la fonction proportionnelle

$$\forall x \in [0,1],\ \tilde{f}_{\alpha,\beta}(x) = x^{\alpha-1}(1-x)^{\beta-1}.$$

Pour $\alpha = 2.7$ et $\beta = 6.3$, une simple étude de la fonction $\ln(\tilde{f}_{\alpha,\beta})$ montre que $M = 0.021$ suffit et la figure 1.8 donne le résultat d'une simulation par la méthode du rejet avec une loi candidate uniforme sur $A = [0,1] \times [0,M]$. ■

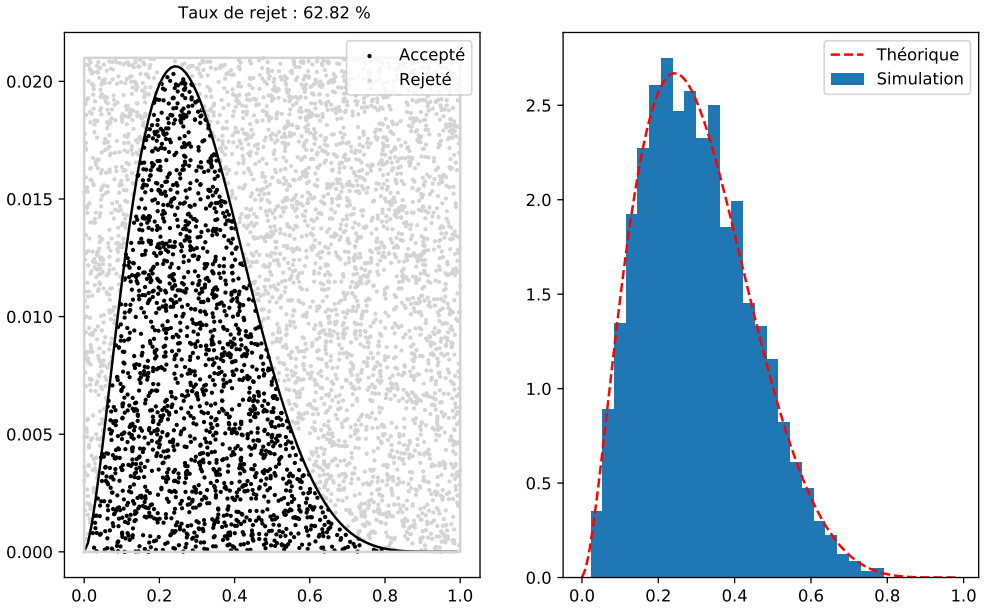

FIGURE 1.8 – Simulation de la loi $\mathscr{B}e(2.7, 6.3)$ par la méthode du rejet (les points acceptés et rejetés à gauche et la distribution de l'échantillon généré à droite).

Un taux de rejet important comme celui qui apparaît sur la figure 1.8 a un impact négatif sur les performances de cet algorithme. Idéalement, la surface de A doit être au plus proche de celle de $\tilde{B}_\mu$ pour éviter ce problème mais cette condition est difficile à satisfaire en toute généralité pour des lois candidates uniformes sur des pavés. Le lemme suivant propose une sorte de réciproque au lemme 1.18 qui nous permettra de contourner cette difficulté.

Lemme 1.21. *Soit ν une loi sur $\mathbb{R}$ de densité f_ν par rapport à la mesure de Lebesgue. Si X est une variable aléatoire de loi ν et $U \sim \mathscr{U}([0,1])$ est indépendante de X, alors le vecteur aléatoire $(X, f_\nu(X)U)$ suit la loi uniforme sur l'ensemble*

$$A_\nu = \left\{(x,y) \in \mathbb{R}^2 \text{ tel que } 0 < y < f_\nu(x)\right\}.$$

Démonstration. Soit un ensemble mesurable $C \subset A_\nu$ de mesure de Lebesgue $\lambda(C)$ dont nous considérons les tranches,

$$\forall x \in \mathbb{R},\ C_x = \{y \in \mathbb{R}_+ \text{ tel que } (x,y) \in C\}.$$

Notons que si $f_\nu(x) = 0$, alors C_x est l'ensemble vide et sa mesure de Lebesgue est nulle. Par indépendance entre U et X, nous obtenons

$$\begin{aligned}
\mathbb{P}((X, f_\nu(X)U) \in C) &= \int_{\mathbb{R}} \mathbb{P}((X, f_\nu(X)U) \in C \mid X = x)\, f_\nu(x) dx \\
&= \int_{\mathbb{R}} \mathbb{P}(f_\nu(x)U \in C_x)\, f_\nu(x) \mathbf{1}_{f_\nu(x)>0}\, dx \\
&= \int_{\mathbb{R}} \frac{\lambda(C_x)}{f_\nu(x)} f_\nu(x) \mathbf{1}_{f_\nu(x)>0}\, dx \\
&= \lambda(C).
\end{aligned}$$

Nous en déduisons bien que $(X, f_\nu(X)U)$ suit la loi uniforme sur A_ν puisque f_ν est une densité de probabilité (*i.e.* $\lambda(A_\nu) = 1$). □

Ce lemme nous permet en particulier d'appliquer la méthode du rejet pour simuler une loi continue μ sur $\mathbb{R}$ à partir de n'importe quelle loi candidate ν continue sur $\mathbb{R}$ telle que $B_\mu \subset A_\nu$. Cette méthode porte le nom de *méthode du rejet comparatif* et se formalise par le théorème ci-dessous.

Théorème 1.22 (Méthode du rejet comparatif)**.** *Soient μ et ν deux lois sur $\mathbb{R}$ de densités respectives f_μ et f_ν par rapport à la mesure de Lebesgue telles qu'il existe une constante $M \geqslant 1$ pour laquelle*

$$\forall x \in \mathbb{R},\ f_\mu(x) \leqslant M f_\nu(x).$$

Considérons une suite $(U_n)_{n\geqslant 1}$ de variables aléatoires indépendantes de loi $\mathscr{U}([0,1])$ et une suite $(X_n)_{n\geqslant 1}$ de variables aléatoires indépendantes de loi ν. Nous définissons

$$T = \inf\{n \geqslant 1 \text{ tel que } M f_\nu(X_n) U_n < f_\mu(X_n)\}.$$

Si les suites $(U_n)_{n\geqslant 1}$ et $(X_n)_{n\geqslant 1}$ sont indépendantes, alors,

- *T suit la loi géométrique $\mathscr{G}(1/M)$,*
- *T et X_T sont indépendantes,*
- *X_T suit la loi μ.*

Démonstration. Le lemme 1.21 implique que les vecteurs aléatoires $(X_n, f_\nu(X_n)U_n)$ suivent la loi uniforme sur l'ensemble A_ν. De façon élémentaire, nous pouvons en déduire que les vecteurs aléatoires $(X_n, M f_\nu(X_n)U_n)$ suivent la loi uniforme sur

$$A = \left\{(x,y) \in \mathbb{R}^2 \text{ tel que } 0 < y < M f_\nu(x)\right\}.$$

Par hypothèse, $B_\mu \subset A$ et les conclusions se déduisent du lemme 1.18 et du théorème 1.16. □

La méthode du rejet comparatif ramène le problème de la simulation pour la loi μ à celui pour la loi ν. Bien entendu, pour que cela ait un intérêt, il faut que la loi ν soit plus simple à simuler que la loi μ. Si la loi candidate ν est bien choisie, cette méthode permet de réduire le taux de candidats rejetés et donc d'améliorer les performances en pratique par rapport à une loi candidate uniforme sur un pavé. Plus la constante M sera proche de 1, meilleures seront ces performances, le cas extrême $M = 1$ correspondant à $\nu = \mu$. Comme précédemment, nous pouvons remarquer que la densité f_μ peut être seulement connue à une constante multiplicative près, *i.e.* lorsque nous ne pouvons évaluer qu'une fonction $\tilde{f}_\mu$ telle qu'il existe une constante $K > 0$ avec $f_\mu(x) = K\tilde{f}_\mu(x)$, pour tout $x \in \mathbb{R}$. Dans ce cas, c'est la constante de la méthode qui absorbe ce manque de connaissance quitte à être sur-estimée, augmentant de fait le taux de rejet. En pratique, la recherche d'une bonne valeur pour cette constante peut être coûteuse mais il ne faut pas la sous-estimer sous peine de tronquer la loi à simuler. L'algorithme suivant implante la méthode du rejet comparatif et l'exemple 1.24 illustre ce que nous pouvons gagner par rapport à l'exemple 1.20.

ALGORITHME 1.23 – Méthode du rejet comparatif

Initialisation :
- une fonction $\tilde{f}_\mu$ proportionnelle à la densité de la loi à simuler
- une densité candidate f_ν et une constante $\tilde{M} > 0$ telles que $\tilde{f}_\mu \leqslant \tilde{M} f_\nu$

À l'étape n :

Répéter
  Tirer U_n selon la loi $\mathscr{U}([0,1])$
  Tirer X_n selon la loi ν

Tant que $\tilde{f}_\mu(X_n) \leqslant \tilde{M} f_\nu(X_n) U_n$
Retourner les valeurs $X_1, X_2, \ldots$

EXEMPLE 1.24 (Loi bêta 2). Reprenons le problème de la simulation pour une loi bêta $\mathscr{B}e(2.7, 6.3)$ dont la densité est proportionnelle à

$$\forall x \in [0,1],\ \tilde{f}(x) = x^{1.7}(1-x)^{5.3}.$$

Une simple étude de fonction suffit encore à établir que pour $\tilde{M} = 0.0141$, nous avons

$$\forall x \in [0,1],\ \tilde{f}(x) \leqslant \tilde{M} f_{2,6}(x)$$

où $f_{2,6}$ est la densité d'une loi $\mathscr{B}e(2,6)$. Les lois bêta de paramètres entiers sont très simples à simuler à partir de réalisations indépendantes d'une loi exponentielle telles celles que nous avons produites dans l'exemple 1.6. En effet, si $E_1, \ldots, E_8$ sont des variables aléatoires indépendantes de loi $\mathscr{E}(1)$, alors il est possible de montrer que

$$\frac{E_1 + E_2}{E_1 + \cdots + E_8} \sim \mathscr{B}e(2,6).$$

Nous pouvons donc employer la méthode du rejet comparatif pour simuler la loi $\mathscr{B}e(2.7, 6.3)$ et la figure 1.9 illustre les résultats. Nous remarquons en particulier le taux de rejet meilleur que celui obtenu dans l'exemple 1.20. ■

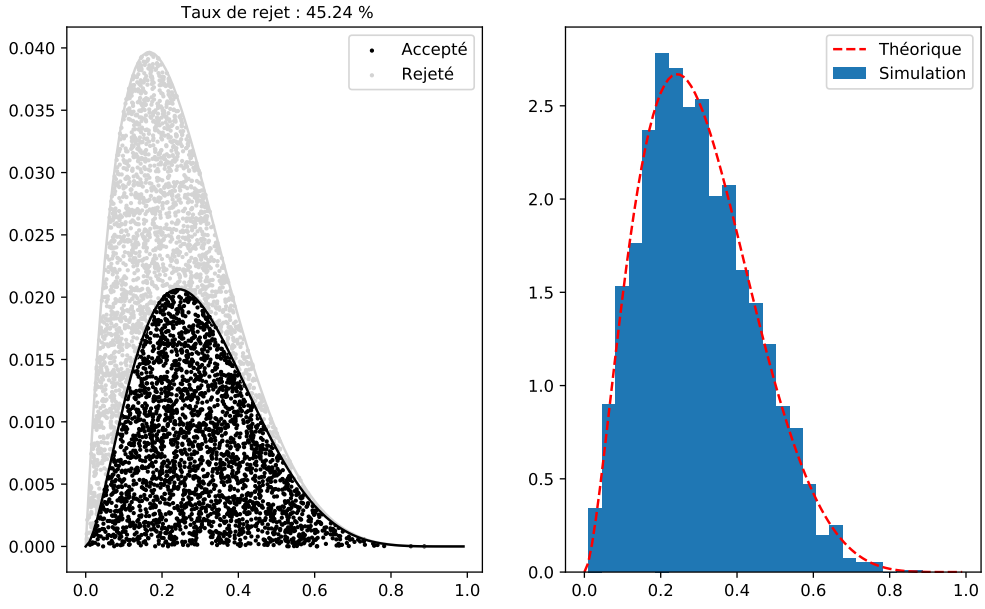

FIGURE 1.9 – Simulation de la loi $\mathscr{B}e(2.7, 6.3)$ par la méthode du rejet comparatif à partir d'une loi candidate $\mathscr{B}e(2,6)$ (les points acceptés et rejetés à gauche et la distribution de l'échantillon généré à droite).

La méthode du rejet est largement utilisée en pratique. Elle présente l'avantage de produire une simulation exacte de la loi d'intérêt et elle se généralise facilement à des lois continues sur

$\mathbb{R}^d$. Le fait de pouvoir utiliser cette méthode pour des lois dont la densité n'est connue qu'à une constante multiplicative près est un atout particulièrement utiles dans différents domaines. Par exemple, en statistique bayésienne, la constante de normalisation de la loi a posteriori n'est pas toujours explicite et cette loi pourra tout de même être simulée par la méthode du rejet. L'inconvénient de cette approche réside dans la difficulté à bien calibrer la constante de la procédure. Si celle-ci est sous-estimée, la loi simulée se retrouve tronquée et la loi des réalisations ne correspond pas à ce qui est attendu. D'autre part, si la constante est sur-estimée, le taux de rejet sera important et du temps de calcul sera gâché en vain. Nous retrouverons le même principe de loi candidate et de réalisations à valider dans les méthodes étudiées au chapitre 2.

1.3 Estimation d'intégrales

Lorsque une variable aléatoire $X \in \mathscr{X}$ de loi μ_X intervient dans un problème de modélisation ou de statistique, nous sommes souvent intéressés par des quantités qui s'expriment comme des intégrales de fonctions $h : \mathscr{X} \to \mathbb{R}$,

$$\mathbb{E}[h(X)] = \int_{\mathscr{X}} h(x) d\mu_X(x).$$

Il n'est pas rare non plus que ces quantités d'intérêt dépendent d'autres paramètres $\theta \in \Theta$ du problème considéré, conduisant de fait à des fonctions $H : \Theta \to \mathbb{R}$ définies comme des intégrales de $h : \Theta \times \mathscr{X} \to \mathbb{R}$,

$$\forall \theta \in \Theta,\ H(\theta) = \mathbb{E}[h(\theta, X)] = \int_{\mathscr{X}} h(\theta, x) d\mu_X(x).$$

Les exemples de tels problèmes ne manquent pas (constante de normalisation d'une densité, moments d'une variable aléatoire, estimateur bayésien pour la fonction de perte quadratique, calculs de quantiles, ...) et il est naturel de vouloir approcher ces quantités à l'aide de réalisations de X obtenues par l'expérience ou par simulation. Ces questions se trouveront au cœur des méthodes étudiées dans le chapitre 2 et nous présentons ici deux approches classiques pour y répondre dans des cadres simples mais souvent utilisés en pratique.

1.3.1 Méthode de Monte Carlo

Vers la fin des années 1940, le mathématicien polono-américain Stanislaw Ulam travaillait dans la même équipe que John von Neumann sur la conception de l'arme nucléaire à Los Alamos. C'est dans le cadre de ces travaux que Ulam formalisa une méthode déjà connue pour approcher la valeur d'une intégrale. Étant donnée la nature de ces travaux, Ulam et von Neumann avaient besoin d'un nom de code pour parler de cette méthode et Nicholas Metropolis, un autre collègue de Los Alamos dont nous reparlerons dans le chapitre 2, leur suggéra le nom de *Monte Carlo* en référence au casino de Monaco.

Le principe de la *méthode de Monte Carlo* est relativement simple. Étant données une loi μ sur $\mathbb{R}^d$ de densité de probabilité f par rapport à la mesure de Lebesgue et une fonction $h : \mathbb{R}^d \to \mathbb{R}$, l'objectif est de calculer une valeur numérique approchée de l'intégrale

$$J = \int_{\mathbb{R}^d} h(x) d\mu(x) = \int_{\mathbb{R}^d} h(x) f(x) dx. \tag{1.2}$$

FIGURE 1.10 – Stanislaw Ulam (1909-1984)

Pour cela, la méthode consiste à considérer une suite $(X_n)_{n\in\mathbb{N}}$ de variables aléatoires indépendantes de loi μ et à définir

$$\forall n \in \mathbb{N},\ J_n = \frac{1}{n}\sum_{k=0}^{n} h(X_k).$$

Cette méthode est basée sur la loi des grands nombres qui assure que J_n converge presque sûrement vers J lorsque le nombre de réalisations prises en compte est grand,

$$J_n \xrightarrow[n\to+\infty]{p.s.} J.$$

De plus, si la fonction h est de carré intégrable,

$$\int_{\mathbb{R}^d} h(x)^2 d\mu(x) = \int_{\mathbb{R}^d} h(x)^2 f(x)dx < +\infty,$$

alors la variance $\sigma^2 = \mathbb{V}(h(X_0))$ est bien définie et l'inégalité de Bienaymé-Tchebychev implique

$$\forall \varepsilon > 0,\ \mathbb{P}\left(|J_n - J| \geqslant \sqrt{\frac{\sigma^2}{n\varepsilon}}\right) \leqslant \varepsilon.$$

Autrement dit, si nous connaissons une borne supérieure pour la variance (ou si cette variance peut être estimée), alors nous disposons d'un intervalle de confiance non asymptotique pour mesurer la qualité de l'approximation de J par J_n. Par le théorème central limite, nous savons

également que $\sqrt{n}(J_n - J)$ converge en loi vers la loi normale $\mathcal{N}(0,\sigma^2)$,

$$\sqrt{n}(J_n - J) \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}(0,\sigma^2),$$

et nous connaissons donc aussi le comportement asymptotique de J_n.

ALGORITHME 1.25 – Méthode de Monte Carlo
Initialisation : une loi μ sur $\mathbb{R}^d$ et une fonction intégrable $h : \mathbb{R}^d \to \mathbb{R}$
Pour k allant de 1 à n :
  Tirer X_k selon la loi μ
  Calculer $h(X_k)$
Retourner $(h(X_1) + \cdots + h(X_n))/n$

Bien entendu, lorsque cela est possible, les méthodes de simulation évoquées dans la section précédente peuvent être utilisées pour obtenir des réalisations de la loi μ. Cette méthode se généralise à des suites de variables qui ne sont plus indépendantes et identiquement distribuées et est à la base des *méthodes de Monte Carlo par chaînes de Markov* qui feront l'objet du chapitre 2.

EXEMPLE 1.26 (Valeur approchée d'une intégrale). Soit la fonction h définie sur $[0,1]$ par

$$\forall x \in [0,1],\ h(x) = (\cos(50x) + \sin(20x))^2.$$

Pour obtenir une valeur approchée de l'intégrale $J = \displaystyle\int_0^1 h(x)dx$, nous considérons une suite $(U_n)_{n\in\mathbb{N}}$ de réalisations indépendantes de la loi uniforme sur $[0,1]$ et nous définissons

$$\forall n \in \mathbb{N},\ J_n = \frac{1}{n}\sum_{k=1}^{n} h(U_k).$$

Une majoration brutale de l'intégrale du carré de h nous donne la borne supérieure $\mathbb{V}(h(U_0)) \leqslant 16$ et, par l'inégalité de Bienaymé-Tchebychev, nous déduisons qu'avec une probabilité au moins égale à 90%,

$$|J_n - J| \leqslant \frac{12.65}{\sqrt{n}}.$$

Cet encadrement n'est pas très fin mais il nous assure que le premier chiffre après la virgule est exact avec une probabilité de 0.9 dès lors que le nombre de réalisations dépassent 27723. Le théorème central limite peut aussi être invoqué pour déduire un intervalle de confiance sur la valeur de J. La figure 1.11 montre l'évolution de J_n en fonction de n. ■

EXEMPLE 1.27 (Fonction de répartition de la loi normale). La fonction de répartition de la loi normale standard est donnée par

$$\forall x \in \mathbb{R},\ \varphi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.$$

Cette fonction a souvent besoin d'être évaluée en statistique, par exemple pour obtenir les quantiles de la loi $\mathcal{N}(0,1)$ apparaissant dans les bornes d'un intervalle de confiance asymptotique. Cependant, la fonction φ n'admet pas de forme explicite et il est nécessaire d'en calculer une valeur approchée en pratique. En considérant une suite $(Z_n)_{n\in\mathbb{N}}$ de variables aléatoires

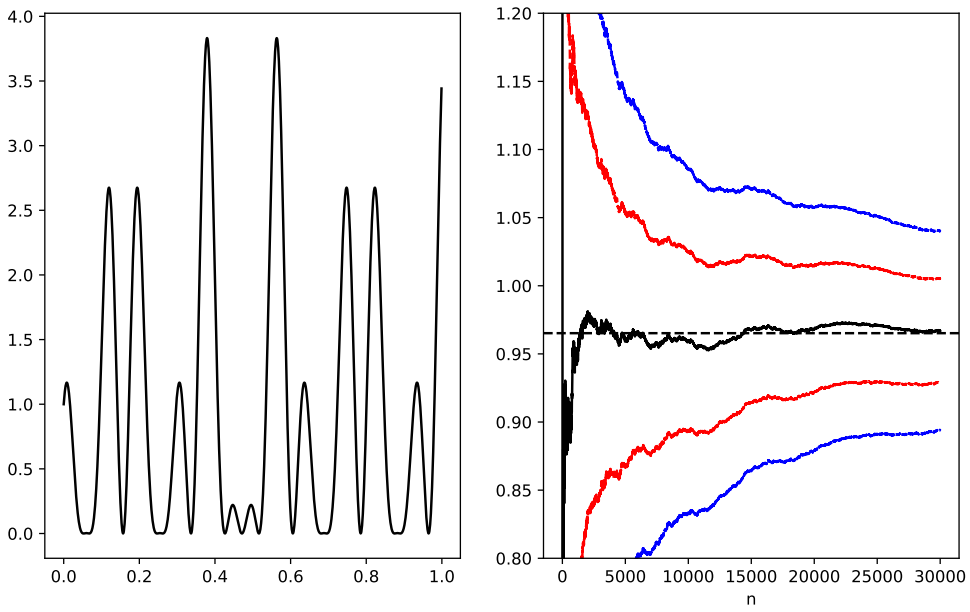

FIGURE 1.11 – Fonction $h(x) = (\cos(50x) + \sin(20x))^2$ (à gauche) et la valeur de son intégrale sur $[0,1]$ approchée par la méthode de Monte Carlo pour différentes valeurs de n avec les bornes de confiance à 90% (à droite, Bienaymé-Tchebychev en bleu et TCL en rouge).

indépendantes de loi $\mathcal{N}(0,1)$, la méthode de Monte Carlo nous amène à considérer la fonction de répartition empirique,

$$\forall n \in \mathbb{N},\ \forall x \in \mathbb{R},\ \varphi_n(x) = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{Z_k \leqslant x}.$$

Ponctuellement, la loi des grands nombres nous assure que $\varphi_n(x)$ converge presque sûrement vers $\varphi(x)$, pour tout $x \in \mathbb{R}$. Le théorème de Glivenko-Cantelli donne le même résultat uniformément sur $\mathbb{R}$,

$$\sup_{x\in\mathbb{R}} |\varphi_n(x) - \varphi(x)| \xrightarrow[n\to+\infty]{p.s.} 0.$$

La figure 1.12 montre l'estimation de la fonction φ par φ_n pour quelques valeurs de n. ■

1.3.2 Échantillonnage préférentiel

La méthode de Monte Carlo présentée ci-dessus s'applique dans un cadre beaucoup plus général et il n'est pas toujours nécessaire de pouvoir générer des réalisations de la loi μ contre laquelle la fonction h est intégrée. En effet, avec les mêmes notations que (1.2), si g est une autre densité de probabilité sur $\mathbb{R}^d$ telle que

$$\forall x \in \mathbb{R}^d,\ f(x) > 0 \Rightarrow g(x) > 0, \tag{1.3}$$

alors l'intégrale à calculer s'écrit également

$$J = \int_{\mathbb{R}^p} \frac{h(x)f(x)}{g(x)} g(x)dx.$$

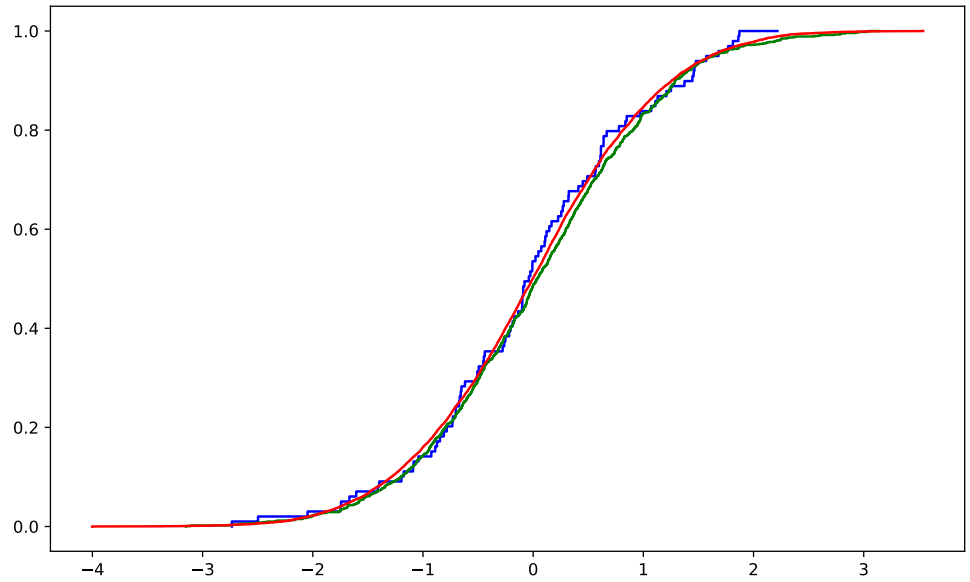

FIGURE 1.12 – Estimation de la fonction de répartition de la loi $\mathcal{N}(0,1)$ par la méthode de Monte Carlo pour $n = 100$ (en bleu), $n = 1000$ (en vert) et $n = 10000$ (en rouge).

Dans le cas où la loi ν de densité g est plus simple à simuler que la loi μ de densité f, il est possible de tirer parti de cette écriture pour obtenir une valeur approchée de J en suivant un raisonnement similaire à ce que nous avons vu pour la méthode de Monte Carlo. En effet, en considérant une suite $(X'_n)_{n\in\mathbb{N}}$ de variables aléatoires indépendantes de loi ν, nous pouvons définir les moyennes empiriques

$$\forall n \in \mathbb{N},\ J'_n = \frac{1}{n}\sum_{k=1}^{n} h(X'_k)\omega(X'_k)$$

où les poids sont donnés par $\omega(x) = f(x)/g(x)$. La loi des grands nombres assure encore que J'_n converge presque sûrement vers l'intégrale J à évaluer. Cette méthode est l'*échantillonnage préférentiel* (ou *importance sampling* en anglais) et la densité auxiliaire g est généralement appelée la *fonction d'importance*.

Un atout important de l'échantillonnage préférentiel apparaît lorsque la constante de normalisation de la densité initiale f est trop coûteuse à évaluer, *i.e.* lorsque nous ne disposons que d'une fonction $\tilde{f}$ telle qu'il existe une constante $K > 0$ avec $f(x) = K^{-1}\tilde{f}(x)$. Dans ce cas, l'intégrale J s'écrit

$$J = K^{-1}\int_{\mathbb{R}^d} \frac{h(x)\tilde{f}(x)}{g(x)} g(x)dx \quad \text{où} \quad K = \int_{\mathbb{R}^d} \tilde{f}(x)dx = \int_{\mathbb{R}^d} \frac{\tilde{f}(x)}{g(x)} g(x)dx$$

et nous pouvons poser

$$\forall n \in \mathbb{N},\ \tilde{J}_n = \frac{1}{n}\sum_{k=1}^{n} h(X'_k)\tilde{\omega}(X'_k) \quad \text{et} \quad \tilde{K}_n = \frac{1}{n}\sum_{k=1}^{n} \tilde{\omega}(X'_k)$$

avec $\tilde{\omega}(x) = \tilde{f}(x)/g(x)$. Toujours grâce à la loi des grands nombres, nous savons

$$\tilde{J}_n \xrightarrow[n\to+\infty]{p.s.} \int_{\mathbb{R}^d} \frac{h(x)\tilde{f}(x)}{g(x)} g(x)dx \quad \text{et} \quad \tilde{K}_n \xrightarrow[n\to+\infty]{p.s.} K.$$

Par continuité de $(x,y) \mapsto x/y$ sur $\mathbb{R} \times \mathbb{R}_+^*$, nous en déduisons que les rapports entre $\tilde{J}_n$ et $\tilde{K}_n$ fournissent des valeurs approchées de l'intégrale cible sans avoir à évaluer explicitement la densité f,

$$\frac{\tilde{J}_n}{\tilde{K}_n} \xrightarrow[n\to+\infty]{p.s.} J.$$

Il n'est pas simple d'obtenir des intervalles de confiance non asymptotiques dans le cadre de l'échantillonnage préférentiel où la constante de normalisation de la densité n'est pas disponible à cause de l'aléa présent au numérateur et au dénominateur du rapport $\tilde{J}_n/\tilde{K}_n$. Cependant, du point de vue asymptotique, il est possible de montrer que

$$\forall 0 < \alpha < 1, \lim_{n\to+\infty} \mathbb{P}\left(\left| \frac{\tilde{J}_n}{\tilde{K}_n} - J \right| \geqslant q_\alpha \sqrt{\frac{\hat{\sigma}_n^2}{n}} \right) \leqslant \alpha$$

où q_α est le α-quantile d'une loi $\mathcal{N}(0,1)$ et l'estimateur de la variance est donné par

$$\hat{\sigma}_n^2 = \frac{1}{n^2 \tilde{K}_n^2} \sum_{k=1}^{n} \tilde{\omega}(X_k')^2 (h(X_k') - \tilde{J}_n/\tilde{K}_n)^2.$$

Ce résultat sort du cadre de cette introduction et il est laissé à titre d'exercice pour le lecteur intéressé. Il fait uniquement appel au théorème central limite et à la méthode delta.

Comme dans le cas de la méthode du rejet comparatif, nous voyons ici apparaître une loi auxiliaire ν que nous pouvons choisir librement tant que les densités respectent (1.3). Bien que n'importe quelle loi vérifiant cette condition sur les supports conduise à la convergence vers l'intégrale à évaluer, il faut faire attention au choix de ν pour assurer de bonnes propriétés à l'échantillonnage préférentiel. En plus de la capacité à obtenir facilement des réalisations de ν, il est primordial que la densité g ne soit pas trop éloignée de f. En effet, si trop de poids $\tilde{\omega}(X_k')$ sont faibles, seul un petit nombre de poids porteront presque toute la masse, induisant une très forte influence de quelques réalisations seulement. Idéalement, il faut assurer que le facteur de variance suivant reste le plus faible possible pour éviter ce problème,

$$\int_{\mathbb{R}^d} h(x)^2 \tilde{\omega}(x)^2 g(x)dx = \int_{\mathbb{R}^d} h(x)^2 \frac{\tilde{f}(x)^2}{g(x)} dx.$$

Ce sera le cas, par exemple, si le rapport $|h(x)\tilde{f}(x)/g(x)|$ reste borné pour tout $x \in \mathbb{R}$.

ALGORITHME 1.28 – Échantillonnage préférentiel

Initialisation :
- une loi μ sur $\mathbb{R}^d$ de densité proportionnelle à $\tilde{f}$
- une loi ν sur $\mathbb{R}^d$ de densité g telle que support$(\tilde{f}) \subset$ support(g)
- une fonction intégrable $h : \mathbb{R}^d \to \mathbb{R}$

Pour k allant de 1 à n :
- Tirer X_k selon la loi ν
- Calculer $h(X_k)$

Calculer $\tilde{\omega}(X_k) = \tilde{f}(X_k)/g(X_k)$
Retourner $(h(X_1)\tilde{\omega}(X_1) + \cdots + h(X_n)\tilde{\omega}(X_n))/(\tilde{\omega}(X_1) + \cdots + \tilde{\omega}(X_n))$

EXEMPLE 1.29 (Estimateur bayésien de la moyenne). Nous nous plaçons dans un cadre bayésien pour estimer la moyenne $\theta \in \mathbb{R}^p$ d'un échantillon gaussien $X_1, \ldots, X_n \sim \mathcal{N}_p(\theta, I_p)$ avec la fonction de perte quadratique. Au lieu de choisir une loi a priori conjuguée, nous proposons d'utiliser une loi uniforme sur la sphère de rayon r. Au travers de ce choix, nous n'imposons que la symétrie sphérique pour la loi a priori, ce qui est plus faible qu'une hypothèse de normalité. Soit $\alpha, \beta > 0$, nous prenons une loi a priori gamma de paramètre (α, β), notée $\Gamma(\alpha, \beta)$, pour le rayon r, *i.e.* la loi sur $\mathbb{R}_+^*$ de densité

$$\forall r > 0,\ f(r) = \frac{\beta^\alpha}{\Gamma(\alpha)} r^{\alpha-1} e^{-\beta r}.$$

La loi a priori π s'obtient par intégration et cela conduit à la loi a posteriori

$$\pi(\theta \mid \mathbf{X}) \propto \ell(\theta \mid \mathbf{X})\pi(\theta)$$

où $\ell(\theta \mid \mathbf{X}) \propto \exp(-(\|X_1 - \theta\|^2 + \cdots + \|X_n - \theta\|^2)/2)$ désigne la fonction de vraisemblance. Ainsi, l'estimateur bayésien de θ s'obtient par

$$\hat{\theta} = \int_{\mathbb{R}^p} \theta \pi(\theta \mid \mathbf{X}) d\theta = \frac{\displaystyle\int_{\mathbb{R}^p} \theta \ell(\theta \mid \mathbf{X})\pi(\theta) d\theta}{\displaystyle\int_{\mathbb{R}^p} \ell(\theta \mid \mathbf{X})\pi(\theta) d\theta}.$$

Bien que des calculs fastidieux conduisent à des formules explicites (voir l'exemple 10.26 de [**?**]), la solution fait appel à des fonctions hypergéométriques confluentes qui sont difficiles à évaluer en pratique et qui ne permettent pas de travailler directement avec cette loi a posteriori. Pour contourner cette difficulté, il peut être intéressant de générer des réalisations indépendantes $R_1, \ldots, R_m \sim \Gamma(\alpha, \beta)$, puis, pour $k \in \{1, \ldots, m\}$, de tirer θ_k uniformément sur la sphère de rayon R_k. Les variables aléatoires $\theta_1, \ldots, \theta_m$ sont des réalisations indépendantes de la loi a priori sur θ et nous approchons la moyenne a posteriori par

$$\tilde{\theta}_m = \frac{\displaystyle\sum_{k=1}^m \theta_k \exp\left(-\frac{1}{2}\sum_{i=1}^n \|X_i - \theta_k\|^2\right)}{\displaystyle\sum_{k=1}^m \exp\left(-\frac{1}{2}\sum_{i=1}^n \|X_i - \theta_k\|^2\right)}.$$

Autrement dit, l'estimateur $\hat{\theta}$ est approché par échantillonnage préférentiel avec la loi a priori comme fonction d'importance. La figure 1.13 illustre cette procédure en donnant l'erreur quadratique commise pour estimer θ par $\tilde{\theta}_m$ en fonction de différentes valeurs de α et β. ■

1.4 Optimisation de fonctions

Le troisième problème pour lequel nous considérerons des algorithmes stochastiques dans le cadre de ce cours est celui de l'optimisation d'une fonction $f : \mathbb{R}^d \to \mathbb{R}$. Il existe différentes

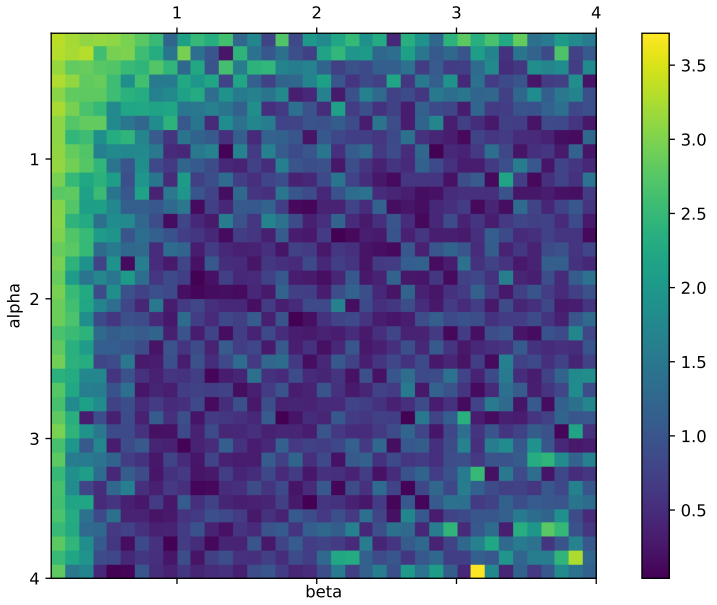

FIGURE 1.13 – Norme des résidus de l'estimateur bayésien de la moyenne d'une loi gaussienne calculé par échantillonnage préférentiel avec un a priori uniforme sur une sphère dont le rayon suit une loi $\Gamma(\alpha,\beta)$ ($m = 100$, $n = 32$ et $p = 2$).

façons de formuler cette question et nous développerons ici la recherche d'un minimiseur θ^* de f sur un ensemble $\Theta \subset \mathbb{R}^d$,

$$\theta^* \in \underset{\theta\in\Theta}{\operatorname{argmin}}\ f(\theta). \tag{1.4}$$

D'un point de vue physique, si la fonction f correspond à une mesure d'énergie, un minimiseur θ^* correspond à un état stable du système considéré. En statistique, rechercher un extremum d'une fonction est une méthode commune pour définir certains estimateurs comme l'estimateur du maximum de vraisemblance (EMV) ou l'estimateur du maximum a posteriori (MAP). La minimisation de critères empiriques est également une problématique récurrente dans de nombreux problèmes d'apprentissage automatique (*e.g.* optimisation des poids d'un réseau de neurones).

1.4.1 Recherche aléatoire

Une approche naïve pour rechercher un minimiseur θ^* d'une fonction f sur un ensemble borné Θ consiste à tirer au hasard un grand nombre de réalisations indépendantes $\theta_1,\dots,\theta_n$ selon la loi uniforme sur Θ et à définir

$$\hat{\theta}_n^* \in \underset{\theta\in\{\theta_1,\dots,\theta_n\}}{\operatorname{argmin}}\ f(\theta).$$

Si la fonction f est continue, alors nous savons que cette procédure convergera au sens où

$$f(\hat{\theta}_n^*) \xrightarrow[n\to+\infty]{p.s.} \min_{\theta\in\Theta} f(\theta).$$

Cependant, cette approche aura de mauvaises performances en pratique car une grande partie du temps de calcul sera gâché en tirant des points éloignés d'un minimiseur et elle souffrira également du *fléau de la dimension* lorsque la dimension d sera grande. En effet, si $\Theta = [0,1]^d$, pour tout $0 < \varepsilon < 1/2$, la probabilité de tirer un point dans $[\varepsilon, 1-\varepsilon]^d$ avec la loi uniforme vaut $(1-2\varepsilon)^d$ qui converge vers 0 lorsque d tend vers l'infini. Autrement dit, la loi uniforme se concentre sur les bords de Θ quand la dimension grandit et, si le minimiseur θ^* se trouve « au milieu » de Θ, alors le nombre de points à générer (et donc le temps à attendre) pour tomber dans voisinage de θ^* deviendra tellement grand qu'il ne sera plus raisonnable d'employer cette méthode en pratique.

Une meilleure approche pour rechercher aléatoirement un minimiseur consiste à tirer les points $\theta_1, \ldots, \theta_n$ selon une loi μ sur Θ adaptée à la fonction f à minimiser. Idéalement, une telle loi doit donner une grande probabilité aux régions dans lesquelles f prend des valeurs petites pour concentrer les réalisations autour de bons candidats. En pratique, il peut être compliqué de trouver ou de simuler une loi avec de telles propriétés. Lorsque cela est faisable, une possibilité consiste à considérer la loi de densité

$$\forall \theta \in \mathbb{R}^d,\ g_\mu(\theta) \propto \exp\left(-\frac{f(\theta)}{T}\right)$$

où $T > 0$ est appelé la *température*. Lorsque la température T est basse, la mesure μ se concentre sur les minima de la fonction f et les réalisations peuvent se retrouver « gelées » dans des minima locaux qui ne représentent pas notre objectif. D'un autre côté, quand la température T est grande, la séquence de points générés selon μ sera plus dispersée et nous aurons plus de chance de tomber dans un voisinage de θ^* au prix d'un plus grand nombre de réalisations inutiles. En pratique, ce paramètre ne reste pas constant au cours de la procédure et il convient de refroidir progressivement la température pour explorer les régions intéressantes dans un premier temps avant de ce concentrer sur les candidats à la minimisation. Nous verrons au chapitre 3 des stratégies d'évolution de la température pour rechercher un minimiseur de la fonction f de cette façon.

ALGORITHME 1.30 – Recherche aléatoire

Initialisation : une fonction $f : \mathbb{R}^d \to \mathbb{R}$ à minimiser sur $\Theta \subset \mathbb{R}^d$ et une loi μ sur Θ
Pour k allant de 1 à n :
  Tirer θ_k selon la loi μ
  Calculer $h(\theta_k)$
Retourner $\hat{\theta}^*$ tel que $h(\hat{\theta}^*) = \min\{f(\theta_k),\ 1 \leqslant k \leqslant n\}$

EXEMPLE 1.31 (Maximum d'un fonction). Reprenons la fonction de l'exemple 1.26,

$$\forall x \in [0,1],\ f(x) = (\cos(50x) + \sin(20x))^2 .$$

Chercher le maximum de cette fonction revient à chercher le minimum de $x \mapsto -f(x)$. Pour cela, une première approche simple consiste à considérer des variables aléatoires $U_1, \ldots, U_n$ indépendantes de même loi $\mathscr{U}([0,1])$ et à ne retenir que la réalisation $U^* \in \{U_1, \ldots, U_n\}$ telle que

$$f(U^*) = \max\{f(U_k),\ k \in \{1, \ldots, n\}\} .$$

La fonction f étant positive, il est également possible de la voir comme une densité de probabilité à une constante multiplicative près. En utilisant une méthode du rejet, nous pouvons

simuler des réalisations $X_1, \ldots, X_n$ dont la loi admet cette densité et poser $X^* \in \{X_1, \ldots, X_n\}$ telle que

$$f(X^*) = \max\{f(X_k),\ k \in \{1, \ldots, n\}\}.$$

Pour une température $T > 0$ donnée, nous pouvons aussi utiliser une méthode du rejet pour simuler des réalisations $Y_1, \ldots, Y_n$ dont la loi admet une densité proportionnelle à $x \mapsto \exp(f(x)/T)$. Le taux de rejet sera plus important que dans le cas précédent mais les points seront plus concentrés autour des maxima locaux. De même, nous retiendrons alors la réalisation $Y^* \in \{Y_1, \ldots, Y_n\}$ telle que

$$f(Y^*) = \max\{f(Y_k),\ k \in \{1, \ldots, n\}\}.$$

La figure 1.14 montre le résultat de ces trois approches avec $n = 1000$. L'utilisation des variables uniformes demande moins de temps de calcul car elle n'a pas besoin de faire appel à une méthode du rejet mais la grande dispersion des réalisations a un impact négatif sur la précision du résultat. D'autre part, les réalisations obtenues avec la densité proportionnelle à f et celle proportionnelle à $\exp(f/T)$ sont plus concentrées dans les régions où f prend de grandes valeurs, ce qui induit une meilleure précision au prix d'un temps de calcul supérieur dû aux méthodes du rejet invoquées. ■

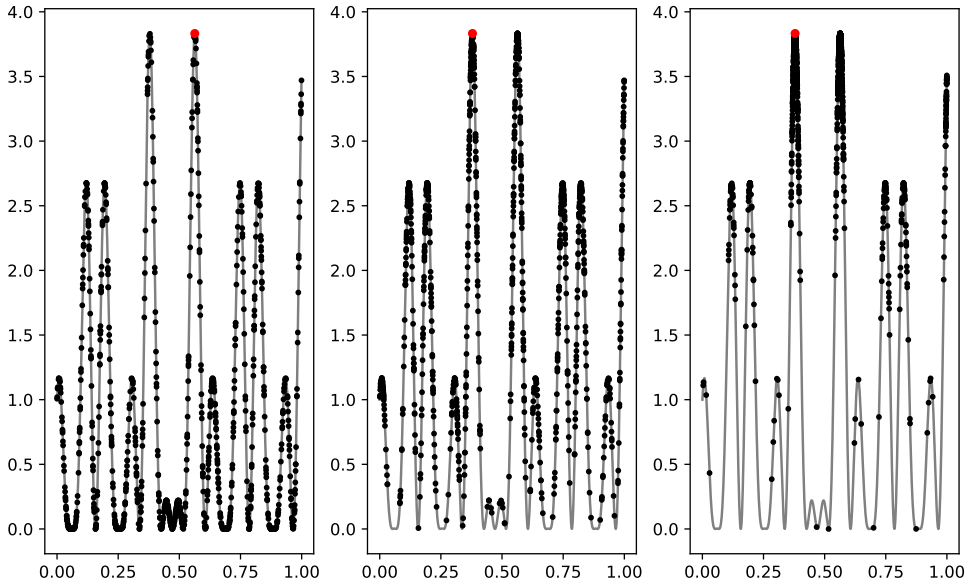

FIGURE 1.14 – Recherche aléatoire du maximum de la fonction $f(x) = (\cos(50x) + \sin(20x))^2$ à partir de $n = 1000$ points générés uniformément sur $[0,1]$ (à gauche), selon une loi de densité proportionnelle à f (au milieu) et selon une loi de densité proportionnelle à $\exp(2f)$ (à droite).

1.4.2 Méthode du gradient stochastique

Une limite de la recherche aléatoire en pratique est la difficulté à trouver une loi adaptée à la fonction à minimiser ou à simuler des réalisations selon elle. Une approche stochastique

différente pour résoudre (1.4) consiste à explorer la fonction f de manière locale plutôt que de générer des points globalement dans Θ. Pour cela, le principe est de définir une suite $(\theta_n)_{n\in\mathbb{N}}$ telle que la transition aléatoire vers θ_{n+1} dépende de θ_n et de la fonction f de façon à ce que cette suite converge vers un minimiseur θ^*. En termes mathématiques, une telle suite est une *chaîne de Markov* et nous discuterons de certaines propriétés de ces objets dans le chapitre 2.

D'un point de vue déterministe, la suite $(\theta_n)_{n\in\mathbb{N}}$ peut être construite, par exemple, grâce à la *méthode de Newton-Raphson* qui utilise le gradient de f au point θ_n pour définir le point θ_{n+1} par récurrence,

$$\forall n \in \mathbb{N},\ \theta_{n+1} = \theta_n - \gamma_n \nabla f(\theta_n) \tag{1.5}$$

avec un point de départ $\theta_0 \in \Theta$ arbitraire et une suite $(\gamma_n)_{n\in\mathbb{N}}$ à valeurs dans $\mathbb{R}_+^*$. Dans le cas favorable où Θ est un ensemble convexe, la fonction f est convexe et la suite $(\gamma_n)_{n\in\mathbb{N}}$ est bien choisie, alors θ_n converge vers θ^* quand n tend vers l'infini. Pour des fonctions f moins régulières, ce résultat ne demeure plus nécessairement vrai et la suite $(\theta_n)_{n\in\mathbb{N}}$ peut se retrouver piégée dans des minima locaux de la fonction f.

L'approche dite du *gradient stochastique* est basée sur le même principe de « petites » perturbations de θ_n pour définir θ_{n+1}. Ces perturbations aléatoires ne seront pas isotropes, comme pourrait l'être un bruit gaussien standard, mais dirigées dans une direction liée à celle du gradient. Les raisons qui peuvent motiver une approche stochastique pour un problème d'optimisation tel que (1.4) sont diverses : une fonction f non explicite ou seulement disponible par le biais d'une espérance, un gradient ∇f difficile à évaluer, ... L'évaluation du gradient est, par exemple, un problème récurrent lorsque nous sommes amenés à travailler avec des jeux de données volumineux, *i.e.* un cadre dit *big data*, où toute évaluation du gradient est extrêmement coûteuse.

Afin d'introduire cette *méthode du gradient stochastique* qui fera l'objet d'une partie importante du chapitre 3, un exemple simple est celui de l'estimation de la moyenne $\theta^* \in \mathbb{R}$ d'une variable aléatoire réelle X. En effet, cette question peut être présentée comme un problème d'optimisation,

$$\theta^* = \operatorname*{argmin}_{\theta\in\mathbb{R}} f(\theta) \quad \text{où} \quad \forall \theta \in \mathbb{R},\ f(\theta) = \frac{1}{2}\mathbb{E}\left[(X-\theta)^2\right].$$

La fonction f ne peut pas être évaluée sans connaître explicitement la loi de X, ce qui n'est généralement pas le cas dans un cadre statistique. Étant donnée une suite $(X_k)_{k\geqslant 1}$ de variables aléatoires indépendantes de même loi que X, l'estimation de θ^* se fait classiquement par la moyenne empirique,

$$\forall n \geqslant 1,\ \theta_n = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

Dans ce cas, en posant $\theta_0 = 0$, la récurrence qui relie θ_n et θ_{n+1} est simple à établir,

$$\begin{aligned}\forall n \in \mathbb{N},\ \theta_{n+1} &= \frac{1}{n+1}\sum_{k=1}^{n+1} X_k \\ &= \theta_n + \frac{1}{n+1}(X_{n+1} - \theta_n).\end{aligned}$$

La nouvelle information apportée par X_{n+1} se traduit donc par un terme additif dirigé par l'écart à la prédiction précédente θ_n et pondéré par un facteur $1/(n+1)$. Le lien avec le gradient de f (ici, il s'agit de la simple fonction dérivée f') s'exprime au travers de la décomposition suivante,

$$\forall n \in \mathbb{N},\ \theta_{n+1} = \theta_n - \gamma_{n+1} f'(\theta_n) + \gamma_{n+1}\Delta M_{n+1} \tag{1.6}$$

où nous avons posé $\gamma_{n+1} = 1/(n+1)$ et $\Delta M_{n+1} = X_{n+1} - \theta^*$. La récurrence qui apparaît est donc la même que dans (1.5) à un terme additif aléatoire près donné par les éléments de la suite $(\Delta M_n)_{n\geqslant 1}$. Cette suite a des propriétés dont nous reparlerons au chapitre 3, il s'agit en particulier d'*incréments de martingale* qui vérifient

$$\forall n \geqslant 1,\ \mathbb{E}[\Delta M_{n+1} \mid X_1, \ldots, X_n] = 0.$$

En tant que moyenne empirique, nous savons ici que θ_n converge presque sûrement vers θ^* par la loi des grands nombres et que la loi de $\sqrt{n}(\theta_n - \theta^*)$ tend vers celle d'une loi normale par le théorème central limite. En généralisant la forme de la recurrence (1.6) à d'autres suites construites comme des méthodes du gradient perturbées par des incréments de martingale, nous serons amenés à préciser quelles doivent être les hypothèses satisfaites par la fonction f, la suite $(\gamma_n)_{n\geqslant 1}$ et les variables $(\Delta M_n)_{n\geqslant 1}$ pour assurer que θ_n converge vers un minimiseur θ^* et présente d'autres propriétés intéressantes telles que la normalité asymptotique. Nous verrons en particulier que le choix des valeurs γ_n sera crucial et que le comportement de la suite $(\theta_n)_{n\in\mathbb{N}}$ en dépendra énormément comme l'illustre la figure 1.15.

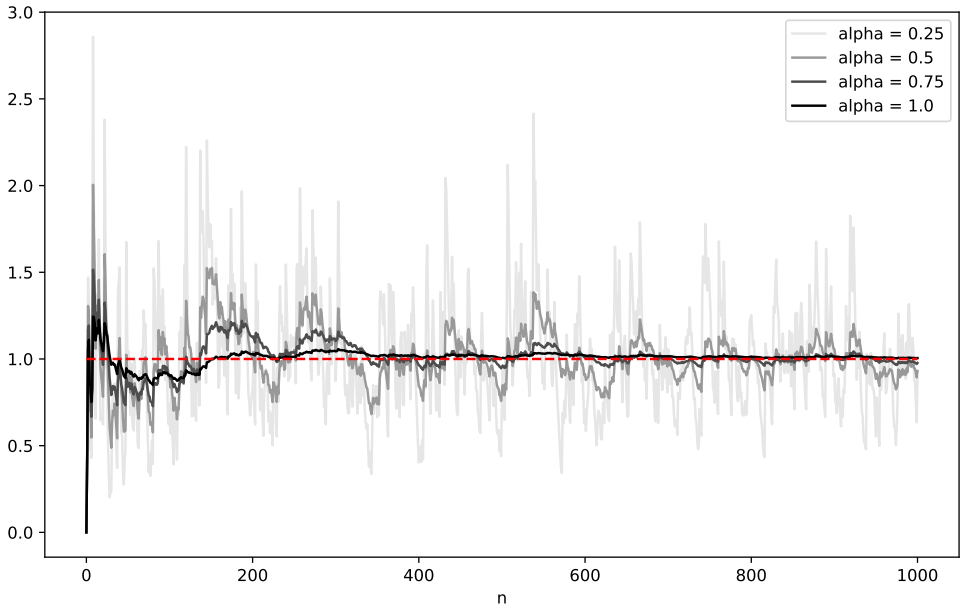

FIGURE 1.15 – Estimation de la moyenne d'une loi $\mathscr{E}(1)$ par les $n = 1000$ premières itérations de la méthode du gradient stochastique (1.6) avec $\gamma_n = n^{-\alpha}$ pour différentes valeurs de $\alpha \in [0,1]$.

2 — Méthodes de Monte Carlo par chaînes de Markov

2.1 Premier exemple : échantillonneur par tranche

Dans le chapitre précédent, nous avons présenté plusieurs algorithmes pour résoudre le problème de la simulation pour une loi μ donnée. Afin de motiver l'introduction des objets mathématiques qui joueront un rôle central dans la suite, nous considérons ici un cas particulier de l'*échantillonneur de Gibbs* (voir la section 2.4) qui permet également de générer des réalisations d'une loi « proche » de μ. La principale différence avec les méthodes évoquées au chapitre précédent réside dans le fait que ces réalisations ne seront plus indépendantes.

2.1.1 Une marche aléatoire

Nous commençons par rappeler le résultat du lemme 1.18. Étant donnée une loi μ sur $\mathbb{R}$ de densité f par rapport à la mesure de Lebesgue, si nous savons générer un vecteur aléatoire (X,Y) de loi uniforme sur

$$B = \left\{(x,y) \in \mathbb{R}^2 \text{ tel que } 0 < y < \tilde{f}(x)\right\}$$

où $\tilde{f} : \mathbb{R} \to \mathbb{R}_+$ est une fonction proportionnelle à f, alors la variable X suit la loi μ. Le principe que nous allons développer consiste à construire une *marche aléatoire* sur B, *i.e.* une suite $((X_n,Y_n))_{n\in\mathbb{N}}$ de vecteurs aléatoires à valeurs dans B construite par récurrence. L'idée sous-jacente est que si cette marche aléatoire explore B de façon « adéquate », nous pouvons espérer que la loi de (X_n,Y_n) soit proche de la loi uniforme sur B quand n devient grand et donc que la loi de la variable X_n soit proche de μ. Si la marche aléatoire est telle que, pour tout $n \in \mathbb{N}$, le point (X_{n+1},Y_{n+1}) ne dépend que de la position précédente (X_n,Y_n), il s'agit alors d'un exemple simple du concept de *chaîne de Markov* que nous développerons dans la section suivante.

Il existe plusieurs manières de construire une marche aléatoire sur B mais une façon relativement naturelle consiste à alterner les déplacements verticaux et horizontaux tout en restant dans B. Ainsi, en partant d'un point $(X_0,Y_0) \in B$ arbitraire, nous pouvons considérer un déplacement vertical en tirant une nouvelle réalisation Y_1 selon une loi uniforme sur les ordonnées,

$$Y_1 \mid X = X_0 \sim \mathscr{U}\left(\{y \in \mathbb{R} \text{ tel que } 0 < y < \tilde{f}(X_0)\}\right)$$

puis un déplacement horizontal en tirant une nouvelle réalisation X_1 selon une loi uniforme sur les abcisses,

$$X_1 \mid Y = Y_1 \sim \mathscr{U}\left(\{x \in \mathbb{R} \text{ tel que } 0 < Y_1 < \tilde{f}(x)\}\right).$$

Une itération de la marche aléatoire consiste alors à passer du point (X_0, Y_0) au point (X_1, Y_1). Il est important de remarquer ici que la valeur Y_1 est obtenue conditionnellement à l'événement $\{X = X_0\}$ puis que la valeur X_1 est obtenue conditionnellement à l'événement $\{Y = Y_1\}$ et non pas à l'événement $\{Y = Y_0\}$. En itérant cette procédure, nous construisons une marche aléatoire $((X_n, Y_n))_{n\in\mathbb{N}}$ sur B qui correspond à l'*échantillonneur par tranche* tel qu'il fut introduit dans [**?**] et [**?**]. La figure 2.1 donne un exemple des premières itérations de cette méthode.

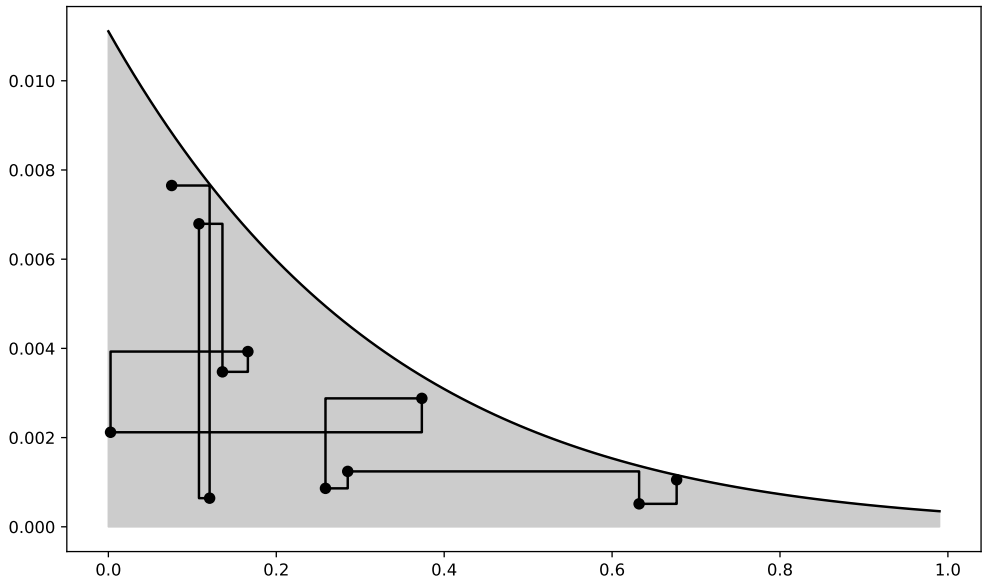

FIGURE 2.1 – Dix premières itérations de l'échantillonneur par tranche pour la loi $\mathscr{N}(-3,1)$ restreinte au segment $[0,1]$.

ALGORITHME 2.1 – Échantillonneur par tranche

Initialisation :

- une fonction $\tilde{f} : \mathbb{R} \to \mathbb{R}_+$ proportionnelle à la densité de la loi μ
- un point arbitraire $(X_0, Y_0) \in \mathbb{R}^2$ tel que $0 < Y_0 < \tilde{f}(X_0)$

À l'étape $n \geqslant 1$:

Tirer Y_n uniformément dans $[0, \tilde{f}(X_{n-1})]$

Tirer X_n uniformément dans

$$A_n = \{x \in \mathbb{R} \text{ tel que } 0 < Y_n < \tilde{f}(x)\}$$

Retourner les valeurs $X_0, X_1, \ldots$

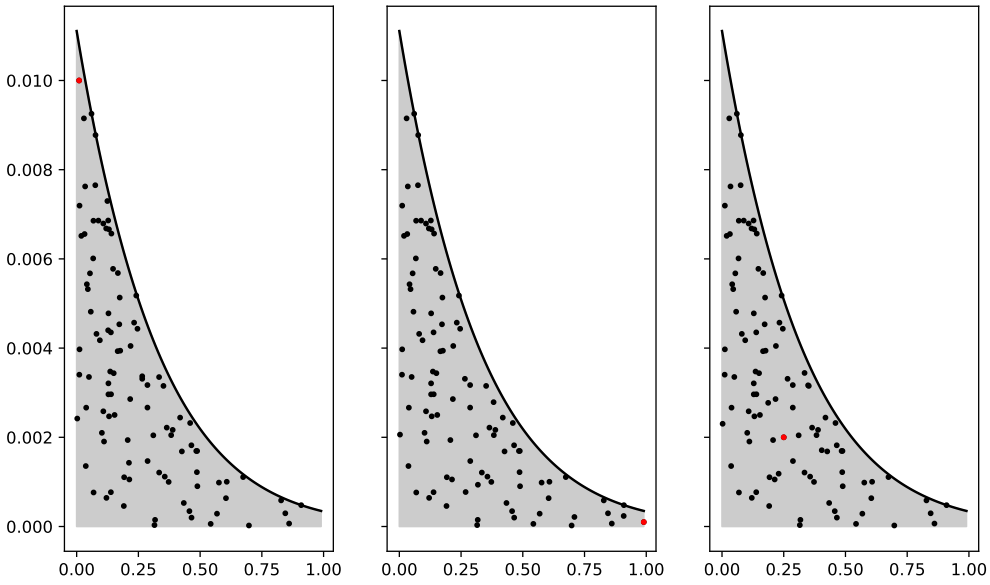

FIGURE 2.2 – Cent premières itérations de l'échantillonneur par tranche pour la loi $\mathcal{N}(-3,1)$ restreinte au segment $[0,1]$ obtenues avec la même graine et un point initial en $(0.01, 0.01)$ (gauche), $(0.99, 0.0001)$ (milieu) et $(0.25, 0.002)$ (droite).

2.1.2 Quelques propriétés de l'échantillonneur par tranche

Une première remarque concernant cet algorithme est que nous ne retenons que la suite $(X_n)_{n\in\mathbb{N}}$ en sortie et pas la suite $(Y_n)_{n\in\mathbb{N}}$ des variables dites *auxiliaires* car, sauf cas particuliers, elles n'ont pas d'interprétation propre. Une autre remarque que nous pouvons faire par rapport à des méthodes comme celle du rejet est que nous ne perdons pas de temps de calcul ici en rejetant des valeurs puisque le résultat de chaque itération est intégré à la marche aléatoire. Enfin, pour tout $n \in \mathbb{N}$, la variable X_{n+1} ne dépend que de la valeur précédente X_n et l'échantillonneur par tranche est donc une méthode qui produit une *chaîne de Markov* au sens donné ci-dessus. En pratique, tirer une variable uniforme dans A_n peut s'avérer compliqué et des variantes existent pour tenter de contourner cette difficulté. Afin de garder cette introduction simple, nous ne discuterons pas cette question ici mais le lecteur intéressé trouvera quelques pistes dans le chapitre 8 de [**?**].

Le point de départ (X_0, Y_0) de l'échantillonneur par tranche peut être fixé de façon arbitraire ou tiré selon une loi quelconque sur B. L'impact de ce choix sur le reste de la suite n'est pas fort et il s'agit d'une propriété de *perte de mémoire* de la chaîne de Markov que nous formaliserons ultérieurement. Pour illustrer ce phénomène, la figure 2.2 montre les premiers points d'un échantillonneur par tranche obtenu à partir de trois points de départ différents. Par souci d'intégrité, la même graine a été utilisée pour chaque simulation et nous observons que les points sont relativement bien répartis dans B et très similaires d'une figure à l'autre malgré une initialisation différente.

Le caractère markovien de la suite $(X_n)_{n\in\mathbb{N}}$ est au cœur des outils développés dans la suite de ce chapitre. Il se formalise au travers de la *loi de transition* qui, pour une intération $n \in \mathbb{N}$ donnée, décrit les valeurs que peut prendre X_{n+1} conditionnellement à X_n. Dans le cas de

l'échantillonneur par tranche, il est plus simple de décrire cette loi à partir des variables $\tilde{f}(X_n)$ pour lesquelles, pour tout $t, v \in \mathbb{R}_+$, nous avons

$$\begin{aligned}\mathbb{P}(\tilde{f}(X_{n+1}) \leqslant t \mid \tilde{f}(X_n) = v) &= \int_0^v \mathbb{P}(\tilde{f}(X_{n+1}) \leqslant t \mid \tilde{f}(X_n) = v \text{ et } Y_{n+1} = u)\frac{du}{v} \\ &= \frac{1}{v}\int_0^v \frac{m(u) - m(t)}{m(u)} \mathbf{1}_{u<t}\, du \\ &= \frac{1}{v}\int_0^v \left(1 - \frac{m(t)}{m(u)}\right)_+ du\end{aligned}$$

où $x_+ = \max\{x, 0\}$ et, pour tout $u > 0$, $m(u)$ désigne la mesure de Lebesgue de l'ensemble des points $x \in \mathbb{R}$ tels que $\tilde{f}(x) > u$. Il est intéressant de remarquer que cette loi de transition ne dépend pas de n et demeure donc valide à chaque itération. Nous parlerons de chaîne de Markov *homogène* pour désigner cette propriété.

Comme nous l'avons mentionné précédemment, nous sommes intéressés par la loi du vecteur (X_n, Y_n) lorsque n tend vers l'infini. Une conséquence de l'homogénéité est que si ce vecteur converge en loi, alors la loi limite π doit nécessairement être préservée lors de la transition de (X_n, Y_n) à (X_{n+1}, Y_{n+1}). Autrement dit, si la loi π existe et que $(X_n, Y_n) \sim \pi$, alors $(X_{n+1}, Y_{n+1}) \sim \pi$. Dans le vocabulaire des chaînes de Markov, la loi π est alors appelée *probabilité invariante* et des résultats d'existence et d'unicité d'une telle loi occuperont une partie importante de la section 2.2. Pour l'échantillonneur par tranche, il convient de remarquer que la loi uniforme sur B est invariante, ce qui en fait un bon candidat pour la loi limite et valide notre motivation initiale. En effet, si (X_n, Y_n) suit la loi uniforme sur B, alors nous savons que X_n suit la loi μ et la densité du vecteur (X_n, Y_{n+1}) est donnée par

$$(t, y) \in \mathbb{R}^2 \mapsto f(t) \times \frac{\mathbf{1}_{0<y<\tilde{f}(t)}}{\tilde{f}(t)} \propto \mathbf{1}_{0<y<\tilde{f}(t)}$$

car il existe $C > 0$ tel que, pour tout $t \in \mathbb{R}$, $\tilde{f}(t) = f(t)/C$. De même, la densité du vecteur (X_n, X_{n+1}, Y_{n+1}) s'écrit

$$(t, x, y) \in \mathbb{R}^3 \mapsto C\mathbf{1}_{0<y<\tilde{f}(t)} \times \frac{\mathbf{1}_{y<\tilde{f}(x)}}{m(y)}.$$

En intégrant par rapport à X_n, nous obtenons la densité du vecteur (X_{n+1}, Y_{n+1}),

$$(x, y) \in \mathbb{R}^2 \mapsto C\mathbf{1}_{0<y<\tilde{f}(x)} \int_{\mathbb{R}} \frac{\mathbf{1}_{y<\tilde{f}(t)}}{m(y)} dt \propto \mathbf{1}_{0<y<\tilde{f}(x)}.$$

Ainsi, le vecteur aléatoire (X_{n+1}, Y_{n+1}) suit également la mesure uniforme sur B qui est donc invariante pour l'échantillonneur par tranche.

Soit $x \in \mathbb{R}$, nous notons $\mathbb{P}^x$ la loi de l'échantillonneur par tranche issu de $X_0 = x$. Si la fonction $\tilde{f}$ est bornée et que son support est également borné, il est possible de montrer à partir des résultats précédents que la chaîne de Markov $(X_n)_{n\in\mathbb{N}}$ est *uniformément ergodique*, c'est-à-dire qu'elle satisfait

$$\forall x \in \mathbb{R},\ \|\mathbb{P}^x(X_n \in \cdot) - \mu\|_{VT} = \sup_A |\mathbb{P}^x(X_n \in A) - \mu(A)| \xrightarrow[n\to+\infty]{} 0 \tag{2.1}$$

où le supremum est pris sur tous les sous-ensemble $A \subset \mathbb{R}$ mesurables. La norme $\|\cdot\|_{VT}$ qui apparaît ici s'appelle la *norme en variation totale* et sera utilisée dans la suite pour caractériser

la convergence vers la probabilité invariante. La preuve de ce résultat dépasse le cadre de cette introduction mais il nous permet de résoudre approximativement le problème de la simulation pour μ en considérant, par exemple, K réalisations indépendantes $(X_{1,n})_{n\in\mathbb{N}}, \dots, (X_{K,n})_{n\in\mathbb{N}}$ de l'échantillonneur par tranche issu de $x \in \mathbb{R}$ et en approchant μ par la n-ème itération de chacune de ces suites,

$$\forall A \subset \mathbb{R} \text{ mesurable}, \ \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}_{X_{k,n}\in A} \simeq \mathbb{P}^x(X_n \in A) \simeq \mu(A)$$

pour n et K assez grands. La figure 2.3 illustre un résultat de cette procédure de simulation. La marche aléatoire $(X_n)_{n\in\mathbb{N}}$ vérifie une autre propriété dont la formulation ressemble à la précédente, il s'agit du *théorème ergodique* que nous présenterons ultérieurement. Ce résultat remarquable généralise la loi des grands nombres à la convergence de la moyenne empirique le long d'une trajectoire de la chaîne de Markov,

$$\frac{1}{n}\sum_{k=0}^{n} h(X_n) \xrightarrow[n\to+\infty]{p.s.} \int_{\mathbb{R}} h(x) f(x) dx \tag{2.2}$$

pour toute fonction h intégrable.

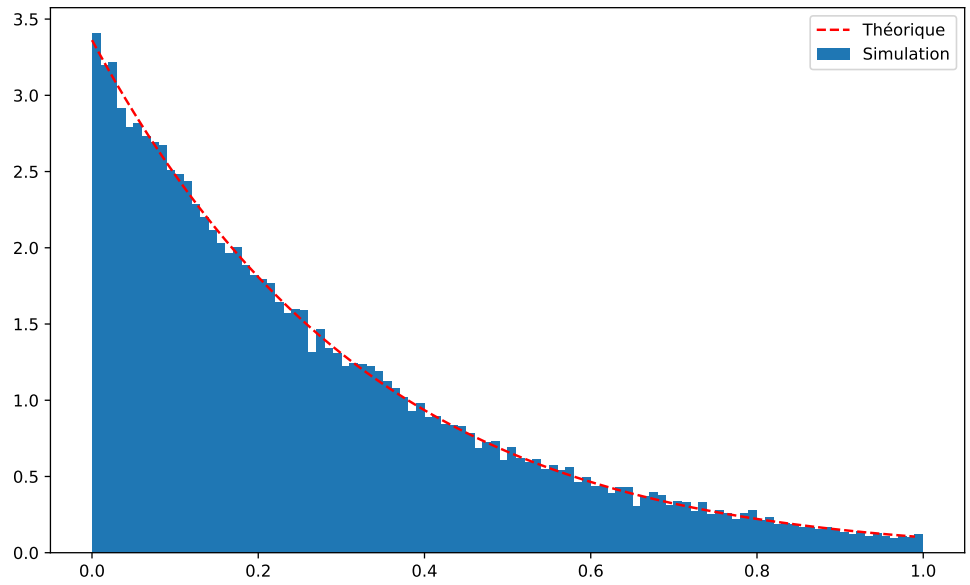

FIGURE 2.3 – Simulation de la loi $\mathcal{N}(-3,1)$ restreinte au segment $[0,1]$ approchée par la loi de la centième itération d'un échantillonneur par tranche.

Des algorithmes tels que l'échantillonneur par tranche qui produisent des chaînes de Markov dont la probabilité invariante est la *loi cible* μ constituent l'objet principal de ce chapitre et sont appelées des *méthodes de Monte Carlo par chaînes de Markov* (ou *méthodes MCMC* pour *Monte Carlo Markov Chains*). La référence à la méthode de Monte Carlo introduite à la sous-section 1.3.1 se justifie par des propriétés telles que (2.2).

Définition 2.2. Nous appelons *méthode de Monte Carlo par chaînes de Markov* (ou *méthode MCMC*) tout algorithme stochastique de simulation d'une loi μ qui produise une chaîne de Markov ergodique de probabilité invariante μ.

2.2 Quelques propriétés des chaînes de Markov

Cette section a pour objectif de présenter des objets probabilistes importants, appelés *chaînes de Markov* en référence au mathématicien russe Andreï Markov (1856-1922) qui les introduisit vers 1906, ainsi que quelques unes de leurs propriétés qui seront utiles pour la suite. Il ne s'agit ici que d'une introduction à ce vaste sujet mathématique et le lecteur intéressé trouvera les preuves des résultats admis et plus de détails dans [**?**] et [**?**] ainsi que dans l'ouvrage de référence [**?**].

FIGURE 2.4 – Andreï Markov (1856-1922)

2.2.1 Définitions

Par souci de simplicité, nous limiterons notre présentation à des variables aléatoires à valeurs dans un ensemble E au plus dénombrable muni de la tribu de ses parties $\mathscr{P}(E)$, ce qui rend toute fonction $f : E \to \mathbb{R}$ continue et donc mesurable. Par définition, une mesure de probabilité μ sur $(E, \mathscr{P}(E))$ vérifie $\mu(E) = 1$ et

$$\forall A \in \mathscr{P}(E),\ \mu(A) = \sum_{x \in A} \mu(x)$$

où la valeur de la probabilité au point $x \in E$ est notée $\mu(x) = \mu(\{x\})$. Nous dirons que la mesure de probabilité μ *charge* l'état $x \in E$ lorsque $\mu(x) > 0$.

Une fonction $f : E \to \mathbb{R}$ est dite *μ-intégrable* si la somme

$$\sum_{x\in E} |f(x)|\mu(x)$$

converge et, dans ce cas, nous utiliserons les notations suivantes pour désigner l'*intégrale de f par rapport à μ*,

$$\mu f = \int_E f(x)d\mu(x) = \sum_{x\in E} f(x)\mu(x).$$

La loi d'une suite $(X_n)_{n\in\mathbb{N}}$ de variables aléatoires à valeurs dans E est caractérisée par les lois marginales des vecteurs aléatoires $(X_0,\dots,X_m)$ pour tout $m \in \mathbb{N}$, *i.e.* par la donnée des probabilités $\mathbb{P}(X_0 = x_0,\dots,X_m = x_m)$ pour tout $x_0,\dots,x_m \in E$. De façon équivalente, il est possible de définir la loi de la suite $(X_n)_{n\in\mathbb{N}}$ grâce à la loi de X_0 et aux probabilités conditionnelles $\mathbb{P}(X_{n+1} = x_{n+1} \mid X_0 = x_0,\dots,X_n = x_n)$ pour tout $n \in \mathbb{N}$ et $x_0,\dots,x_{n+1} \in E$ car nous avons par conditionnements successifs

$$\mathbb{P}(X_0 = x_0,\dots,X_m = x_m) = \mathbb{P}(X_0 = x_0) \times \prod_{n=0}^{m-1} \mathbb{P}(X_{n+1} = x_{n+1} \mid X_0 = x_0,\dots,X_n = x_n).$$

Il est courant d'utiliser un vocabulaire spatio-temporel pour parler des éléments d'une suite $(X_n)_{n\in\mathbb{N}}$ de variables aléatoires. Ainsi, pour $n \in \mathbb{N}$, la variable X_n représente la position spatiale à l'intant n, la tribu engendrée par les variables $X_0,\dots,X_{n-1}$ représente son passé et celle engendrée par $X_{n+1}, X_{n+2},\dots$ son futur. Dans ce cadre, une *chaîne de Markov* est souvent présentée comme une suite de variables aléatoires dont le passé et le futur sont indépendants conditionnellement au présent, ce qui se formalise par la définition suivante.

Définition 2.3. Une suite $(X_n)_{n\in\mathbb{N}}$ de variables aléatoires à valeurs dans un ensemble E dénombrable est appelée *chaîne de Markov d'espace d'états E* lorsque pour tout $n \in \mathbb{N}$ et $x_0,\dots,x_{n+1} \in E$,

$$\mathbb{P}(X_{n+1} = x_{n+1} \mid X_0 = x_0,\dots,X_n = x_n) = \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n). \tag{2.3}$$

De plus, la chaîne est dite *homogène* si la loi conditionnelle de X_{n+1} sachant X_n ne dépend pas de n,

$$\forall x, y \in E,\ \mathbb{P}(X_{n+1} = y \mid X_n = x) = \mathbb{P}(X_1 = y \mid X_0 = x).$$

Dans le cadre de ce cours, toute chaîne de Markov $(X_n)_{n\in\mathbb{N}}$ sera supposée homogène. De fait, nous pouvons introduire une fonction $P : E \times E \to [0,1]$, appelée *noyau de transition* de la chaîne, pour encoder la probabilité de passer d'un état $x \in E$ à un état $y \in E$ par

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

Par construction, le noyau de transition vérifie

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

et, si ν désigne la loi de X_0, les probabilités marginales de la chaîne de Markov sont données par

$$\forall m \in \mathbb{N},\ \forall x_0,\dots,x_m \in E,\ \mathbb{P}(X_0 = x_0,\dots,X_m = x_m) = \nu(x_0) \prod_{n=0}^{m-1} P(x_n,x_{n+1}).$$

La loi de la chaîne de Markov est donc entièrement caractérisée par la *loi initiale* ν et le noyau de transition P.

EXEMPLE 2.4 (Variables indépendantes). Une suite $(X_n)_{n\in\mathbb{N}}$ de variables aléatoires indépendantes et identiquement distribuées selon une loi μ sur E constitue un exemple élémentaire de chaîne de Markov. Dans ce cas, le noyau de transition ne dépend plus de l'état de départ,

$$\forall x,y\in E,\ P(x,y)=\mu(y).$$

■

EXEMPLE 2.5 (Récurrence aléatoire). Considérons une variable aléatoire X_0 à valeurs dans un espace E dénombrable et une suite $(\xi_n)_{n\in\mathbb{N}}$ de variables aléatoires à valeurs dans un espace mesurable F indépendantes, identiquement distribuées et indépendantes de X_0. Une façon canonique de construire une chaîne de Markov homogène issue de X_0 consiste à considérer une fonction $g : E\times F\to E$ et à représenter l'innovation au temps n par ξ_{n+1} au travers de la relation de récurrence

$$\forall n\in\mathbb{N},\ X_{n+1}=g(X_n,\xi_{n+1}).$$

Soient $n\in\mathbb{N}$ et $x_0,\dots,x_{n+1}\in E$, l'événement $\{X_0=x_0,\dots,X_n=x_n\}$ ne dépend que des variables $X_0,\xi_1,\dots,\xi_n$ et la propriété (2.3) est satisfaite par indépendance,

$$\begin{aligned}\mathbb{P}(X_{n+1}=x_{n+1}\mid X_0=x_0,\dots,X_n=x_n)&=\frac{\mathbb{P}\left(\{g(x_n,\xi_{n+1})=x_{n+1}\}\cap\{X_0=x_0,\dots,X_n=x_n\}\right)}{\mathbb{P}(X_0=x_0,\dots,X_n=x_n)}\\&=\frac{\mathbb{P}(g(x_n,\xi_{n+1})=x_{n+1})\times\mathbb{P}(X_0=x_0,\dots,X_n=x_n)}{\mathbb{P}(X_0=x_0,\dots,X_n=x_n)}\\&=\mathbb{P}(g(x_n,\xi_{n+1})=x_{n+1})\\&=\mathbb{P}(X_{n+1}=x_{n+1}\mid X_n=x_n).\end{aligned}$$

Le passage de l'état x_n à l'état x_{n+1} correspond à l'événement $\{g(x_n,\xi_{n+1})=x_{n+1}\}$ qui ne dépend que de la variable ξ_{n+1}, la chaîne est donc homogène et ce raisonnement donne le noyau de transition,

$$\forall x,y\in E,\ P(x,y)=\mathbb{P}(g(x,\xi_0)=y).$$

■

EXEMPLE 2.6 (Marche aléatoire simple sur le cercle). Soit un entier $k>0$, considérons l'espace $E=\mathbb{Z}/k\mathbb{Z}$, une variable aléatoire X_0 de loi ν sur E et une suite $(R_n)_{n\in\mathbb{N}}$ de variables aléatoires indépendantes, identiquement distribuées selon la loi de Rademacher $\mathscr{R}(1/2)$ et indépendantes de X_0. La suite $(X_n)_{n\in\mathbb{N}}$ de variables aléatoires définies par la relation de récurrence

$$\begin{aligned}\forall n\in\mathbb{N},\ X_{n+1}&=X_n+R_{n+1}\pmod k\\&=X_0+R_1+\cdots+R_{n+1}\pmod k\end{aligned}$$

est une chaîne de Markov. En effet, nous retrouvons l'exemple de la récurrence aléatoire avec $g(x,r)=x+r\pmod k$ et le noyau de transition est donné par

$$\forall x,y\in\mathbb{Z}/k\mathbb{Z},\ P(x,y)=\begin{cases}1/2 & \text{si } |x-y|=1\pmod k,\\ 0 & \text{sinon.}\end{cases}$$

Dans ce contexte, le noyau de transition P peut être vu comme une matrice carrée de taille $k \times k$,

$$P = \begin{pmatrix} 0 & \frac{1}{2} & 0 & \dots & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & \dots & 0 \\ 0 & \frac{1}{2} & 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \ddots & \ddots & \frac{1}{2} \\ \frac{1}{2} & 0 & \dots & 0 & \frac{1}{2} & 0 \end{pmatrix}.$$

■

EXEMPLE 2.7 (Processus de Bernoulli). Considérons une suite $(B_n)_{n\in\mathbb{N}}$ de variables aléatoires indépendantes et identiquement distribuées selon la loi de Bernoulli $\mathscr{B}(p)$ avec $p \in [0,1]$. La suite $(X_n)_{n\in\mathbb{N}}$ définie par $X_0 = 0$ et par la relation de récurrence

$$\forall n \in \mathbb{N},\ X_{n+1} = X_n + B_{n+1} = B_1 + \cdots + B_{n+1}$$

est une chaîne de Markov d'espace d'états $E = \mathbb{N}$ de noyau de transition donné par

$$\forall x, y \in \mathbb{N},\ P(x,y) = \begin{cases} p & \text{si } y = x+1, \\ 1-p & \text{si } y = x, \\ 0 & \text{sinon.} \end{cases}$$

Cette chaîne est appelée *processus de Bernoulli* et toute réalisation correspond à une trajectoire en escalier dans $\mathbb{N}^2$ telle que celle représentée dans la figure 2.5. ■

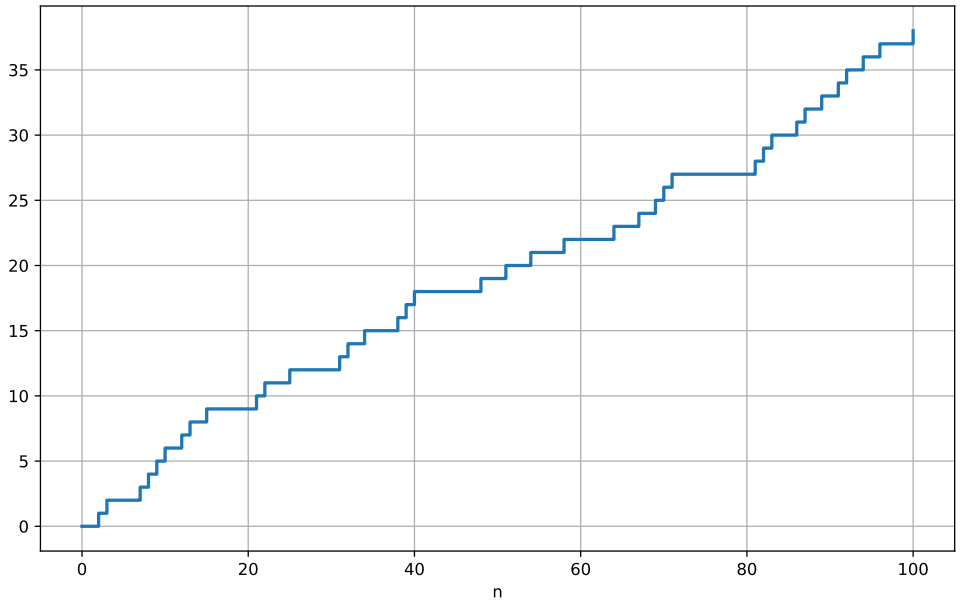

FIGURE 2.5 – Exemple des 100 premiers pas d'un processus de Bernoulli ($p = 0.4$).

EXEMPLE 2.8 (File d'attente). Nous parlons de *file d'attente* pour désigner un ensemble d'éléments à traiter séquentiellement (clients à une caisse, tâches d'impression, ...) dont le nombre évolue au cours du temps à la baisse lorsque des traitements se terminent ou à la hausse lorsque de nouveaux éléments arrivent. Un modèle simple pour représenter le nombre d'éléments dans une file d'attente consiste à supposer qu'à chaque temps $n \in \mathbb{N}$, un élément est traité si la file n'est pas vide et ξ_{n+1} nouveaux éléments arrivent. En supposant que la file soit initialement vide et que les variables aléatoires de la suite $(\xi_n)_{n\in\mathbb{N}}$ soient indépendantes et de même loi sur $\mathbb{N}$, le nombre X_n d'éléments dans la file au temps $n \in \mathbb{N}$ est donné par $X_0 = 0$ et

$$\forall n \in \mathbb{N},\ X_{n+1} = (X_n - 1)_+ + \xi_{n+1}$$

où $x_+ = \max\{x, 0\}$. En tant que récurrence aléatoire, la suite $(X_n)_{n\in\mathbb{N}}$ est une chaîne de Markov à valeurs dans $\mathbb{N}$ de noyau de transition

$$\forall x, y \in \mathbb{N},\ P(x, y) = \mathbb{P}(\xi_0 = y - (x-1)_+).$$

Selon la loi des arrivées ξ_n, le comportement de X_n lorsque n grandit peut être très différent comme l'illustre la figure 2.6. ■

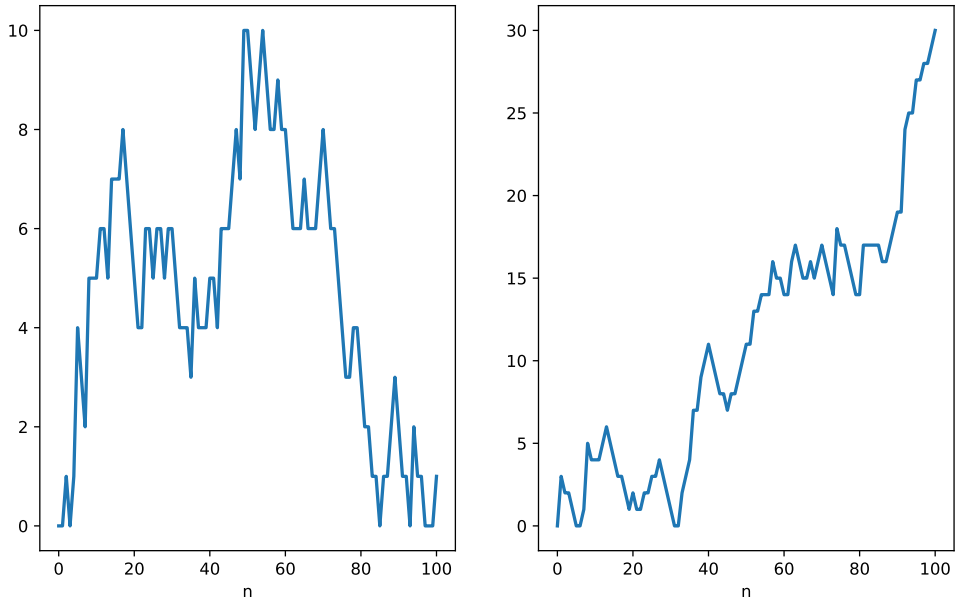

FIGURE 2.6 – Deux trajectoires du nombre d'éléments dans une file d'attente au cours du temps où les arrivées sont distribuées selon une loi de Poisson de paramètre λ. À gauche, $\lambda = 0.8$ et le taux de service semble suffisant pour résorber la demande. À droite, $\lambda = 1.2$ et la taille de la file semble diverger avec le temps, les arrivées sont trop importantes.

2.2.2 Lois instantanées

Il est possible de définir la notion de noyau de transition indépendamment du concept de chaîne de Markov.

Définition 2.9. Soit E un espace dénombrable, une application $P : E \times E \to [0,1]$ est un *noyau de transition* si
$$\forall x \in E,\ \sum_{y \in E} P(x,y) = 1.$$

Il est intéressant de remarquer que l'ensemble des noyaux de transition sur un espace E est convexe, *i.e.* si P et Q sont deux noyaux de transition sur E, alors, pour tout $\alpha \in [0,1]$, l'application $\alpha P + (1-\alpha)Q$ est aussi un noyau de transition sur E. De plus, le produit de P et de Q, noté PQ et défini par
$$\forall x,y \in E,\ PQ(x,y) = \sum_{z \in E} P(x,z)Q(z,y),$$
est encore un noyau de transition. L'élément neutre I pour ce produit est donné par le noyau trivial
$$\forall x,y \in E,\ I(x,y) = \begin{cases} 1 & \text{si } x = y, \\ 0 & \text{sinon.} \end{cases} \tag{2.4}$$
Lorsque l'espace E est fini et que les noyaux de transition sont vus comme des matrices (voir l'exemple 2.6), le produit PQ correspond au produit matriciel entre P et Q et I à la matrice identité.

Considérons une chaîne de Markov $(X_n)_{n \in \mathbb{N}}$ de noyau de transition P sur E. En posant $P^0 = I$ et $P^1 = P$, nous pouvons écrire que la probabilité de passer d'un état $x \in E$ à un état $y \in E$ après n itérations de la chaîne est donnée par le noyau de transition P^n. En effet, par conditionnements successifs,
$$\begin{aligned} \mathbb{P}(X_n = y \mid X_0 = x) &= \sum_{x_1,\dots,x_{n-1} \in E} \mathbb{P}(X_1 = x_1, \dots, X_{n-1} = x_{n-1}, X_n = y \mid X_0 = x) \\ &= \sum_{x_1,\dots,x_{n-1} \in E} P(x,x_1)P(x_1,x_2)\dots P(x_{n-1},y) \\ &= \sum_{x_2,\dots,x_{n-1} \in E} P^2(x,x_2)P(x_2,x_3)\dots P(x_{n-1},y) \\ &= \dots \\ &= P^n(x,y) \end{aligned}$$
Nous retrouvons également par ce calcul les formules bien connues dans le cadre matriciel,
$$P^n = PP^{n-1} = P^{n-1}P.$$

Pour désigner la loi de la chaîne issue de $X_0 = x$ dans la suite, nous utiliserons la notation $\mathbb{P}^x$. Ainsi, pour tout $x,y \in E$ et $n \in \mathbb{N}$, $\mathbb{P}^x(X_n = y) = P^n(x,y)$. Plus généralement, si la loi de X_0 est ν, nous noterons la loi de la chaîne $\mathbb{P}^\nu$,
$$\forall y \in E,\ \mathbb{P}^\nu(X_n = y) = \sum_{x \in E} P^n(x,y)\nu(x).$$

Nous avons $\mathbb{P}^x = \mathbb{P}^{\delta_x}$ où δ_x est la masse de Dirac en x et ces notations s'étendent naturellement aux espérances, pour tout $x \in E$, toute loi ν sur E et toute fonction intégrable $f : E \to \mathbb{R}$,
$$\mathbb{E}^x[f(X_n)] = \mathbb{E}[f(X_n) \mid X_0 = x] = \sum_{y \in E} f(y)P^n(x,y) = \delta_x P^n f$$
et
$$\mathbb{E}^\nu[f(X_n)] = \sum_{x \in E} \mathbb{E}[f(X_n) \mid X_0 = x]\nu(x) = \nu P^n f.$$

2.2.3 Propriété de Markov forte

Une conséquence immédiate de la définition d'une chaîne de Markov homogène $(X_n)_{n\in\mathbb{N}}$ d'espace d'états E dénombrable, de loi initiale ν et de noyau de transition P est que, pour tous les entiers $m,n \in \mathbb{N}$ avec $n \geqslant 1$ et tous les états $x_0,\ldots,x_m,x'_1,\ldots,x'_n \in E$,

$$\begin{aligned}&\mathbb{P}(X_{m+1}=x'_1,\ldots,X_{m+n}=x'_n \mid X_0=x_0,\ldots,X_m=x_m)\\&=\mathbb{P}(X_{m+1}=x'_1,\ldots,X_{m+n}=x'_n \mid X_m=x_m)\\&=\mathbb{P}(X_1=x'_1,\ldots,X_n=x'_n \mid X_0=x_m).\end{aligned}$$

Cela implique la *propriété de Markov faible* : soit $x \in E$, pour tout $m \in \mathbb{N}$ et conditionnellement à l'événement $\{X_m = x\}$, la suite $(X_{m+n})_{n\in\mathbb{N}}$ est une chaîne de Markov homogène, de loi initiale δ_x, de noyau de transition P et indépendante du passé de X_m. Notre objectif ici est de montrer que cette propriété s'étend à certains décalages en temps aléatoires.

Définition 2.10. Un *temps d'arrêt* T associé à une suite de variables aléatoires $(X_n)_{n\in\mathbb{N}}$ est une variable aléatoire à valeurs dans $\mathbb{N}\cup\{+\infty\}$ telle que, pour tout $n \in \mathbb{N}$, l'événement $\{T = n\}$ ne dépend que des variables $X_0,\ldots,X_n$.

EXEMPLE 2.11 (Temps d'atteinte). Soit $(X_n)_{n\in\mathbb{N}}$ une chaîne de Markov d'espace d'états E. Pour tout $A \subset E$, le premier *temps d'atteinte* de A par la chaîne de Markov est

$$T_A = \inf\{n \in \mathbb{N} \text{ tel que } X_n \in A\}$$

avec éventuellement $T_A = +\infty$ si le sous-ensemble A n'est jamais atteint. Soit $n \in \mathbb{N}$, nous avons

$$T_A = n \iff X_0 \notin A, \ldots X_{n-1} \notin A, X_n \in A$$

et T_A est donc un temps d'arrêt pour la chaîne de Markov. Dans la suite, nous noterons également $T_x = T_{\{x\}}$ lorsque le sous-ensemble est réduit à un singleton. ■

EXEMPLE 2.12 (Temps de premier passage). Soit $(X_n)_{n\in\mathbb{N}}$ une chaîne de Markov d'espace d'états E. Pour tout $A \subset E$, le *temps de premier passage* en A est défini par

$$T_A^+ = \inf\{n \geqslant 1 \text{ tel que } X_n \in A\}$$

avec éventuellement $T_A^+ = +\infty$ si le sous-ensemble A n'est jamais atteint. La variable T_A^+ est un temps d'arrêt pour la chaîne de Markov. De plus, si $X_0 \notin A$, T_A^+ coïncide avec le temps d'atteinte T_A et, si $X_0 \in A$, T_A^+ est également appelé *temps de retour* en A. Dans la suite, nous noterons également $T_x^+ = T_{\{x\}}^+$ lorsque le sous-ensemble est réduit à un singleton. ■

Théorème 2.13 (Propriété de Markov forte)**.** *Soit $(X_n)_{n\in\mathbb{N}}$ une chaîne de Markov homogène d'espace d'états E dénombrable, de loi initiale ν et de noyau de transition P. Si T est un temps d'arrêt, alors, pour tout $x \in E$, conditionnellement à l'événement $\{T < +\infty$ et $X_T = x\}$, la suite $(X_{T+n})_{n\in\mathbb{N}}$ est une chaîne de Markov homogène, de loi initiale δ_x, de noyau de transition P et indépendante du passé de X_T.*

Démonstration. Considérons un événement A qui ne dépend que de $X_0,\dots,X_T$. Fixons $m\in\mathbb{N}$, l'événement $\{T=m\}$ ne dépend que de $X_0,\dots,X_m$ par définition d'un temps d'arrêt et, sachant cet événement, A ne dépend aussi que de $X_0,\dots,X_m$. Ainsi, pour tout $x\in E$, il en va de même pour l'événement

$$A\cap\{T=m\}\cap\{X_T=x\}.$$

Soient un entier $n\geqslant 1$ et des états $x_1,\dots,x_n\in E$. Étant donné que le décalage m est fixé, la propriété de Markov faible implique

$$\begin{aligned}&\mathbb{P}^\nu\left(\{X_{T+1}=x_1,\dots,X_{T+n}=x_n\}\cap A\cap\{T=m\}\cap\{X_T=x\}\right)\\&=\mathbb{P}^\nu\left(\{X_{m+1}=x_1,\dots,X_{m+n}=x_n\}\cap A\cap\{T=m\}\cap\{X_m=x\}\right)\\&=\mathbb{P}(X_1=x_1,\dots,X_n=x_n\mid X_0=x)\mathbb{P}^\nu\left(A\cap\{T=m\}\cap\{X_m=x\}\right).\end{aligned}$$

Il suffit maintenant de sommer ces équations par rapport à m pour obtenir

$$\begin{aligned}&\mathbb{P}^\nu\left(\{X_{T+1}=x_1,\dots,X_{T+n}=x_n\}\cap A\cap\{T<+\infty\}\cap\{X_T=x\}\right)\\&=\mathbb{P}(X_1=x_1,\dots,X_n=x_n\mid X_0=x)\mathbb{P}^\nu\left(A\cap\{T<+\infty\}\cap\{X_T=x\}\right).\end{aligned}$$

Nous obtenons la conclusion du théorème en divisant par $\mathbb{P}^\nu(T<+\infty \text{ et } X_T=x)$,

$$\begin{aligned}&\mathbb{P}^\nu\left(\{X_{T+1}=x_1,\dots,X_{T+n}=x_n\}\cap A\mid T<+\infty \text{ et } X_T=x\right)\\&=\mathbb{P}(X_1=x_1,\dots,X_n=x_n\mid X_0=x)\mathbb{P}^\nu\left(A\mid T<+\infty \text{ et } X_T=x\right).\end{aligned}$$

□

La propriété de Markov forte joue un rôle très important dans l'étude des chaînes de Markov comme l'illustre l'exemple suivant.

Exemple 2.14 (Nombre d'excursions). Soit $(X_n)_{n\in\mathbb{N}}$ une chaîne de Markov homogène d'espace d'états E dénombrable. Étant donné un état $y\in E$, une portion de la trajectoire de la chaîne entre deux visites de y est appelée une *excursion* issue de y (voir la figure 2.7). Le nombre de telles excursions correspond à la variable aléatoire suivante qui prend ses valeurs dans $\mathbb{N}\cup\{+\infty\}$,

$$N_y=\sum_{n\geqslant 1}\mathbf{1}_{X_n=y}.$$

Si la chaîne n'atteint jamais l'état y, il peut n'y avoir aucune excursion et $N_y=0$. En partant d'un état $x\in E$, le nombre moyen d'excursions issues de y est donné par la fonction G définie par

$$\forall x,y\in E,\ G(x,y)=\mathbb{E}^x[N_y]=\sum_{n\geqslant 1}P^n(x,y).$$

Par définition, une excursion est liée au temps de retour T_y^+. Pour considérer cette relation, nous introduisons la série

$$G_\lambda(x,y)=\sum_{n\geqslant 1}e^{-\lambda n}P^n(x,y)$$

où $\lambda>0$. Cette série converge puisque, pour tout $n\geqslant 1$, $P^n(x,y)\in[0,1]$ et comme tous les termes sont positifs, nous avons

$$G_\lambda(x,y)=\sum_{n\geqslant 1}e^{-\lambda n}\mathbb{P}^x(X_n=y)=\mathbb{E}^x\left[\sum_{n\geqslant 1}e^{-\lambda n}\mathbf{1}_{X_n=y}\right].$$

La définition du temps d'arrêt T_y^+ et la propriété de Markov forte donnent

$$\begin{aligned} G_\lambda(x,y) &= \mathbb{E}^x\left[e^{-\lambda T_y^+} + \sum_{n\geqslant 1} e^{-\lambda(T_y^+ + n)} \mathbf{1}_{X_{T_y^+ + n} = y}\right] \\ &= \mathbb{E}^x\left[e^{-\lambda T_y^+}\right] + \sum_{n\geqslant 1} e^{-\lambda n} \mathbb{E}^x\left[e^{-\lambda T_y^+} \mathbf{1}_{X_{T_y^+ + n} = y}\right] \\ &= \left(1 + \sum_{n\geqslant 1} e^{-\lambda n} P^n(y,y)\right) \mathbb{E}^x\left[e^{-\lambda T_y^+}\right] \\ &= (1 + G_\lambda(y,y))\, \mathbb{E}^x\left[e^{-\lambda T_y^+}\right]. \end{aligned}$$

Nous obtenons ainsi la transformée de Laplace de T_y^+,

$$\mathbb{E}^x\left[e^{-\lambda T_y^+}\right] = \frac{G_\lambda(x,y)}{1 + G_\lambda(y,y)},$$

et en faisant tendre λ vers 0, nous déduisons

$$\mathbb{P}^x(T_y^+ < +\infty) = \frac{G(x,y)}{1 + G(y,y)}.$$

■

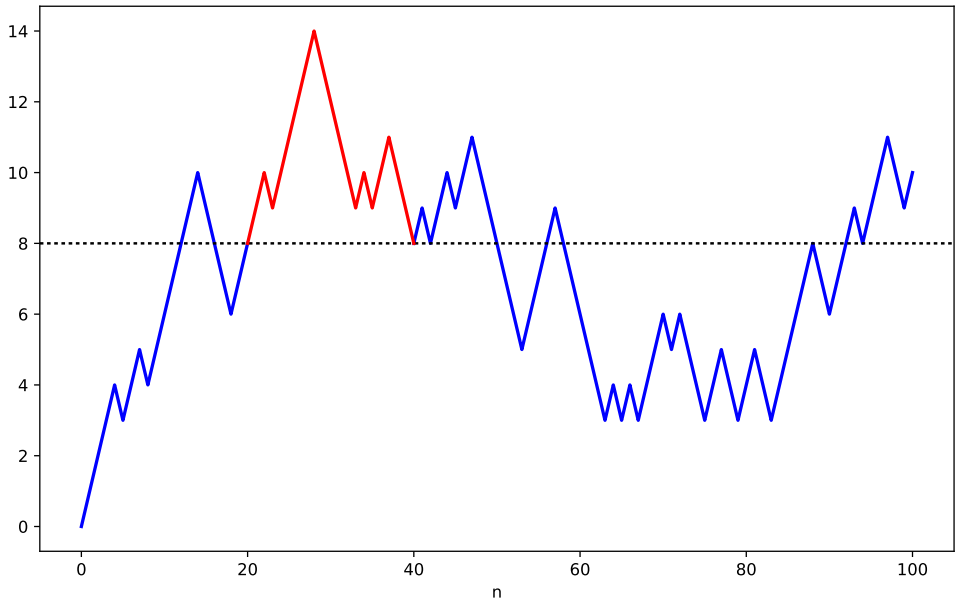

FIGURE 2.7 – Exemple de trajectoire d'une chaîne de Markov à valeurs dans $\mathbb{Z}$ où une excursion issue de $y = 8$ est tracée en rouge.

2.2.4 Mesures de probabilité invariantes

Soit $(X_n)_{n\in\mathbb{N}}$ une chaîne de Markov homogène d'espace d'états E dénombrable et de noyau de transition P. Pour tout $n \in \mathbb{N}$, si μ_n est la loi de la variable X_n, nous avons vu que la loi de X_{n+1} s'obtient par

$$\forall y \in E,\ \mu_{n+1}(y) = \sum_{x\in E} P(x,y)\mu_n(x).$$

Les mesures de probabilités invariantes sous cette action du noyau de transition joueront par la suite un rôle clef dans le comportement asymptotique des chaînes de Markov comme nous l'avons vu avec l'exemple de l'échantillonneur par tranche.

Définition 2.15. Une mesure de probabilité μ sur E est dite *invariante* pour la chaîne de Markov $(X_n)_{n\in\mathbb{N}}$ de noyau de transition P si

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

Si la loi initiale de X_0 est une probabilité invariante μ alors, pour tout $n \in \mathbb{N}$, la loi de X_n est aussi $\mu_n = \mu$. Autrement dit, une mesure de probabilité invariante correspond à un état stationnaire de la chaîne de Markov.

EXEMPLE 2.16 (Variables indépendantes). Si $(X_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires indépendantes et identiquement distribuées selon une loi μ sur E, alors $(X_n)_{n\in\mathbb{N}}$ est une chaîne de Markov et μ est l'unique mesure de probabilité invariante. ■

EXEMPLE 2.17 (Chaîne à deux états). Un noyau de transition P sur l'espace à deux états $E = \{0,1\}$ s'écrit

$$P(0,1) = 1 - P(0,0) = p_0 \quad \text{et} \quad P(1,0) = 1 - P(1,1) = p_1$$

où $p_0, p_1 \in [0,1]$. Vu comme une matrice, ce noyau de transition correspond à

$$P = \begin{pmatrix} 1-p_0 & p_0 \\ p_1 & 1-p_1 \end{pmatrix}.$$

Une mesure de probabilité μ sur E est invariante pour ce noyau de transition si et seulement si $p_0\mu(0) = p_1\mu(1)$. Lorsque $p_0 + p_1 > 0$, l'unique probabilité invariante est donnée par

$$\mu(0) = \frac{p_1}{p_0+p_1} \quad \text{et} \quad \mu(1) = \frac{p_0}{p_0+p_1}.$$

Si $p_0 = p_1 = 0$, la chaîne est constante et toutes les mesures de probabilité sont invariantes. ■

EXEMPLE 2.18 (Absence de probabilité invariante). Pour le processus de Bernoulli de paramètre $p \in [0,1]$ de l'exemple 2.7, une mesure de probabilité μ sur $\mathbb{N}$ est invariante si et seulement si $\mu(0) = (1-p)\mu(0)$ et $\mu(y) = p\mu(y-1) + (1-p)\mu(y)$, pour tout $y > 0$. Ainsi, si $p > 0$, il n'existe aucune mesure de probabilité invariante et, si $p = 0$, la chaîne est constante et toutes les mesures de probabilité sont invariantes. ■

Théorème 2.19. *Pour toute chaîne de Markov homogène sur un espace d'états E fini, il existe une mesure de probabilité invariante.*

Démonstration. Notons $|E|$ le cardinal fini de E et considérons une mesure de probabilité ν sur E. Nous pouvons définir une suite $(\nu_n)_{n\geqslant 1}$ de mesures de probabilité par

$$\forall n \geqslant 1,\ \forall y \in E,\ \nu_n(y) = \frac{1}{n}\sum_{k=1}^{n}\sum_{x\in E} P^k(x,y)\nu(x).$$

Les vecteurs $(\nu_n(y))_{y\in E}$ prennent leurs valeurs dans l'ensemble compact $[0,1]^{|E|}$. Il existe donc une suite extraite $(\nu_{n_k})_{k\geqslant 1}$ qui converge vers une mesure μ sur E,

$$\forall y \in E,\ \lim_{k\to+\infty} \nu_{n_k}(y) = \mu(y).$$

Comme E est fini, la mesure μ est une probabilité et nous avons, pour tout $n \geqslant 1$ et $z \in E$,

$$\begin{aligned}\sum_{y\in E} P(y,z)\nu_n(y) &= \frac{1}{n}\sum_{k=1}^{n}\sum_{x\in E} P^{k+1}(x,z)\nu(x)\\ &= \nu_n(z) + \frac{1}{n}\sum_{k=1}^{n}\sum_{x\in E}\left(P^{k+1}(x,z) - P^k(x,z)\right)\nu(x)\\ &= \nu_n(z) + \frac{1}{n}\sum_{x\in E}\left(P^{n+1}(x,z) - P(x,z)\right)\nu(x).\end{aligned}$$

Étant donné que $|P^{n+1}(x,z) - P(x,z)| \leqslant 1$, nous en déduisons que μ est invariante car

$$\forall z \in E,\ \sum_{y\in E} P(y,z)\mu(y) - \mu(z) = \lim_{k\to+\infty}\sum_{y\in E} P(y,z)\nu_{n_k}(y) - \nu_{n_k}(z) = 0.$$

□

Le résultat précédent est propre aux espaces d'états finis et ne peut pas être généralisé aux espaces dénombrables (voir l'exemple 2.18 pour un contre-exemple). Il faut également remarquer que ce théorème ne dit rien sur l'unicité de la mesure de probabilité invariante. Nous aurons besoin de plus de structure sur la chaîne de Markov pour établir de tels résultats.

2.2.5 Irréductibilité

Dans les exemples de la sous-section précédente, nous avons vu que l'existence de probabilités invariantes pour une chaîne de Markov dépend de son noyau de transition et que certains noyaux de transition n'admettent qu'une unique probabilité invariante. Pour discuter de ce lien entre noyaux de transition et mesures de probabilité invariantes, nous devons décrire comment un noyau de transition met en relation les différents états que peut prendre la chaîne de Markov. Cette structure relationnelle aura des conséquences sur le support d'une mesure invariante.

Définition 2.20. Soit P un noyau de transition sur un espace d'états E dénombrable. Pour tout $x, y \in E$, nous disons que

- $x \to y$: « x communique avec y » si il existe $n \in \mathbb{N}$ tel que $P^n(x,y) > 0$,
- $x \leftrightarrow y$: « x et y communiquent » si $x \to y$ et $y \to x$.

La relation de communication « $\leftrightarrow$ » est une relation d'équivalence qui partitionne l'espace d'états E en une union disjointe de classes d'équivalence appelées *classes irréductibles*.

Définition 2.21. Soit P un noyau de transition sur un espace d'états E dénombrable. Le noyau de transition P est dit *irréductible* si l'espace d'état E est l'unique classe irréductible pour la relation « $\leftrightarrow$ », *i.e.* si pour tous les états x et y, il existe $n \in \mathbb{N}$ tel que $P^n(x,y) > 0$. Par extension, une chaîne de Markov homogène de noyau de transition irréductible est appelée *chaîne de Markov irréductible*.

Toutes les chaînes de Markov ne sont pas irréductibles et il est possible de considérer des classes irréductibles de natures distinctes sur lesquelles la chaîne aura un comportement particulier. Par exemple, lorsqu'une classe irréductible est réduite à un singleton $\{x_0\}$, la chaîne de Markov ne peut pas s'en échapper et l'état x_0 est dit *absorbant*. Cependant, pour les applications qui nous intéressent dans le cadre de ce cours, nous limiterons notre étude aux chaînes irréductibles.

EXEMPLE 2.22 (Variables indépendantes). Si $(X_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires indépendantes et identiquement distribuées selon une loi μ sur un espace E dénombrable, nous savons que $(X_n)_{n\in\mathbb{N}}$ est une chaîne de Markov. De plus, si nous supposons que μ charge tous les états, alors le noyau de transition P est irréductible puisque, pour tout $x,y \in E$, $P(x,y) = \mu(y) > 0$. Dans cet exemple trivial où noyau de transition et loi initiale se confondent, nous pouvons caractériser l'unique probabilité invariante μ. En effet, pour tout état $x \in E$, la loi du temps de retour en x est géométrique de paramètre $\mu(x) > 0$,

$$\forall n \geqslant 1,\ \mathbb{P}^x(T_x^+ = n) = (1-\mu(x))^{n-1}\mu(x).$$

En particulier, nous avons $\mathbb{E}^x[T_x^+] = 1/\mu(x) < +\infty$ et donc

$$\forall x \in E,\ \mu(x) = \frac{1}{\mathbb{E}^x[T_x^+]}.$$

■

La caractérisation de la mesure invariante dans l'exemple précédent se généralise à toute chaîne de Markov d'espace d'états fini sous l'hypothèse d'irréductibilité. Ce théorème permet d'établir un lien entre la probabilité invariante et la fréquence à laquelle la chaîne visite les différents états.

Théorème 2.23. *Toute chaîne de Markov irréductible sur un espace d'états E fini vérifie*

$$\forall x,y \in E,\ \mathbb{E}^x[T_y^+] < +\infty$$

et l'unique mesure de probabilité invariante est donnée par

$$\forall x \in E,\ \mu(x) = \frac{1}{\mathbb{E}^x[T_x^+]}.$$

Démonstration. Admis. □

2.2.6 Réversibilité

Par définition, trouver une mesure de probabilité invariante pour un noyau de transition P sur un espace d'états E dénombrable revient à résoudre le système d'équations linéaires

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

dont le nombre d'inconnues est proportionnel au cardinal de E. Dans le cas fini, le théorème 2.23 caractérise μ mais ne donne pas une solution analytique. En pratique, l'espace d'états est souvent très grand (parfois infini) et il est difficile de déterminer μ explicitement. La notion de *réversibilité* représente une condition suffisante et facile à vérifier pour assurer l'existence d'une mesure de probabilité invariante.

Définition 2.24. Un noyau de transition P sur un espace d'états E dénombrable est dit *réversible* par rapport à une mesure de probabilité μ si il vérifie

$$\forall x,y \in E,\ \mu(x)P(x,y) = \mu(y)P(y,x).$$

Par extension, une chaîne de Markov homogène de noyau de transition réversible par rapport à μ est dite *réversible*.

La réversibilité caractérise un système à l'équilibre au sens où une trajectoire de la chaîne de Markov a la même probabilité que la trajectoire inverse. Autrement dit, si $(X_n)_{n\in\mathbb{N}}$ est une chaîne de Markov d'espace d'états E dénombrable et réversible par rapport à une probabilité μ, alors, pour tout $n \in \mathbb{N}$ et $x_0,\dots,x_n \in E$,

$$\mathbb{P}^\mu(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n) = \mathbb{P}^\mu(X_0 = x_n, X_1 = x_{n-1}, \dots, X_n = x_0).$$

En effet, dans ce cadre, la propriété de Markov faible donne

$$\begin{aligned}
\mathbb{P}^\mu(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n) &= \mu(x_0)P(x_0,x_1)P(x_1,x_2)\dots P(x_{n-1},x_n) \\
&= P(x_1,x_0)\mu(x_1)P(x_1,x_2)\dots P(x_{n-1},x_n) \\
&= \dots \\
&= P(x_1,x_0)P(x_2,x_1)\dots P(x_n,x_{n-1})\mu(x_n) \\
&= \mathbb{P}^\mu(X_0 = x_n, X_1 = x_{n-1}, \dots, X_n = x_0).
\end{aligned}$$

EXEMPLE 2.25 (Marche aléatoire sur le cercle). Soient un entier $k > 0$ et $p \in [0,1]$, nous considérons le noyau de transition suivant sur l'espace d'états $E = \mathbb{Z}/k\mathbb{Z}$,

$$\forall x,y \in \mathbb{Z}/k\mathbb{Z},\ P(x,y) = \begin{cases} p & \text{si } y = x+1 \pmod k, \\ 1-p & \text{si } y = x-1 \pmod k, \\ 0 & \text{sinon.} \end{cases}$$

Quelle que soit p, ce noyau de transition admet la mesure de probabilité uniforme $\mu(x) = 1/k$ comme unique probabilité invariante,

$$\forall y \in \mathbb{Z}/k\mathbb{Z},\ \sum_{x \in \mathbb{Z}/k\mathbb{Z}} P(x,y)\mu(x) = \frac{1}{k}(p+1-p) = \frac{1}{k} = \mu(y).$$

Remarquons que grâce au théorème 2.23, nous pouvons en déduire que $\mathbb{E}^x[T_x^+] = k$. La chaîne est réversible par rapport à μ si

$$\forall x \in \mathbb{Z}/k\mathbb{Z},\ \mu(x)P(x,x+1) = \mu(x+1)P(x+1,x),$$

c'est-à-dire si $p = 1-p$, ce qui n'est vrai que dans le cas de la marche aléatoire simple où $p = 1/2$ (voir l'exemple 2.6). En effet, si $p \neq 1/2$, la chaîne aura tendance à tourner dans le même sens et une trajectoire en sens inverse aura une probabilité moindre. Pour $p = 1/2$, la marche aléatoire est symétrique et les trajectoires dans un sens ou dans l'autre seront équiprobables. Cet exemple montre qu'il existe des probabilités invariantes pour lesquelles le noyau de transition n'est pas réversible. ■

Théorème 2.26. *Si un noyau de transition P sur un espace d'états E dénombrable est réversible par rapport à une mesure de probabilité μ, alors μ est une mesure de probabilité invariante.*

Démonstration. Soit $y \in E$, le fait que $P(y,\cdot)$ soit une mesure de probabilité et la réversibilité impliquent

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

Par conséquent, μ est une mesure de probabilité invariante. □

2.2.7 Espace d'états dénombrable

Lorsque l'espace d'états est fini, le théorème 2.19 assure l'existence d'une mesure de probabilité invariante et, sous l'hypothèse d'irréductibilité, le théorème 2.23 donne l'unicité d'une telle probabilité. Dans le cas d'un espace d'états dénombrable infini, l'existence d'une probabilité invariante n'est pas assurée (voir l'exemple 2.18) et l'irréductibilité n'est pas suffisante non plus (voir l'exemple ci-dessous). Il faut donc imposer plus de structure sur la façon dont la chaîne visite ses états pour assurer de telles propriétés quand l'espace est infini.

Exemple 2.27 (Chaîne irréductible sans probabilité invariante). Considérons l'espace d'états $E = \mathbb{N}$ et le noyau de transition P donné par

$$\forall k \in \mathbb{N},\ P(k,0) = p_k \quad \text{et} \quad P(k,k+1) = 1-p_k$$

où $0 < p_k < 1$. Quitte à passer par 0, il existe toujours une trajectoire de probabilité strictement positive qui relie deux entiers donnés et ce noyau de transition est donc irréductible. Si une probabilité μ sur $\mathbb{N}$ est invariante, elle doit vérifier, pour tout $k > 0$,

$$\mu(k) = \sum_{\ell\in\mathbb{N}} P(\ell,k)\mu(\ell) = (1-p_{k-1})\mu(k-1) = \cdots = \mu(0) \times \prod_{\ell=0}^{k-1}(1-p_\ell),$$

et $\mu(0) = \sum_{k\in\mathbb{N}} p_k\mu(k)$. Notons $R_0 = 1$ et, pour tout $k > 0$, $R_k = \prod_{\ell=0}^{k-1}(1-p_\ell)$, nous obtenons

$$\mu(0) = \mu(0) \times \sum_{k\in\mathbb{N}} p_k R_k = \mu(0) \times \sum_{k\in\mathbb{N}} (R_k - R_{k+1}).$$

La dernière somme se télescope et cette équation ne peut être satifsaite que si R_k converge vers 0 quand k tend vers l'infini. Dans le cas contraire, il n'y a pas de probabilité invariante pour cette chaîne irréductible. ■

Avec le théorème 2.23, nous avons vu que la fréquence à laquelle une chaîne de Markov $(X_n)_{n\in\mathbb{N}}$ d'espace d'états E fini visite un état $x \in E$ se mesure à l'aide de propriétés sur le temps de premier passage,

$$T_x^+ = \inf\{n \geqslant 1 \text{ tel que } X_n = x\}.$$

C'est également cette variable que nous considérerons dans le cas d'un espace d'états dénombrable où il convient de distinguer plusieurs cas.

Définition 2.28. Soit $(X_n)_{n\in\mathbb{N}}$ une chaîne de Markov sur un espace d'états E dénombrable. Un état $x \in E$ est appelé

- *transitoire* si $\mathbb{P}^x(T_x^+ < +\infty) < 1$,
- *récurrent* si $\mathbb{P}^x(T_x^+ < +\infty) = 1$.

Un état récurrent peut être

- *récurrent nul* si $\mathbb{E}^x[T_x^+] = +\infty$,
- *récurrent positif* si $\mathbb{E}^x[T_x^+] < +\infty$.

Dans le cas où tous les états $x \in E$ sont récurrents positifs, la chaîne est dite *récurrente positive*.

D'après le théorème 2.23, une chaîne de Markov irréductible sur un espace d'états fini est récurrente positive. Cette propriété est celle qui nous permettra d'étendre l'existence et l'unicité de la probabilité invariante à des chaînes de Markov d'espace d'états infini dénombrable. Avant d'énoncer ce résultat, il convient de discuter comment une trajectoire d'une chaîne de Markov $(X_n)_{n\in\mathbb{N}}$ d'espace d'états E dénombrable peut se décomposer en une suite d'excursions issues d'un état $x \in E$ donné. Pour cela, supposons que $X_0 = x$ et définissons les temps de retour successifs en x,

$$T_x^{(0)} = 0 \quad \text{et} \quad \forall k \geqslant 1,\ T_x^{(k)} = \inf\left\{n \geqslant T_x^{(k-1)} + 1 \text{ tel que } X_n = x\right\}.$$

Par la propriété de Markov forte, toutes les excursions $(X_n,\ T_x^{(k-1)} \leqslant n \leqslant T_x^{(k)})_{k\geqslant 1}$ sont indépendantes et suivent la même loi. Si l'état x est transitoire, il y a une probabilité strictement positive que la chaîne ne revienne pas en x et nous savons donc que la séquence d'excursions issues de x sera interrompue par cet absence de retour. Cette remarque est la base de la preuve que la chaîne de Markov ne visite presque sûrement d'un nombre fini de fois un état transitoire. Réciproquement, le nombre de visites d'un état récurrent sera presque sûrement infini et, pour tout $k \geqslant 1$, la longueur $T_x^{(k)} - T_x^{(k-1)}$ de la k-ème excursion sera presque sûrement finie. Cette longueur suit la même loi que le temps T_x^+ par la propriété de Markov forte et sa moyenne n'est finie que si l'état x est récurrent positif. Les états récurrents nuls sont donc des états visités un nombre infini de fois mais pour lesquels le temps entre deux visites consécutives est infini en moyenne. Ainsi, la fréquence $1/\mathbb{E}^x[T_x^+]$ de visite de l'état x qui donne la mesure de probabilité invariante de la chaîne lorsque l'espace E est fini ne chargera que les états récurrents positifs dans le cas d'un espace d'états dénombrable.

Théorème 2.29. *Pour toute chaîne de Markov irréductible sur un espace d'états E dénombrable, les propositions suivantes sont équivalentes :*

- *La chaîne est récurrente positive.*
- *Il existe une unique mesure de probabilité invariante donnée par*

$$\forall x \in E,\ \mu(x) = \frac{1}{\mathbb{E}^x[T_x^+]}.$$

Démonstration. Admis. □

2.2.8 Convergence

Comme nous l'avons évoqué dans l'exemple de l'échantillonneur par tranche, ce sont des propriétés de convergence qui motivent cette introduction aux chaînes de Markov. En particulier, la référence à la méthode de Monte Carlo dans le nom des méthodes MCMC est liée à la généralisation suivante de la loi des grands nombres pour les chaînes de Markov.

Théorème 2.30 (Théorème ergodique)**.** *Soit $(X_n)_{n\in\mathbb{N}}$ une chaîne de Markov irréductible et récurrente positive sur un espace d'états E dénombrable de loi initiale ν et de noyau de transition P. Notons μ son unique probabilité invariante et considérons deux fonctions $f : E \to \mathbb{R}$ et $g : E \times E \to \mathbb{R}$ telles que*

$$\sum_{x\in E} |f(x)|\mu(x) < +\infty \quad \text{et} \quad \sum_{x,y\in E} |g(x,y)|P(x,y)\mu(x) < +\infty.$$

Alors, les moyennes le long des trajectoires convergent presque sûrement,

$$\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \xrightarrow[n\to+\infty]{p.s.} \sum_{x\in E} f(x)\mu(x)$$

et

$$\frac{1}{n}\sum_{k=1}^{n} g(X_{k-1},X_k) \xrightarrow[n\to+\infty]{p.s.} \sum_{x,y\in E} g(x,y)P(x,y)\mu(x).$$

Démonstration. Admis. □

Étant donné un état $x \in E$, une conséquence immédiate du théorème ergodique appliqué à la fonction indicatice $f(y) = \mathbf{1}_{y=x}$ permet de faire le lien entre la probabilité invariante μ et la fréquence des visites de l'état x,

$$\frac{1}{n}\sum_{k=0}^{n-1} \mathbf{1}_{X_k=x} \xrightarrow[n\to+\infty]{p.s.} \mu(x) = \frac{1}{\mathbb{E}^x[T_x^+]}. \tag{2.5}$$

Cette convergence reste vraie pour les chaînes de Markov admettant des états $x \in E$ récurrents nuls ou transitoires pour lesquels $\mathbb{E}^x[T_x^+] = +\infty$. Dans ce cas, la fréquence empirique des visites de x tend presque sûrement vers 0.

Ces résultats sont intéressants pour obtenir des valeurs approchées d'intégrales par rapport à μ qui sont difficiles à calculer explicitement comme avec les méthodes présentées dans la

section 1.3. Cependant, cela n'est pas suffisant dans la perspective de résoudre le problème de la simulation pour la loi μ pour lequel nous avons besoin de résultats de convergence sur la loi de la variable X_n. Afin d'illustrer cela, reprenons l'exemple 2.17 de la chaîne à deux états. Si $p_0, p_1 \in]0,1[$, nous avons vu que l'unique probabilité invariante est donnée par

$$\mu(0) = \frac{p_1}{p_0+p_1} \quad \text{et} \quad \mu(1) = \frac{p_0}{p_0+p_1}.$$

En considérant le noyau de transition P comme une matrice, celui-ci se diagonalise facilement,

$$P = \begin{pmatrix} 1-p_0 & p_0 \\ p_1 & 1-p_1 \end{pmatrix} = \begin{pmatrix} 1 & -\mu(1) \\ 1 & \mu(0) \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1-p_0-p_1 \end{pmatrix} \begin{pmatrix} \mu(0) & \mu(1) \\ -1 & 1 \end{pmatrix}.$$

Les puissances de P en découlent et nous pouvons en déduire le résultat de convergence suivant,

$$P^n = \begin{pmatrix} 1 & -\mu(1) \\ 1 & \mu(0) \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & (1-p_0-p_1)^n \end{pmatrix} \begin{pmatrix} \mu(0) & \mu(1) \\ -1 & 1 \end{pmatrix} \xrightarrow[n\to+\infty]{} \begin{pmatrix} \mu(0) & \mu(1) \\ \mu(0) & \mu(1) \end{pmatrix}.$$

Autrement dit, lorsque n tend vers l'infini, les probabilités de transition convergent exponentiellement vite vers la mesure de probabilité invariante,

$$\forall x, y \in \{0,1\}, \ \lim_{n\to+\infty} \mathbb{P}^x(X_n = y) = \mu(y).$$

Comme pour le théorème ergodique, la limite ne dépend pas de la loi de l'état initial. En temps long, la chaîne « oublie » la position dont elle est issue et, dans le vocabulaire des chaînes de Markov, nous parlons de *perte de mémoire*.

La convergence en temps long vers la probabilité invariante μ est une propriété générale des chaînes de Markov $(X_n)_{n\in\mathbb{N}}$ récurrentes positives sur un espace d'états E dénombrable. Afin d'établir des résultats généraux et de quantifier la vitesse de cette convergence, il est cependant nécessaire de se prémunir de certains cas pathologiques et de vérifier quelques conditions. En effet, en considérant l'espérance de (2.5), nous savons que

$$\forall x, y \in E, \ \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{P}^x(X_k = y) \xrightarrow[n\to+\infty]{} \mu(y)$$

mais cela ne suffit pas à établir que

$$\forall x, y \in E, \ \lim_{n\to+\infty} \mathbb{P}^x(X_n = y) = \mu(y). \tag{2.6}$$

Pour s'en convaincre, il suffit de reprendre d'exemple 2.6 de la marche aléatoire simple sur le cercle avec un nombre pair $k = 2\ell$ de sites. Si cette marche aléatoire part de 0, la loi de X_{2n} ne chargera que les sites pairs et celle de X_{2n+1} ne chargera que les sites impairs. Pour pouvoir obtenir des résultats de convergence tels que (2.6), il faut donc au minimum éviter de considérer des noyaux de transition qui présentent de telles propriétés periodiques.

Définition 2.31. Une chaîne de Markov $(X_n)_{n\in\mathbb{N}}$ irréductible de noyau de transition P sur un espace d'état E dénombrable est dite *apériodique* si, pour tout $x, y \in E$, il existe $n(x,y) \in \mathbb{N}$ tel que

$$\forall n \geqslant n(x,y), \ \mathbb{P}^x(X_n = y) = P^n(x,y) > 0.$$

Cette définition permet d'éviter les problèmes mentionnés ci-dessus. De plus, si une chaîne de Markov irréductible de noyau P n'est pas apériodique, il est simple de se ramener à une chaîne de Markov apériodique en considérant une variante dont le noyau de transition est donné par

$$Q = \frac{I+P}{2}$$

où I est le noyau de transition trivial (2.4). À chaque temps, cette variante reste sur place avec une probabilité $1/2$ ou saute à un nouvel état selon P avec probabilité $1/2$. Il est évident qu'une probabilité invariante pour P est également invariante pour Q et ces deux chaînes admettent le même comportement asymptotique. En pratique, l'irréductibilité d'une chaîne de Markov permet d'établir facilement son apériodicité par la proposition suivante.

Proposition 2.32. *Soit une chaîne de Markov $(X_n)_{n\in\mathbb{N}}$ irréductible de noyau de transition P sur un espace d'états E dénombrable. Si un état $x \in E$ est apériodique, c'est-à-dire qu'il existe un entier n_x tel que*

$$\forall n \geqslant n_x,\ \mathbb{P}^x(X_n = x) = P^n(x,x) > 0,$$

alors la chaîne est apériodique.

Démonstration. Considérons deux états $y, z \in E$. Par irréductibilité, il existe deux entiers r et s tels que $P^r(y,x) > 0$ et $P^s(x,z) > 0$. Nous pouvons en déduire que, pour tout $n \geqslant n_x$,

$$P^{r+n+s}(y,z) \geqslant P^r(y,x)P^n(x,x)P^s(x,z) > 0.$$

La chaîne est donc apériodique. □

Une conséquence importante de l'apériodicité est donnée par le résultat suivant qui assure la convergence vers la probabilité invariante.

Théorème 2.33. *Si $(X_n)_{n\in\mathbb{N}}$ est une chaîne de Markov irréductible et apériodique qui admet une unique mesure de probabilité invariante μ sur un espace d'états E dénombrable, alors, pour toute loi initiale ν, la loi de X_n converge vers μ quand n tend vers l'infini,*

$$\forall x \in E,\ \lim_{n\to+\infty} \mathbb{P}^\nu(X_n = x) = \mu(x).$$

Démonstration. Admis. □

L'interprétation de ce résultat est que, pour tout état $x \in E$, la valeur de la probabilité invariante $\mu(x)$ peut être approchée par $\mathbb{P}^\nu(X_n = x)$ quand n est grand quelle que soit la loi initiale ν. Cette probabilité $\mathbb{P}^\nu(X_n = x)$ peut elle-même être approchée par la loi des grands nombres en considérant plusieurs réalisations indépendantes $(X_n^{(1)})_{n\in\mathbb{N}}, \ldots, (X_n^{(K)})_{n\in\mathbb{N}}$ de la chaîne de Markov,

$$\frac{1}{K}\sum_{k=1}^{K} \mathbf{1}_{X_n^{(k)}=x} \xrightarrow[n\to+\infty]{p.s.} \mathbb{P}^\nu(X_n = x).$$

De fait, il existe deux approches pour estimer la mesure de probabilité invariante μ :

- pour un état x donné, prendre la moyenne empirique (2.5) du nombre de visites de x le long d'une trajectoire par le théorème ergodique,
- pour un temps n fixé, construire l'histogramme des valeurs prises au temps n par plusieurs réalisations indépendantes de la chaîne.

Grâce à l'apériodicité, les probabilités de transition convergent donc vers la mesure de probabilité invariante. La manipulation d'une chaîne de Markov de probabilité invariante μ sera d'autant plus facile que cette convergence sera rapide. Une vitesse de convergence correspond à une borne supérieure de l'écart entre la loi de la n-ème itération de la chaîne et la mesure invariante. Il existe diverses façons de quantifier cet écart et, dans le cadre de ce cours, nous serons intéressés par la *distance en variation totale* introduite dans (2.1).

Définition 2.34. Étant données deux mesures de probabilité μ et ν sur un espace mesurable E, la *distance en variation totale* entre μ et ν est définie par

$$\|\mu-\nu\|_{VT} = \sup_{A\subset E} |\mu(A)-\nu(A)|$$

où le supremum est pris sur tous les événements mesurables. Si l'espace E est dénombrable, cette définition est équivalente à

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

La convergence d'une suite $(\mu_n)_{n\in\mathbb{N}}$ de mesures de probabilité sur un espace E au sens de la distance en variation totale implique la convergence en loi. En effet, si il existe une mesure de probabilité μ sur E telle que

$$\lim_{n\to+\infty} \|\mu_n-\mu\|_{VT} = 0$$

alors une suite $(X_n)_{n\in\mathbb{N}}$ de variables aléatoires telle que, pour tout $n\in\mathbb{N}$, $X_n\sim\mu_n$ converge en loi vers μ quand n tend vers l'infini,

$$X_n \xrightarrow[n\to+\infty]{\mathscr{L}} \mu.$$

Du point de vue des chaînes de Markov, la distance en variation totale est également intéressante car elle satisfait un principe de contraction.

Proposition 2.35 (Contraction markovienne)**.** *Pour tout noyau de transition P et toute mesure de probabilité μ sur un espace E dénombrable, nous pouvons considérer la mesure de probabilité définie par*

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

Alors, pour toutes mesures de probabilité μ et ν sur E, la distance en variation totale vérifie

$$\|\mu P-\nu P\|_{VT} \leqslant \|\mu-\nu\|_{VT}.$$

Démonstration. Ce résultat découle de l'inégalité triangulaire et du fait que $P(x,\cdot)$ soit une mesure de probabilité sur E car

$$\begin{aligned}\|\mu P - \nu P\|_{VT} &= \frac{1}{2}\sum_{y\in E}|\mu P(y) - \nu P(y)| \\ &= \frac{1}{2}\sum_{y\in E}\left|\sum_{x\in E}P(x,y)(\mu(x)-\nu(x))\right| \\ &\leqslant \frac{1}{2}\sum_{x\in E}\sum_{y\in E}P(x,y)|\mu(x)-\nu(x))| \\ &= \|\mu-\nu\|_{VT}.\end{aligned}$$

□

En particulier, si μ est une probabilité invariante pour un noyau de transition P sur un espace E dénombrable, alors, pour toute mesure de probabilité ν sur E, la contraction markovienne implique

$$\forall n\in\mathbb{N},\ \|\nu P^n - \mu\|_{VT} \leqslant \|\nu-\mu\|_{VT}.$$

Si $(X_n)_{n\in\mathbb{N}}$ est une chaîne de Markov de noyau de transition P et de loi initiale ν sur E, nous savons que

$$\forall n\in\mathbb{N},\ \forall y\in E,\ \mathbb{P}^\nu(X_n=y) = \nu P^n(y).$$

Ainsi, la loi de la chaîne de Markov ne peut pas s'éloigner d'une probabilité invariante au sens de la distance en variation totale. En quantifiant cet écart entre la loi de X_n et la mesure invariante, nous pouvons donc établir une vitesse de convergence comme nous le verrons dans la suite. Afin d'établir des vitesses de convergence générales, il est nécessaire de faire des hypothèses additionnelles sur la chaîne de Markov telles que la *condition de Doeblin*.

Définition 2.36. Une chaîne de Markov $(X_n)_{n\in\mathbb{N}}$ de noyau de transition P sur un espace d'états E dénombrable vérifie la *condition de Doeblin* si il existe $r\geqslant 1$, $\delta>0$ et une mesure de probabilité ν sur E tels que

$$\forall x,y\in E,\ P^r(x,y)\geqslant \delta\nu(y).$$

Sous cette condition, il est possible de montrer la convergence exponentielle des probabilités de transition d'une chaîne de Markov irréductible et apériodique vers la mesure de probabilité invariante.

Théorème 2.37. *Soit une chaîne de Markov $(X_n)_{n\in\mathbb{N}}$ irréductible et apériodique sur un espace d'états E dénombrable. Si cette chaîne vérifie la condition de Doeblin, alors elle admet une mesure de probabilité invariante μ vers laquelle les probabilités de transition convergent exponentiellement vite quelle que soit la loi initiale ν,*

$$\forall n\in\mathbb{N},\ \|\nu P^n-\mu\|_{VT} = \frac{1}{2}\sum_{x\in E}|\mathbb{P}^\nu(X_n=x)-\mu(x)| \leqslant (1-\delta)^{\lfloor n/r\rfloor}$$

où $r\geqslant 1$ et $\delta>0$ sont donnés par la définition 2.36 et $\lfloor\cdot\rfloor$ désigne la partie entière.

Démonstration. Admis. □

Dans le cas d'un espace E fini, une chaîne de Markov irréductible et apériodique de noyau de transition P vérifie toujours la condition de Doeblin. En effet, par hypothèse, pour tout $x,y \in E$, il existe $r(x,y) \in \mathbb{N}$ tel que $P^n(x,y) > 0$ pour tout $n \geqslant r(x,y)$. En posant

$$r = \max_{x,y\in E} r(x,y) \geqslant 1,$$

nous avons $P^r(x,y) > 0$ pour tout $x,y \in E$. Il suffit alors de choisir

$$\delta = \sum_{y\in E} \min_{x\in E} P^r(x,y) > 0 \quad \text{et} \quad \forall y \in E,\ \nu(y) = \delta^{-1} \min_{x\in E} P^r(x,y).$$

La convergence vers la mesure de probabilité invariante a donc toujours lieu à vitesse exponentielle dans ce cas.

EXEMPLE 2.38 (Vitesses de convergence pour la marche aléatoire sur le cercle). Reprenons l'exemple 2.6 de la marche aléatoire simple sur le cercle $E = \mathbb{Z}/k\mathbb{Z}$ pour $k > 1$. La variante apériodique de cette chaîne de Markov est donnée par le noyau de transition suivant,

$$\forall x,y \in E,\ P(x,y) = \begin{cases} 1/2 & \text{si } x = y, \\ 1/4 & \text{si } |x-y| = 1 \pmod k, \\ 0 & \text{sinon.} \end{cases}$$

Pour $r = \lfloor k/2 \rfloor \geqslant 1$, nous avons $P^r(x,y) > 0$ pour tout $x,y \in E$ et la chaîne vérifie la condition de Doeblin avec $\delta = k4^{-\lfloor k/2 \rfloor} \in]0,1[$. Pour $n \in \mathbb{N}$, le théorème 2.37 donne la borne supérieure suivante sur la distance en variation totale,

$$(1-\delta)^{\lfloor n/r \rfloor} = \left(1 - k4^{-\lfloor k/2 \rfloor}\right)^{\lfloor n/\lfloor k/2 \rfloor \rfloor} \leqslant \exp\left(-\left\lfloor \frac{n}{\lfloor k/2 \rfloor} \right\rfloor k4^{-\lfloor k/2 \rfloor}\right)$$

où nous avons utilisé le fait que $\ln(1-x) \leqslant -x$ pour tout $x \in]0,1[$. Les propriétés de la partie entière donnent alors

$$(1-\delta)^{\lfloor n/r \rfloor} \leqslant \exp\left(-\frac{2n-k}{2^k}\right).$$

Cette vitesse n'est pas pertinente en pratique car elle n'a d'intérêt que lorsque n est au moins de l'ordre de 2^k qui peut être très grand si le nombre de sites est important. Une alternative consiste à considérer la variante apériodique de la marche aléatoire simple sur E qui ne reste pas en place avec probabilité $1/2$ comme la précédente mais saute plutôt uniformément sur tous les sites. Cette chaîne admet encore la loi uniforme sur E comme unique mesure de probabilité invariante et son noyau de transition est donné par

$$\forall x,y \in E,\ P'(x,y) = \frac{1}{2k} + \frac{1}{4}\mathbf{1}_{|x-y|=1}.$$

Pour satisfaire la condition de Doeblin, il suffit maintenant de prendre $r' = 1$ et $\delta' = 1/2$. Le théorème 2.37 donne la borne supérieure

$$(1-\delta')^{\lfloor n/r' \rfloor} = \frac{1}{2^n}$$

qui est une vitesse de convergence bien plus rapide et qui ne dépend pas du nombre de sites k. Cette modification des probabilités de transition qui permet d'accélérer la vitesse de convergence correspond à celle introduite dans l'algorithme *PageRank* développé à partir de 1995 par les cofondateurs de Google, Larry Page et Sergey Brin. ■

2.3 Algorithme de Metropolis-Hastings

Dans le cadre des recherches sur les armes nucléaires à Los Alamos à la fin des années 1940, Nicholas Metropolis et Stanislaw Ulam ont été amenés à travailler avec la *distribution de Boltzmann* qui est une mesure de probabilité utilisée en physique statistique pour décrire l'état d'un système en fonction de son énergie et de sa température. L'ensemble E des états accessibles par le système étant supposé fini, cette distribution attribue à un état $x \in E$ une probabilité proportionnelle à $\exp(-\varepsilon_x/(k_B T))$ où ε_x désigne l'énergie de l'état x, k_B la *constante de Boltzmann* et T la température. La constante de normalisation de cette distribution est donc donnée par

$$Z(T) = \sum_{x \in E} \exp\left(-\frac{\varepsilon_x}{k_B T}\right).$$

Bien que l'ensemble E soit fini, son cardinal est généralement très grand et cette somme peut être coûteuse (voire impossible) à calculer. Cette difficulté a conduit Metropolis et Ulam à développer un algorithme spécifique pour résoudre le problème de la simulation pour la distribution de Boltzmann qui sera correctement décrit en 1953 par Metropolis *et al.* dans [**?**]. Cette méthode sera généralisée en 1970 par le statisticien Wilfred K. Hastings, dans une perspective bien plus pacifiste, pour simuler une mesure de probabilité μ quelconque dans [**?**]. L'algorithme porte le nom de *Metropolis-Hastings* et il s'agit d'une méthode MCMC relativement universelle puisque elle n'impose aucune hypothèse sur la loi cible μ.

FIGURE 2.8 – Nicholas Metropolis (1915-1999)

2.3.1 Construction de l'algorithme

Nous considérons donc une mesure de probabilité μ sur un espace E fini pour laquelle nous souhaitons résoudre le problème de la simulation. Quitte à nous restreindre au support de

FIGURE 2.9 – Wilfred Keith Hastings (1930-2016)

μ, nous pouvons supposer que tous les états sont chargés, *i.e.* $\mu(x) > 0$ pour tout $x \in E$. Le principe général de l'algorithme de Metropolis-Hastings consiste à construire une chaîne de Markov $(X_n)_{n\in\mathbb{N}}$ irréductible et apériodique sur E qui admette μ comme unique probabilité invariante. Pour une telle chaîne, le théorème 2.33 assure que la suite des états convergent en loi vers μ,

$$X_n \xrightarrow[n\to+\infty]{\mathscr{L}} \mu. \tag{2.7}$$

Cette construction se fait par la définition d'un noyau de transition P réversible par rapport à μ. Le théorème 2.26 implique que μ est invariante pour ce noyau et, comme l'espace d'état E est fini, l'unicité de la probabilité invariante découle de l'irréductibilité.

L'idée de Metropolis pour cela est d'introduire un noyau de transition auxiliaire Q irréductible et idéalement facile à simuler tel que

$$\forall x,y \in E,\ Q(x,y) > 0 \Rightarrow Q(y,x) > 0. \tag{2.8}$$

Soient $x,y \in E$, la probabilité de transition de l'état x à l'état y pour la chaîne de Markov auxiliaire de loi initiale μ et de noyau de transition Q est donnée par $\mu(x)Q(x,y)$ et, réciproquement, la transition de y à x admet une probabilité $\mu(y)Q(y,x)$. Le rapport entre ces probabilités quantifie quelle transition est la plus vraisemblable en étant issu d'une réalisation de la loi μ. Ainsi, une itération de l'algorithme consiste à privilégier d'autant plus le passage de l'état x à l'état y que sa vraisemblance est élevée et nous sommes amenés à définir les *probabilités d'acceptation* à l'aide de ces rapports,

$$\forall x,y \in E,\ \alpha(x,y) = \min\left\{1, \frac{\mu(y)Q(y,x)}{\mu(x)Q(x,y)}\right\}.$$

La quotient dans ces probabilités est important car il permet de ne considérer qu'une fonction $\tilde{\mu} : E \to \mathbb{R}_+^*$ proportionnelle à μ pour appliquer l'algorithme. En effet, si il existe $C > 0$ tel

que $\tilde{\mu}(x) = C\mu(x)$ pour tout $x \in E$, alors

$$\forall x,y \in E,\ \frac{\mu(y)Q(y,x)}{\mu(x)Q(x,y)} = \frac{\tilde{\mu}(y)Q(y,x)}{\tilde{\mu}(x)Q(x,y)}.$$

De plus, dans le cas particulier d'un noyau de transition Q symétrique, *i.e.* pour tout $x,y \in E$, $Q(x,y) = Q(y,x)$, les probabilités d'acceptation se simplifient et il ne reste que le rapport $\mu(y)/\mu(x)$ pour quantifier de façon plus intuitive combien un état candidat $y \in E$ est plus probable que l'état courant $x \in E$ pour la loi cible. Il s'agit de la forme initiale de l'algorithme tel qu'il fut proposé par Metropolis.

ALGORITHME 2.39 – Metropolis-Hastings

Initialisation :
- une fonction $\tilde{\mu}$ proportionnelle à la loi cible μ sur E fini
- un noyau de transition Q irréductible sur E vérifiant (2.8)
- un état initial $X_0 \in E$

À l'étape $n \geqslant 1$:
- Tirer Y selon la loi $Q(X_{n-1},\cdot)$
- Calculer $\alpha = \alpha(X_{n-1},Y)$
- Tirer U selon la loi $\mathscr{U}([0,1])$
- Règle de rejet :
  - Si $U \leqslant \alpha$, le candidat est accepté : $X_n = Y$
  - Si $U > \alpha$, le candidat est rejeté : $X_n = X_{n-1}$

Retourner les valeurs $X_0, X_1, X_2, \ldots$

Pour tout $n \in \mathbb{N}$, l'état X_{n+1} ne dépend que de X_n et cet algorithme produit bien une chaîne de Markov d'espace d'états E fini. Cette chaîne peut se comprendre comme une marche aléatoire sur E qui reste en place quand le candidat Y est rejeté et qui a tendance à se diriger vers les états de forte probabilité sinon. Autrement dit, cette marche aléatoire est encouragée à visiter les zones les plus probables de E tout en continuant à explorer le reste des états avec une probabilité moindre. Ce compromis entre l'exploration de l'espace et la concentration autour des états les plus probables est un aspect central de l'algorithme de Metropolis-Hastings et le principe s'étend au-delà du problème de la simulation, *e.g.* pour les problèmes d'optimisation présentés au chapitre 3, l'exploration est une façon d'échapper aux pièges des extrema locaux.

2.3.2 Noyau de Metropolis-Hastings

L'expression du noyau de transition de la chaîne de Markov produite par l'algorithme de Metropolis-Hastings se déduit de la définition,

$$\forall x,y \in E,\ P(x,y) = \begin{cases} Q(x,y)\alpha(x,y) = \min\left\{Q(x,y), \dfrac{\mu(y)Q(y,x)}{\mu(x)}\right\} & \text{si } x \neq y, \\ 1 - \displaystyle\sum_{z \neq x} Q(x,z)\alpha(x,z) & \text{sinon.} \end{cases}$$

Le noyau de transition P est appelé *noyau de Metropolis-Hastings* et l'objectif de cette sous-section est de présenter certaines de ses propriétés.

Proposition 2.40. *Le noyau de Metropolis-Hastings est réversible par rapport à la mesure de probabilité μ.*

Démonstration. Soient deux états $x, y \in E$ distincts, le résultat découle de la définition,

$$\mu(x)P(x,y) = \min\{\mu(x)Q(x,y), \mu(y)Q(y,x)\} = \mu(y)P(y,x).$$

□

La réversibilité donnée par la proposition précédente reste vraie dans un cadre plus général pour toute fonction $h : \mathbb{R}_+^* \to [0,1]$ telle que

$$\forall u > 0,\ h(u) = uh(1/u) \tag{2.9}$$

dès lors que la probabilité d'acceptation de l'algorithme de Metropolis-Hastings est définie par

$$\forall x, y \in E,\ \alpha(x,y) = h\left(\frac{\mu(y)Q(y,x)}{\mu(x)Q(x,y)}\right).$$

En effet, nous pouvons écrire, pour tous les états distincts $x, y \in E$,

$$\begin{aligned}
\mu(x)P(x,y) &= \mu(x)Q(x,y)h\left(\frac{\mu(y)Q(y,x)}{\mu(x)Q(x,y)}\right) \\
&= \mu(x)Q(x,y) \times \frac{\mu(y)Q(y,x)}{\mu(x)Q(x,y)} h\left(\frac{\mu(x)Q(x,y)}{\mu(y)Q(y,x)}\right) \\
&= \mu(y)Q(y,x)h\left(\frac{\mu(x)Q(x,y)}{\mu(y)Q(y,x)}\right) \\
&= \mu(y)P(y,x).
\end{aligned}$$

Cette remarque est utile en pratique pour définir des variantes de l'algorithme de Metropolis-Hastings où la forme des probabilités d'acceptation est plus régulière que notre choix initial $u \mapsto \min\{1,u\}$. Un exemple de fonction h infiniment dérivable sur $\mathbb{R}_+^*$ vérifiant (2.9) est donné par $h(u) = u/(1+u)$ et le noyau de transition P associé est appelé *noyau de Barker*

Par réversibilité, la loi cible μ est invariante pour le noyau de Metropolis-Hastings. Comme mentionné dans la construction de l'algorithme, la convergence des probabilités de transition vers cette probabilité μ se déduit alors de la proposition suivante.

Proposition 2.41. *Si il existe deux états $x_0, y_0 \in E$ distincts tels que*

$$Q(x_0,y_0) > 0 \quad \textit{et} \quad \alpha(x_0,y_0) < 1$$

alors le noyau de Metropolis-Hastings est irréductible et apériodique.

Démonstration. Considérons deux états $x, y \in E$ distincts, l'irréductibilité de Q implique l'existence d'un entier $m \in \mathbb{N}$ et d'états distincts intermédiaires $z_0, \ldots, z_m \in E$ tels que $z_0 = x$ et $z_m = y$ avec $Q(z_{k-1}, z_k) > 0$ pour tout $k \in \{1, \ldots, m\}$. La propriété (2.8) donne $Q(z_k, z_{k-1}) > 0$

et, puisque la probabilité μ est supposée charger tous les états, nous avons $\mu(z_k) > 0$. Il s'en suit que $\alpha(z_{k-1}, z_k) > 0$ et donc

$$P^m(x,y) \geqslant \prod_{k=1}^{m} P(z_{k-1}, z_k) = \prod_{k=1}^{m} Q(z_{k-1}, z_k)\alpha(z_{k-1}, z_k) > 0.$$

Le noyau de Metropolis-Hastings est donc irréductible. De plus, par hypothèse, nous savons que la probabilité de rester en x_0 vérifie

$$\begin{aligned} P(x_0,x_0) &= 1 - \sum_{z \neq x_0} Q(x_0,z)\alpha(x_0,z) \\ &= 1 - Q(x_0,y_0)\alpha(x_0,y_0) - \sum_{z \neq x_0,y_0} Q(x_0,z)\alpha(x_0,z) \\ &> 1 - Q(x_0,y_0) - \sum_{z \neq x_0,y_0} Q(x_0,z) \\ &= 1 - \sum_{z \neq x_0} Q(x_0,z) = Q(x_0,x_0) \geqslant 0. \end{aligned}$$

L'inégalité stricte permet de déduire $P(x_0,x_0) > 0$ et donc l'apériodicité de l'état x_0. Par irréductibilité, nous obtenons que le noyau de Metropolis-Hastings est apériodique par la proposition 2.32. □

2.3.3 Vitesse de convergence

La convergence des probabilités de transition vers la probabilité invariante μ est particulièrement intéressante en pratique si sa vitesse est rapide. En effet, pour pouvoir considérer les itérations de l'algorithme de Metropolis-Hastings comme des réalisations approximative de μ par (2.7), il convient d'attendre que la chaîne ait suffisamment évolué pour avoir « oublié » ses conditions initiales et s'être approchée de μ. Nous parlons de *temps de mélange* pour désigner ce phénomène et celui-ci sera d'autant plus court que la convergence a lieu rapidement. Les premières itérations de la chaîne devront être ignorées et le temps de calcul pour les produire est donc nécessairement perdu. Le fait de sacrifier ces valeurs avant que la chaîne soit assez mélangée est souvent désigner comme une période de *burn-in*.

L'espace d'états E étant supposé de cardinal $K = |E|$ fini, nous pouvons l'identifier à $\{1,\dots,K\}$ et le noyau de Metropolis-Hastings P se manipule alors comme une matrice carrée de taille $K \times K$. Pour établir la vitesse de convergence de l'algorithme de Metropolis-Hastings, nous allons étudier le spectre de cette matrice, *i.e.* l'ensemble de ses valeurs propres. La motivation pour cela réside dans le fait que si nous notons $\mu = (\mu_1,\dots,\mu_K)' \in \mathbb{R}_+^*$ le vecteur donné par les probabilités $\mu_k = \mu(k) > 0$ pour $k \in E$, alors la propriété d'invariance se traduit par

$$\mu' P = \mu',$$

c'est-à-dire que μ est un vecteur propre à gauche de P associé à la valeur propre 1. Pour tout $k,\ell \in E$, la réversibilité de P par rapport à μ se traduit par $\mu_k P_{k\ell} = \mu_\ell P_{\ell k}$ et, en introduisant la matrice diagonale D donnée par $D_{kk} = \sqrt{\mu_k}$ pour tout $k \in E$, nous obtenons que la matrice DPD^{-1} est symétrique car

$$\forall k,\ell \in E,\ \left(DPD^{-1}\right)_{k\ell} = \frac{\sqrt{\mu_k}}{\sqrt{\mu_\ell}} P_{k\ell} = \frac{\sqrt{\mu_k}}{\sqrt{\mu_\ell}} \times \frac{\mu_\ell}{\mu_k} P_{\ell k} = \frac{\sqrt{\mu_\ell}}{\sqrt{\mu_k}} P_{\ell k} = \left(DPD^{-1}\right)_{\ell k}.$$

Lemme 2.42. *Les matrices $I+DPD^{-1}$ et $I-DPD^{-1}$ sont positives.*

Démonstration. Par définition d'un noyau de transition, nous savons que

$$\forall k \in E,\ \sum_{\ell \in E} P_{k\ell} = 1.$$

Soit $x \in \mathbb{R}^K$, cette propriété et la réversibilité de P par rapport à μ donne

$$\begin{aligned} x'(I+DPD^{-1})x &= \frac{1}{2}\sum_{k,\ell\in E} P_{k\ell}x_k^2 + P_{\ell k}x_\ell^2 + 2\frac{\sqrt{\mu_k}}{\sqrt{\mu_\ell}}P_{k\ell}x_k x_\ell \\ &= \frac{1}{2}\sum_{k,\ell\in E} \mu_k P_{k\ell}\frac{x_k^2}{\mu_k} + \mu_\ell P_{\ell k}\frac{x_\ell^2}{\mu_\ell} + 2\mu_k P_{k\ell}\frac{x_k}{\sqrt{\mu_k}}\frac{x_\ell}{\sqrt{\mu_\ell}} \\ &= \frac{1}{2}\sum_{k,\ell\in E} \mu_k P_{k\ell}\frac{x_k^2}{\mu_k} + \mu_k P_{k\ell}\frac{x_\ell^2}{\mu_\ell} + 2\mu_k P_{k\ell}\frac{x_k}{\sqrt{\mu_k}}\frac{x_\ell}{\sqrt{\mu_\ell}} \\ &= \frac{1}{2}\sum_{k,\ell\in E} \mu_k P_{k\ell}\left(\frac{x_k}{\sqrt{\mu_k}} + \frac{x_\ell}{\sqrt{\mu_\ell}}\right)^2 \geqslant 0. \end{aligned}$$

Un calcul similaire conduit à la positivité de $I-DPD^{-1}$,

$$x'(I-DPD^{-1})x = \frac{1}{2}\sum_{k,\ell\in E} \mu_k P_{k\ell}\left(\frac{x_k}{\sqrt{\mu_k}} - \frac{x_\ell}{\sqrt{\mu_\ell}}\right)^2 \geqslant 0.$$

□

Nous pouvons maintenant énoncer le théorème qui permet de faire le lien entre les propriétés d'un noyau de transition réversible sur un espace d'états fini et la matrice associée.

Théorème 2.43. *Soit P une matrice donnée par un noyau de transition réversible pour une probabilité μ sur un espace d'états E fini telle que $\mu(k) > 0$ pour tout $k \in E$. Les propositions suivantes sont toutes satisfaites :*

(i) Le spectre de P est inclus dans l'intervalle $[-1,1]$.

(ii) Le noyau de transition est irréductible si et seulement si 1 est une valeur propre simple de P.

(iii) Si il existe un état $k \in E$ tel que $P_{kk} > 0$ et que le noyau est irréductible, alors -1 n'est pas une valeur propre de P.

Démonstration. Remarquons tout d'abord que P et DPD^{-1} ont les même valeurs propres car si il existe $\lambda \in \mathbb{R}$ et $v \in \mathbb{R}^K$ non nul tels que $Pv = \lambda v$, alors

$$DPD^{-1}(Dv) = DPv = \lambda(Dv). \tag{2.10}$$

Si $\lambda \in \mathbb{R}$ est une valeur propre de DPD^{-1}, alors $1+\lambda$ est valeur propre de $I+DPD^{-1}$ et $1-\lambda$ est valeur propre de $I-DPD^{-1}$. Par le lemme précédent, nous en déduisons que $1+\lambda \geqslant 0$ et $1-\lambda \geqslant 0$, *i.e.* $\lambda \in [-1,1]$.

Si un vecteur non nul $v \in \mathbb{R}^K$ est tel que $DPD^{-1}v = v$, alors $(I - DPD^{-1})v = 0$ et nous en déduisons comme dans la preuve du lemme précédent que

$$v'(I - DPD^{-1})v = \frac{1}{2} \sum_{k,\ell \in E} \mu_k P_{k\ell} \left(\frac{v_k}{\sqrt{\mu_k}} - \frac{v_\ell}{\sqrt{\mu_\ell}} \right)^2 = 0.$$

L'espace propre associé à la valeur propre 1 est donc l'espace des vecteurs v tels que

$$\forall k, \ell \in E,\ P_{k\ell} > 0 \Rightarrow \frac{v_k}{\sqrt{\mu_k}} = \frac{v_\ell}{\sqrt{\mu_\ell}}$$

Si il existe un chemin de probabilité strictement positive entre deux états $k, \ell \in E$ disctincts, alors toutes les composantes associées aux états visités par ce chemin vérifient cette égalité. Autrement dit, la dimension de l'espace propre associé à la valeur propre 1 (qui n'est pas vide car μ lui appartient) est donnée par le nombre de classes irréductibles du noyau de transition.

Supposons que -1 soit une valeur propre de P, alors il existe un vecteur non nul $v \in \mathbb{R}^K$ tel que $DPD^{-1}v = -v$. Nous obtenons comme ci-dessus que

$$\forall k, \ell \in E,\ P_{k\ell} > 0 \Rightarrow \frac{v_k}{\sqrt{\mu_k}} = -\frac{v_\ell}{\sqrt{\mu_\ell}}.$$

En particulier, si il existe un état $k \in E$ tel que $P_{kk} > 0$, alors $v_k = 0$ et tous les états $\ell \in E$ tels que $P_{k\ell} > 0$ vérifient $v_\ell = 0$. Par irréductibilité, cela implique que le vecteur v est nul, ce qui est une contradiction. Donc pour tout $k \in E$, l'élément diagonal P_{kk} est nul et la dernière proposition est vérifiée. □

Pour le noyau de Metropolis-Hastings, l'existence de deux états distincts $k_0, \ell_0 \in E$ tels que $Q(k_0, \ell_0) > 0$ et $\alpha(k_0, \ell_0) < 1$ implique l'irréductibilité du noyau et son apériodicité par la proposition 2.41. Dans la preuve de ce résultat, nous avons montré que $P_{k_0 k_0} > 0$ et donc que -1 ne peut pas être une valeur propre par le théorème précédent. Nous supposons maintenant que nous sommes dans ce cadre, nous savons alors que :

- la valeur propre 1 est simple et l'espace propre associé est l'espace $\mathbb{R}\sqrt{\mu}$ engendré par l'unique probabilité invariante μ,
- toutes les autres valeurs propres appartiennent à $]-1, 1[$.

Ainsi, nous pouvons ordonner les K valeurs propres de la matrice P par ordre décroissant avec leur multiplicité,

$$\lambda_1 = 1 > \lambda_2 \geqslant \cdots \geqslant \lambda_K > -1.$$

Les matrices DP^nD^{-1} où $n \geqslant 1$ sont symétriques et commutent, elles sont diagonalisables dans une même base orthonormée de vecteurs propres. Notons $\phi_1, \ldots, \phi_K \in \mathbb{R}^K$ ces vecteurs propres associés respectivement aux valeurs propres $\lambda_1, \ldots, \lambda_K$. Pour le cas particulier du premier vecteur propre, nous considérons évidemment

$$\phi_1 = (\sqrt{\mu_1}, \ldots, \sqrt{\mu_K})'.$$

Si $u, v \in \mathbb{R}^K$, alors nous pouvons développer le produit à l'aide des coordonnées dans la base des ϕ_k,

$$u'DP^nD^{-1}v = \sum_{k=1}^{K} (u'\phi_k)(v'\phi_k)\lambda_k^n.$$

Soient $k, \ell \in E$, prenons $u = \mathbf{1}_k/\sqrt{\mu_k}$ et $v = \mathbf{1}_\ell\sqrt{\mu_\ell}$ où $\mathbf{1}_k$ désigne le vecteur de $\mathbb{R}^K$ dont la k-ème composante vaut 1 et les autres sont nulles. À l'aide l'égalité précédente, nous obtenons

$$(P^n)_{k\ell} = \mu_\ell + \frac{\sqrt{\mu_\ell}}{\sqrt{\mu_k}} \sum_{j=2}^{K} \phi_{j,k}\phi_{j,\ell}\lambda_j^n \tag{2.11}$$

et, puisque $|\lambda_j| < 1$ pour tout $j \geqslant 2$, ceci établit la convergence des probabilités de transition vers la loi cible μ à vitesse exponentielle lorsque le nombre d'itérations n tend vers l'infini. Cette formule n'est pas facile à utiliser en pratique et le théorème suivant fournit une forme plus simple.

Théorème 2.44. *Soit P une matrice donnée par un noyau de transition irréductible, apériodique et réversible pour une probabilité μ sur un espace d'états E fini telle que $\mu(k) > 0$ pour tout $k \in E$. Notons α la plus grande valeur propre de P strictement inférieure à 1 et considérons un entier $n \geqslant 1$, alors*

$$\forall k \in E,\ \sum_{\ell\in E} \mu_\ell \left(\frac{(P^n)_{k\ell}}{\mu_\ell} - 1\right)^2 \leqslant \frac{\alpha^{2n}}{\mu_k}.$$

Démonstration. Soient $k, \ell \in E$, en réarrangeant les termes de (2.11), nous avons

$$\sqrt{\mu_\ell}\left(\frac{(P^n)_{k\ell}}{\mu_\ell} - 1\right) = \frac{1}{\sqrt{\mu_k}} \sum_{j=2}^{K} \phi_{j,k}\phi_{j,\ell}\lambda_j^n.$$

Élevons cette égalité au carré et faisons la somme sur tous les états, l'orthonormalité des vecteurs $\phi_1, \ldots, \phi_K$ donne

$$\begin{aligned}
\sum_{\ell\in E} \mu_\ell \left(\frac{(P^n)_{k\ell}}{\mu_\ell} - 1\right)^2 &= \frac{1}{\mu_k} \sum_{\ell\in E} \left(\sum_{j=2}^{K} \phi_{j,k}\phi_{j,\ell}\lambda_j^n\right)^2 \\
&= \frac{1}{\mu_k} \sum_{j=2}^{K}\sum_{j'=2}^{K} \phi_{j,k}\phi_{j',k}\lambda_j^n\lambda_{j'}^n \sum_{\ell\in E} \phi_{j,\ell}\phi_{j',\ell} \\
&= \frac{1}{\mu_k} \sum_{j=2}^{K} \phi_{j,k}^2\lambda_j^{2n}.
\end{aligned}$$

Par hypothèse, $\lambda_j^{2n} \leqslant \alpha^{2n}$ pour tout $j \in \{2, \ldots, K\}$ et nous obtenons

$$\sum_{\ell\in E} \mu_\ell \left(\frac{(P^n)_{k\ell}}{\mu_\ell} - 1\right)^2 \leqslant \frac{\alpha^{2n}}{\mu_k} \sum_{j=2}^{K} \phi_{j,k}^2 \leqslant \frac{\alpha^{2n}}{\mu_k}$$

où la dernière inégalité se déduit de l'orthogonalité de la matrice de passage de la base canonique dans la base orthonormée des ϕ_j car, par transposition, les lignes de cette matrice sont aussi de norme unitaire. □

Ce théorème donne la vitesse de convergence exponentielle des probabilités de transition $P^n(k,\cdot)$ vers la loi μ au sens de la *distance du χ^2* définie pour toutes probabilités ν et μ sur E

avec μ strictement positive par

$$\chi^2(\nu,\mu) = \sum_{\ell\in E} \mu_\ell \left(\frac{\nu_\ell}{\mu_\ell} - 1\right)^2 = \sum_{\ell\in E} \frac{(\nu_\ell - \mu_\ell)^2}{\mu_\ell}.$$

Cette distance n'en est pas une au sens mathématique car elle n'est pas symétrique. Cependant, elle permet d'établir une borne supérieure sur la distance en variation totale par l'inégalité de Cauchy-Schwarz,

$$\|\nu - \mu\|_{VT} = \frac{1}{2}\sum_{\ell\in E} |\nu_\ell - \mu_\ell| = \frac{1}{2}\sum_{\ell\in E} \frac{|\nu_\ell - \mu_\ell|}{\sqrt{\mu_\ell}}\sqrt{\mu_\ell} \leqslant \frac{1}{2}\sqrt{\chi^2(\nu,\mu)}. \tag{2.12}$$

Ainsi, nous pouvons établir une vitesse de convergence vers la loi cible μ au sens de la distance en variation totale pour l'algorithme de Metropolis-Hastings,

$$\forall k \in E,\ \|P^n(k,\cdot) - \mu\|_{VT} \leqslant \frac{\alpha^n}{2\sqrt{\mu(k)}}. \tag{2.13}$$

La valeur $1-\alpha$ est appelée le *trou spectral* et la vitesse de convergence établie ci-dessus sera d'autant meilleure que cette quantité est grande. Le choix du noyau auxiliaire Q est crucial en pratique car c'est essentiellement de lui que dépende les valeurs propres de la matrice P et donc la valeur du trou spectral.

2.3.4 Application : modèle d'Ising

La distribution de Boltzmann introduite au début de cette section est un cas particulier des *mesures de Gibbs* dont nous reparlerons dans la sous-section 3.1.1. De telles mesures de probabilité apparaissent dans de nombreux problèmes en théorie des probabilités ou en physique statistique et sont dénommées ainsi en référence au physico-chimiste américain Josiah Willard Gibbs. Elles sont définies sur un espace E fini et prennent la forme suivante

$$\forall x \in E,\ \mu(x) = \frac{1}{Z_T}\exp(-V(x)/T) \quad \text{avec} \quad Z_T = \sum_{y\in E}\exp(-V(y)/T) \tag{2.14}$$

où $T > 0$ est un paramètre dit de *température* et la fonction $V : E \to \mathbb{R}$ est appelée le *fonction d'énergie*. En pratique, pour calculer la probabilité $\mu(x)$ pour un état $x \in E$, la constante de normalisation Z_T doit être évaluée et cela revient à calculer les valeurs de la fonction d'énergie V pour tous les états d'un ensemble E dont le cardinal est souvent trop important. L'algorithme de Metropolis-Hastings fournit une méthode de simulation de la loi μ sans avoir à évaluer Z_T.

Le *modèle d'Ising* est un cadre théorique simple en physique statistique pour décrire les interactions locales entre des particules à deux états appelées des *spins*. À chaque site k d'un réseau Λ, nous associons un spin $s_k \in \{+1,-1\}$ et nous notons $S = \{s_k\}_{k\in\Lambda}$ une configuration de l'ensemble des spins. Pour une configuration S, nous pouvons définir une énergie à l'aide des interactions locales entre chaque spin et ses plus proches voisins,

$$V(S) = \sum_{k\sim\ell} s_k s_\ell$$

où la somme porte sur toutes les paires de sites $(k,\ell) \in \Lambda^2$ à distance 1 dans le réseau. Pour tenir compte des fluctuations thermiques, un paramètre de température $T > 0$ est introduit et la fonction d'énergie V donne la mesure de Gibbs suivante pour la configuration S,

$$\mu(S) = \frac{1}{Z_T}\exp(-V(S)/T).$$

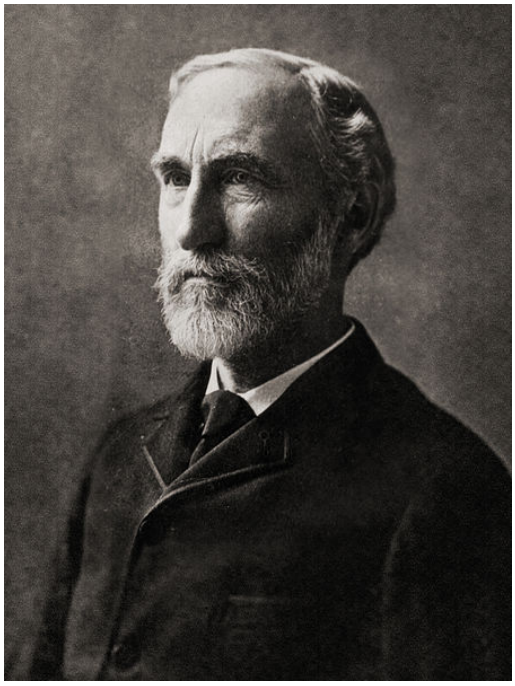

FIGURE 2.10 – Josiah Willard Gibbs (1839-1903)

Pour une grille régulière bidimensionnel $\Lambda = \{1,\dots,L\} \times \{1,\dots,L\}$ avec $L = 128$, l'espace des configurations $E = \{+1,-1\}^{\Lambda}$ est de cardinal $2^{128\times 128} \simeq 10^{4932}$. Il est donc impossible d'énumérer toutes des les configurations pour calculer la constante Z_T et nous allons utiliser l'algorithme de Metropolis-Hastings pour produire des réalisations de la loi μ. Pour cela, nous introduisons la transformation élémentaire du site $k \in \Lambda$ d'une configuration $S \in E$ qui consiste à inverser le signe du spin s_k,

$$\forall \ell \in \Lambda,\ S_\ell^{(k)} = \begin{cases} -s_k & \text{si } \ell = k, \\ s_\ell & \text{si } \ell \neq k. \end{cases}$$

Le noyau de transition auxiliaire Q que nous considérons correspond au tirage d'un site k uniformément dans Λ et à cette transformation élémentaire,

$$\forall S \in E,\ \forall k \in \Lambda,\ Q(S,S^{(k)}) = \frac{1}{|\Lambda|}.$$

Ce noyau vérifie (2.8) et il est symétrique. Il n'apparaît donc plus dans les probabilités d'acceptation comme nous l'avons vu et nous obtenons

$$\begin{aligned} \forall S \in E,\ \forall k \in \Lambda,\ \alpha(S,S^{(k)}) &= \min\left\{1, \exp\left(-\frac{V(S^{(k)}) - V(S)}{T}\right)\right\} \\ &= \min\left\{1, \exp\left(-\frac{2s_k}{T}\sum_{\ell\sim k} s_\ell\right)\right\}. \end{aligned}$$

La figure 2.11 donne un exemple de réalisation de μ par cette méthode pour $L = 128$ et $T = 1.0$.

ALGORITHME 2.45 – Modèle d'Ising bidimensionnel régulier

Initialisation :

- la taille $L > 0$ de la grille et la température $T > 0$
- une configuration initial $S_0 \in E$

À l'étape $n \geqslant 1$:

Tirer k uniformément dans Λ

Calculer $\alpha = \alpha(S, S^{(k)})$

Tirer U selon la loi $\mathscr{U}([0,1])$

Règle de rejet :

- Si $U \leqslant \alpha$, la configuration est acceptée : $S_n = S_{n-1}^{(k)}$
- Si $U > \alpha$, la configuration est rejetée : $S_n = S_{n-1}$

Retourner les configurations $S_0, S_1, S_2, \ldots$

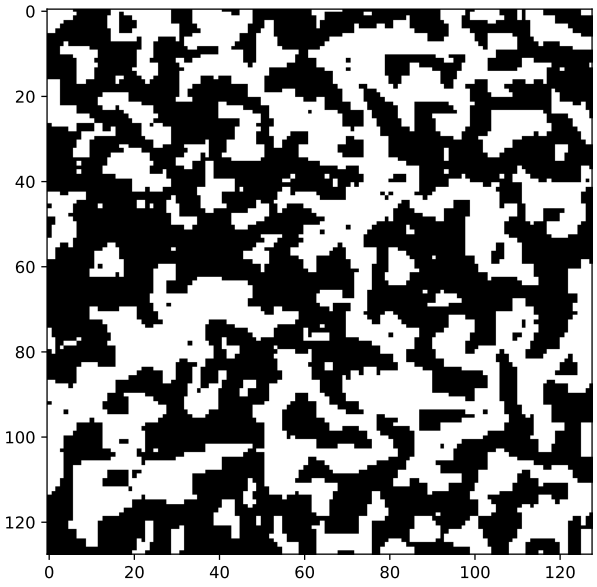

FIGURE 2.11 – Une réalisation du modèle d'Ising obtenue par l'algorithme de Metropolis-Hastings sur une grille régulière de taille $L = 128$ pour une température $T = 1.0$.

2.4 Échantillonneur de Gibbs

Le principe de l'échantillonneur par tranche introduit dans la section 2.1 consiste à produire une marche aléatoire sur un ensemble $B \subset \mathbb{R}^2$ donné par une densité de probabilité en simulant les composantes une par une. Cette démarche se généralise en dimension plus grande avec l'*échantillonneur de Gibbs*. Contrairement à l'algorithme de Metropolis-Hastings qui porte le nom de ses inventeurs, l'échantillonneur de Gibbs n'a pas été introduit par Josiah Willard Gibbs mais par Stuart et Donald Geman en 1984 dans [**?**]. Le nom de cette méthode est une référence aux champs aléatoires de Gibbs utilisés en analyse d'image.

Une difficulté potentielle de l'implantation 2.39 de l'algorithme de Metropolis-Hastings est l'étape de simulation du candidat selon la loi donnée par le noyau de transition auxiliaire. En particulier, lorsque la dimension de l'espace d'états est grande, le fléau de la dimension peut compliquer grandement notre capacité à obtenir des réalisations selon une loi arbitraire. D'autres problèmes techniques peuvent également se poser en grande dimension et des méthodes de simulation composante par composante telles que l'échantillonneur de Gibbs permettent souvent de les contourner.

2.4.1 Le cas bidimensionnel

La méthode du conditionnement (voir le théorème 1.15) permet d'introduire le principe général de l'échantillonneur de Gibbs dans le cas bidimensionnel. Considérons un vecteur aléatoire (X_1, X_2) à valeurs dans un espace $E_1 \times E_2$ dénombrable. La loi μ sur E_1 de X_1 s'exprime comme le mélange des lois conditionnelles $\mu_1(\cdot \mid x_2)$ de X_1 sachant $X_2 = x_2$, pour tout $x_2 \in E_2$, par la loi ν de X_2 sur E_2,

$$\forall x_1 \in E, \mu(x_1) = \sum_{x_2 \in E_2} \mu_1(x_1 \mid x_2)\nu(x_2). \tag{2.15}$$

Comme nous l'avons vu, pour simuler des réalisations de la loi μ, il suffit de pouvoir simuler des réalisations de la loi ν et de toutes les lois conditionnelles $\mu_1(\cdot \mid x_2)$ pour $x_2 \in E_2$. En pratique, la loi de mélange ν n'est pas toujours disponible et la méthode du conditionnement est donc d'un intérêt limité. Cependant, dans les problèmes de modélisation, les lois conditionnelles $\mu_1(\cdot \mid x_2)$ de X_1 sachant $X_2 = x_2 \in E_2$ et $\mu_2(\cdot \mid x_1)$ de X_2 sachant $X_1 = x_1 \in E_1$ sont souvent simples à simuler et l'échantillonneur de Gibbs tire partie de cela en alternant les réalisations de l'une et de l'autre pour éviter la simulation de la loi ν.

ALGORITHME 2.46 – Échantillonneur de Gibbs bidimensionnel

Initialisation :
- une loi cible sur un espace produit $E_1 \times E_2$
- les lois conditionnelles $\mu_1(\cdot \mid x_2)$ et $\mu_2(\cdot \mid x_1)$
- un état initial $\mathbf{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}) \in E_1 \times E_2$

À l'étape $n \geqslant 1$:

Tirer $x_1^{(n)}$ selon la loi $\mu_1(\cdot \mid x_2^{(n-1)})$

Tirer $x_2^{(n)}$ selon la loi $\mu_2(\cdot \mid x_1^{(n)})$

Poser $\mathbf{x}^{(n)} = (x_1^{(n)}, x_2^{(n)})$

Retourner les vecteurs $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)} \ldots$

Comme dans le cas de l'échantillonneur par tranche, nous pouvons remarquer que les valeurs générées ne sont pas rejetées. De plus, il convient d'insister sur le fait que la réalisation $x_2^{(n)}$ est tirée conditionnellement à $x_1^{(n)}$ et non pas à $x_1^{(n-1)}$. Enfin, cet algorithme produit une chaîne de Markov homogène sur $E_1 \times E_2$ et chaque composante est également une chaîne de Markov homogène. En effet, prenons la première composante, la suite $(x_1^{(n)})_{n \in \mathbb{N}}$ ainsi produite est évidemment une chaîne de Markov par construction et son noyau de transition est donné par

$$\forall x_1, x_1' \in E_1,\ P(x_1', x_1) = \sum_{x_2 \in E_2} \mu_1(x_1 \mid x_2)\mu_2(x_2 \mid x_1').$$

Sous des hypothèses adéquates, les probabilités de transition de cette chaîne convergeront vers la loi μ car celle-ci est invariante. En effet, l'identité (2.15) donne

$$\begin{aligned}\forall x_1 \in E_1,\ \sum_{x_1' \in E_1} P(x_1', x_1)\mu(x_1') &= \sum_{x_1' \in E_1} \left(\sum_{x_2 \in E_2} \mu_1(x_1 \mid x_2)\mu_2(x_2 \mid x_1') \right) \mu(x_1') \\ &= \sum_{x_2 \in E_2} \mu_1(x_1 \mid x_2) \left(\sum_{x_1' \in E_1} \mu_2(x_2 \mid x_1')\mu(x_1') \right) \\ &= \sum_{x_2 \in E_2} \mu_1(x_1 \mid x_2)\nu(x_2) = \mu(x_1).\end{aligned}$$

L'échantillonneur de Gibbs bidimensionnel est donc une méthode MCMC pour la loi μ. Les propriétés de cet algorithme seront discutées dans la sous-section suivante qui introduit une version plus générale.

2.4.2 Un algorithme par balayage

Soit un entier $d \geqslant 2$, nous nous plaçons dans un cadre multidimensionnel en considérant une mesure de probabilité μ sur un espace d'états $E = E_1 \times \cdots \times E_d$ dénombrable. Pour chaque $k \in \{1, \ldots, d\}$, il est possible de définir la loi conditionnelle $\mu_k(\cdot \mid x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_d)$ de la k-ème composante sachant les autres. Si le problème de la simulation est résolu pour toutes ces lois, alors l'échantillonneur de Gibbs consiste à mettre à jour successivement chaque composante en simulant la loi conditionnelle associée. Nous parlons de *balayage* pour désigner le parcours des différentes composantes et leurs changements de valeurs. Lorsque ce balayage est fait de la première composante à la dernière, l'algorithme est dit à *balayage systématique*.

ALGORITHME 2.47 – Échantillonneur de Gibbs par balayage systématique

Initialisation :

- une loi cible μ sur un espace produit $E = E_1 \times \cdots \times E_d$
- les lois conditionnelles $\mu_k(\cdot \mid x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_d)$ pour tout $k \in \{1, \ldots d\}$
- un état initial $\mathbf{x}^{(0)} = (x_1^{(0)}, \ldots, x_d^{(0)}) \in E$

À l'étape $n \geqslant 1$:

Tirer $x_1^{(n)}$ selon la loi $\mu_1(\cdot \mid x_2^{(n-1)}, \ldots, x_d^{(n-1)})$
Tirer $x_2^{(n)}$ selon la loi $\mu_2(\cdot \mid x_1^{(n)}, x_3^{(n-1)}, \ldots, x_d^{(n-1)})$
$\vdots$
Tirer $x_d^{(n)}$ selon la loi $\mu_d(\cdot \mid x_1^{(n)}, \ldots, x_{d-1}^{(n)})$
Poser $\mathbf{x}^{(n)} = (x_1^{(n)}, \ldots, x_d^{(n)})$

Retourner les vecteurs $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$

Cette implantation de l'échantillonneur de Gibbs est la plus classique et la suite produite $(\mathbf{x}^{(n)})_{n \in \mathbb{N}}$ est une chaîne de Markov sur E. Il présente cependant le défaut de fournir des marginales $(x_k^{(n)})_{n \in \mathbb{N}}$, $k \in \{1, \ldots, d\}$, qui ne sont pas des chaînes de Markov comme c'était le cas pour l'échantillonneur de Gibbs bidimensionnel. En pratique, il s'agit d'une propriété souhaitable pour approcher la loi marginale d'une partie des composantes comme dans le cas de variables latentes, par exemple. Le problème avec le balayage systématique est que toutes les composantes sont mises à jour à chaque itération, ce qui induit une dépendance entre

elles et ne permet plus aux marginales de demeurer markoviennes. Une alternative consiste à ne modifier qu'une seule composante tirée uniformément à chaque itération de l'algorithme. Nous parlons alors de *balayage aléatoire*.

ALGORITHME 2.48 – Échantillonneur de Gibbs par balayage aléatoire

Initialisation :

- une loi cible μ sur un espace produit $E = E_1 \times \cdots \times E_d$
- les lois conditionnelles $\mu_k(\cdot \mid x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_d)$ pour tout $k \in \{1, \ldots d\}$
- un état initial $\mathbf{x}^{(0)} = (x_1^{(0)}, \ldots, x_d^{(0)}) \in E$

À l'étape $n \geqslant 1$:

Tirer k uniformément parmi $\{1, \ldots, d\}$

Tirer $x_k^{(n)}$ selon la loi $\mu_k(\cdot \mid x_1^{(n-1)}, \ldots, x_{k-1}^{(n-1)}, x_{k+1}^{(n-1)}, \ldots, x_d^{(n-1)})$

Poser $\mathbf{x}^{(n)} = (x_1^{(n-1)}, \ldots, x_{k-1}^{(n-1)}, x_k^{(n)}, x_{k+1}^{(n-1)}, \ldots, x_d^{(n-1)})$

Retourner les vecteurs $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$

Du point de vue mathématique, cette version à balayage aléatoire est également plus simple à étudier car la chaîne de Markov $(\mathbf{x}^{(n)})_{n \in \mathbb{N}}$ est réversible par rapport à la loi cible μ qui est donc invariante. Pour voir cela, nous introduisons les relations $\mathbf{x} \sim_k \mathbf{y}$, pour tout $k \in \{1, \ldots, d\}$, définies par

$$\mathbf{x} \sim_k \mathbf{y} \iff x_k = y_k$$

où $\mathbf{x} = (x_1, \ldots, x_d), \mathbf{y} = (y_1, \ldots, y_d) \in E$. Ainsi, le noyau de transition de l'échantillonneur de Gibbs par balayage aléatoire s'écrit

$$\forall \mathbf{x}, \mathbf{y} \in E,\ P(\mathbf{x}, \mathbf{y}) = \begin{cases} \dfrac{1}{d} \times \dfrac{\mu(\mathbf{y})}{\displaystyle\sum_{\mathbf{z} \sim_k \mathbf{x}} \mu(\mathbf{z})} & \text{si } \mathbf{x} \sim_k \mathbf{y}, \\ 0 & \text{sinon.} \end{cases}$$

La réversibilité s'obtient en remarquant que, pour tout $\mathbf{x}, \mathbf{y} \in E$ tels que $\mathbf{x} \sim_k \mathbf{y}$, nous avons par transitivité

$$\mu(\mathbf{x}) P(\mathbf{x}, \mathbf{y}) = \frac{1}{d} \times \frac{\mu(\mathbf{x})\mu(\mathbf{y})}{\displaystyle\sum_{\mathbf{z} \sim_k \mathbf{x}} \mu(\mathbf{z})} = \frac{1}{d} \times \frac{\mu(\mathbf{x})\mu(\mathbf{y})}{\displaystyle\sum_{\mathbf{z} \sim_k \mathbf{y}} \mu(\mathbf{z})} = \mu(\mathbf{y}) P(\mathbf{y}, \mathbf{x}).$$

L'échantillonneur de Gibbs par balayage aléatoire est donc bien une méthode MCMC pour la mesure de probabilité μ. Sous des hypothèses convenables (*e.g.* espace d'états fini), les probabilités de transition convergent exponentiellement vite vers la probabilité invariante et la chaîne de Markov produite pourra être utilisée pour produire des réalisations approchées de la loi μ ou pour calculer des approximations d'intégrales contre μ par le théorème ergodique.

Comme l'algorithme de Metropolis-Hastings, l'échantillonneur de Gibbs se généralise à des espaces d'états continus en faisant intervenir la densité f de la loi μ sur $\mathbb{R}^d$ ainsi que les densités conditionnelles f_k des lois $\mu_k(\cdot \mid x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_d)$, pour tout $k \in \{1, \ldots, d\}$,

$$\begin{aligned} \forall t \in \mathbb{R},\ f_k(t \mid x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_d) &= \frac{f(x_1, \ldots, x_{k-1}, t, x_{k+1}, \ldots, x_d)}{\int_{\mathbb{R}} f(x_1, \ldots, x_{k-1}, x, x_{k+1}, \ldots, x_d) dx} \\ &\propto f(x_1, \ldots, x_{k-1}, t, x_{k+1}, \ldots, x_d). \end{aligned}$$

Cette densité s'exprime donc facilement à l'aide de la densité cible à une constante multiplicative près. Des réalisations des lois conditionnelles sur $\mathbb{R}$ pourront être obtenues par la méthode du rejet ou par Metropolis-Hastings, par exemple. Les chaînes de Markov à valeurs dans des espaces d'états continus dépassent le cadre de ce cours mais un certains nombre de propriétés se traduisent de façon similaire au cas dénombrable sous de bonnes hypothèses. En particulier, il est possible de prouver dans de nombreuses situations que les probabilités de transition convergent encore exponentiellement vite vers la loi cible pour l'échantillonneur de Gibbs par balayage aléatoire implanté comme ci-dessous.

ALGORITHME 2.49 – Échantillonneur de Gibbs pour une loi continue

Initialisation :
- une loi cible de densité f sur $\mathbb{R}^d$
- les densités conditionnelles $f_k(\cdot \mid x_1,\dots,x_{k-1},x_{k+1},\dots,x_d)$ pour tout $k \in \{1,\dots d\}$
- un état initial $\mathbf{x}^{(0)} = (x_1^{(0)},\dots,x_d^{(0)}) \in \mathbb{R}^d$

À l'étape $n \geqslant 1$:

Tirer k uniformément parmi $\{1,\dots,d\}$

Tirer $x_k^{(n)}$ selon la loi de densité $f_k(\cdot \mid x_1^{(n-1)},\dots,x_{k-1}^{(n-1)},x_{k+1}^{(n-1)},\dots,x_d^{(n-1)})$

Poser $\mathbf{x}^{(n)} = (x_1^{(n-1)},\dots,x_{k-1}^{(n-1)},x_k^{(n)},x_{k+1}^{(n-1)},\dots,x_d^{(n-1)})$

Retourner les vecteurs $\mathbf{x}^{(0)},\mathbf{x}^{(1)},\mathbf{x}^{(2)},\dots$

2.4.3 Liens avec Metropolis-Hastings

Il existe une relation évidente entre Metropolis-Hastings et l'échantillonneur de Gibbs tel que nous l'avons introduit car, dans cette forme simple, l'échantillonneur de Gibbs est un algorithme de Metropolis-Hastings particulier pour lequel les valeurs candidates ne sont jamais rejetées. En effet, le principe de Metropolis-Hastings est de construire un noyau de transition réversible pour la loi cible μ à partir d'un noyau de transition auxiliaire Q. Or, nous avons vu que le noyau de transition P de l'échantillonneur de Gibbs par balayage aléatoire est déjà réversible par rapport à la loi cible μ. Nous pouvons donc considérer un algorithme de Metropolis-Hastings avec $Q = P$ dans ce cas et les probabilités d'acceptation deviennent systématiquement égales à 1,

$$\forall \mathbf{x},\mathbf{y} \in E,\ \alpha(\mathbf{x},\mathbf{y}) = \min\left\{1, \frac{\mu(\mathbf{y})Q(\mathbf{y},\mathbf{x})}{\mu(\mathbf{x})Q(\mathbf{x},\mathbf{y})}\right\} = \min\left\{1, \frac{\mu(\mathbf{x})P(\mathbf{x},\mathbf{y})}{\mu(\mathbf{x})P(\mathbf{x},\mathbf{y})}\right\} = 1.$$

Toute valeur candidate produite par le noyau P est donc acceptée et l'échantillonneur de Gibbs intègre toujours cette valeur dans la chaîne de Markov produite.

Contrairement à l'algorithme de Metropolis-Hastings qui permet de considérer n'importe quelle loi μ, l'échantillonneur de Gibbs nécessite de pouvoir manipuler les lois conditionnelles de chaque composante. De fait, l'échantillonneur de Gibbs est une méthode liée à la structure de la loi cible et il n'est pas possible d'en donner une implantation universelle. Cependant, une comparaison générale entre Metropolis-Hastings et l'échantillonneur de Gibbs n'aurait pas de sens car, selon le problème considéré, un de ces algorithmes peut avoir de meilleures propriétés que l'autre et *vice versa*. Les lois a posteriori en statistique bayésienne prennent souvent la forme d'un mélange de lois simples et l'échantillonneur de Gibbs est une solution parfois évidente dans ce cadre. Les *modèles hiérarchiques* où chaque composante est en interaction avec un faible nombre d'autres composantes est également un cas favorable à

l'échantillonneur de Gibbs qui tirera partie de lois conditionnelles plus simples à simuler là où Metropolis-Hastings pourra se heurter à des problèmes de grande dimension. D'un autre côté, l'échantillonneur de Gibbs ne pourra pas être utilisé dans des problèmes où la dimension d varie (*e.g.* sélection de modèle en statistique) car son principe est structurellement lié à la nature de la loi cible alors que Metropolis-Hastings pourra être adapté grâce à son caractère universel. Le lecteur intéressé par ces aspects pourra lire le chapitre 11 de [**?**] pour plus d'informations.

En pratique, ces deux algorithmes ne sont d'ailleurs pas opposés mais plutôt utilisés conjointement pour définir de nombreuses variantes. Par exemple, lorsque les lois conditionnelles ne sont connues que à une constante multiplicative près, une trajectoire de Metropolis-Hastings peut être générée à chaque itération de l'échantillonneur de Gibbs pour produire la nouvelle réalisation. Il est même possible de limiter cet appel à une seule itération de Metropolis-Hastings à chaque fois car, par construction, cela ne modifie pas la mesure invariante de la chaîne produite et cette approche demeure valable du point de vue des méthodes MCMC.

3 — Optimisation stochastique

3.1 Recuit simulé

Le *recuit simulé* est une méthode stochastique pour traiter des problèmes d'optimisation tels que la minimisation de fonctions présentée dans la section 1.4. Cette approche s'avère particulièrement utile dans les cas difficiles où une recherche systématique est impossible compte tenu de la taille de l'espace à explorer et où la fonction à minimiser présente un grand nombre de minima locaux que nous souhaitons éviter pour trouver un minimum global. Le *recuit simulé* a été introduit indépendamment en 1983 par Kirkpatrick *et al.* dans [**?**] et en 1985 par Černý dans [**?**]. Nous discutons dans la suite essentiellement de son utilisation pour des problèmes d'*optimisation combinatoire* au sens de la minimisation d'une fonction f à valeurs réelles définie sur un espace E fini mais de taille potentiellement très grande.

3.1.1 Mesures de Gibbs

Les *mesures de Gibbs* introduites dans la section 2.3.4 sont des mesures de probabilités sur un espace E fini. Une motivation importante pour considérer une telle classe de probabilités est liée au principe de physique statistique qui considère qu'un système isolé atteint un état d'équilibre lorsque son *entropie* est maximale. La définition suivante permet de faire le lien entre ce point de vue physique et l'approche probabiliste.

Définition 3.1. Considérons un espace E fini et une mesure de probabilité π sur E. L'*entropie* de la loi π est la quantité

$$H(\pi) = -\sum_{k\in E} \pi_k \ln(\pi_k)$$

où, pour tout $k \in E$, $\pi_k = \pi(k)$.

L'entropie $H(\pi)$ caractérise le niveau d'imprédictibilité de la mesure de probabilité π. Elle est minimale et vaut 0 si il n'y a aucune incertitude au sens où il existe un état $k \in E$ tel que $\pi_k = 1$. Inversement, en l'absence d'information additionnelle, la mesure de probabilité qui maximise l'entropie est la loi uniforme sur E. Il est également possible d'apporter de l'information à l'aide d'une fonction $f : E \to \mathbb{R}$ en supposant connue la valeur moyenne de f

pour la loi π,

$$\sum_{k\in E} f(k)\pi_k.$$

La méthode des multiplicateurs de Lagrange donne facilement la loi de probabilité π qui maximise l'entropie sous cette contrainte et elle est de la forme

$$\forall k \in E,\ \pi_k = \alpha \exp(\beta f(k))$$

où $\alpha, \beta \in \mathbb{R}$ sont deux constantes. En physique, l'opposé de β s'interprète comme l'inverse d'une température $T > 0$ et nous retrouvons la définition (2.14) de la mesure de Gibbs de *fonction d'énergie* f à température T,

$$\forall k \in E,\ \pi_k^T = \frac{1}{Z_T}\exp(-f(k)/T)$$

où la constante de normalisation est donnée par

$$Z_T = \sum_{\ell\in E} \exp(-f(\ell)/T)$$

et porte le nom de *fonction de partition*. Une propriété des mesures de Gibbs qui nous intéresse dans le cadre de l'optimisation stochastique est le fait que, à basse température, elles se concentrent sur les états qui minimisent la fonction f.

Proposition 3.2. *Soit f une fonction à valeurs réelles sur un espace E fini, l'ensemble des minimiseurs de f est défini par*

$$E_{\min} = \{k \in E \text{ tel que } \forall \ell \in E,\ f(k) \leqslant f(\ell)\}.$$

Nous notons $|E_{\min}|$ le cardinal de $E_{\min}$ et π^T la mesure de Gibbs de fonction d'énergie f à température $T > 0$. Alors,

$$\forall k \in E,\ \lim_{T\to 0^+} \pi_k^T = \begin{cases} \dfrac{1}{|E_{\min}|} & \text{si } k \in E_{\min}, \\ 0 & \text{sinon.} \end{cases}$$

De plus, lorsque T tend vers l'infini, la mesure de Gibbs π^T converge vers la loi uniforme sur E.

Démonstration. Soit $m = \min_{k\in E} f(k)$, par définition de $E_{\min}$, la fonction de partition s'écrit

$$Z_T = |E_{\min}|e^{-m/T} + \sum_{\ell\in E\setminus E_{\min}} e^{-f(\ell)/T} = e^{-m/T}\left(|E_{\min}| + \sum_{\ell\in E\setminus E_{\min}} e^{-(f(\ell)-m)/T}\right).$$

Si $\ell \in E \setminus E_{\min}$ alors $f(\ell) > m$ et nous savons que

$$e^{-(f(\ell)-m)/T} \xrightarrow[T\to 0^+]{} 0.$$

Soit $k \in E$, nous pouvons en déduire la valeur limite de la probabilité π_k^T à basse température,

- si $k \in E_{\min}$,

$$\pi_k^T = \frac{e^{-m/T}}{Z_T} = \frac{1}{|E_{\min}| + \sum_{\ell \in E \setminus E_{\min}} e^{-(f(\ell)-m)/T}} \xrightarrow[T\to 0^+]{} \frac{1}{|E_{\min}|}.$$

- si $k \notin E_{\min}$,

$$\pi_k^T = \frac{e^{-m/T} e^{-(f(k)-m)/T}}{Z_T} = \frac{e^{-(f(k)-m)/T}}{|E_{\min}| + \sum_{\ell \in E \setminus E_{\min}} e^{-(f(\ell)-m)/T}} \xrightarrow[T\to 0^+]{} 0.$$

La convergence vers la loi uniforme sur E à haute température se déduit de la définition car, pour tout $k \in E$, $\exp(-f(k)/T)$ tend vers 1 quand $T \to +\infty$. □

La conclusion à basse température de la proposition précédente est que la mesure de Gibbs de fonction d'énergie f est proche de la loi uniforme sur $E_{\min}$. Ainsi, une idée naturelle pour minimiser f consiste à simuler la loi π^T pour une température suffisamment basse et obtenir ainsi des réalisations proches d'un minimiseur de f. Grâce à l'algorithme de Metropolis-Hastings, nous savons qu'il n'est pas nécessaire de connaître la valeur de Z_T pour simuler π^T et cela constitue l'idée à la base du recuit simulé.

Pour assurer l'irréductibilité, nous supposons que E est muni d'une structure de graphe connexe régulier, *i.e.* tous les états ont le même nombre de voisins, ce qui est souvent le cas en pratique. Nous pouvons alors considérer le noyau de transition Q de la marche aléatoire symétrique sur E et les probabilités d'acceptation $\alpha(k,\ell) = \min\{1, \pi_\ell^T / \pi_k^T\}$ pour tout $k, \ell \in E$. Ces probabilités se réécrivent simplement à partir de la définition de la loi π^T,

$$\forall k, \ell \in E,\ \alpha(k,\ell) = \begin{cases} 1 & \text{si } f(\ell) \leqslant f(k), \\ e^{-(f(\ell)-f(k))/T} & \text{si } f(\ell) > f(k). \end{cases}$$

ALGORITHME 3.3 – Recuit simulé simple

Initialisation :
- une fonction réelle f à minimiser sur un espace E fini
- une température $T > 0$ et un état initial $X_0 \in E$

À l'étape $n \geqslant 1$:
- Tirer X' uniformément parmi les voisins de X_{n-1}
- Si $f(X') \leqslant f(X_{n-1})$:
  - Poser $X_n = X'$
- Sinon :
  - Calculer $\alpha = e^{-(f(X')-f(X_{n-1}))/T}$
  - Tirer U selon la loi $\mathscr{U}[0,1]$
  - Règle de rejet :
    - Si $U \leqslant \alpha$, le nouvel état est accepté : $X_n = X'$
    - Si $U > \alpha$, le nouvel état est rejeté : $X_n = X_{n-1}$

Retourner les états $X_0, X_1, X_2, \ldots$

Cet algorithme peut se comprendre comme une marche aléatoire qui saute à coup sûr vers les basses valeurs de la fonction f et s'autorise à effectuer des pas dans la mauvaise direction

avec une petite probabilité. Sans cette possibilité d'évoluer vers des états où la fonction f prend une valeur plus grande, l'algorithme risquerait de rester piégé dans des minima locaux, *i.e.* des états $k \in E$ tels que $f(k) < f(\ell)$ pour tous les états $\ell \in E$ voisins de k. Le problème de cette version du recuit simulé est que, à trop basse température, la probabilité de remonter est très faible et l'algorithme peut rester piégé très longtemps autour de minima locaux.

3.1.2 Schémas de température

Le nom du *recuit simulé* (*simulated annealing* en anglais) fait référence au procédé du recuit en métallurgie qui correspond à un cycle de chauffage d'un métal constitué d'une étape de montée graduelle en température suivie d'un refroidissement contrôlé pour améliorer les qualités du métal. L'idée physique est qu'un refroidissement trop brutal peut bloquer le métal dans un état peu favorable. C'est la même idée que nous retrouvons dans l'algorithme du recuit simulé pour éviter que la séquence produite reste piégée dans des minima locaux. Nous considérons ainsi un *schéma de température* (*cooling schedule* en anglais) donné par une suite de températures $(T_n)_{n\geqslant 1}$ décroissante. Le principe reste le même que dans la version simple 3.3 à la différence que la température change en fonction de l'itération, ce qui implique la non homogénéité de la chaîne de Markov produite.

ALGORITHME 3.4 – Recuit simulé avec schéma de température

Initialisation :
- une fonction réelle f à minimiser sur un espace E fini
- un schéma de température $(T_n)_{n\geqslant 1}$ avec $T_n > 0$ pour tout $n \geqslant 1$
- un état initial $X_0 \in E$

À l'étape $n \geqslant 1$:
- Tirer X' uniformément parmi les voisins de X_{n-1}
- Si $f(X') \leqslant f(X_{n-1})$:
  - Poser $X_n = X'$
- Sinon :
  - Calculer $\alpha = e^{-(f(X')-f(X_{n-1}))/T_n}$
  - Tirer U selon la loi $\mathscr{U}[0,1]$
  - Règle de rejet :
    - Si $U \leqslant \alpha$, le nouvel état est accepté : $X_n = X'$
    - Si $U > \alpha$, le nouvel état est rejeté : $X_n = X_{n-1}$

Retourner les états $X_0, X_1, X_2, \ldots$

La difficulté est ici de bien choisir le schéma de température. Il faut s'assurer que la loi converge encore vers la loi uniforme sur l'ensemble $E_{\min}$ des minimiseurs de la fonction f et que le temps d'atteinte de cette loi soit contrôlé. Dans la suite, nous dirons que le recuit simulé *converge* si la chaîne de Markov $(X_n)_{n\in\mathbb{N}}$ produite par l'algorithme 3.4 est telle que

$$\lim_{n\to+\infty} \mathbb{P}(X_n \in E_{\min}) = 1.$$

Le principal obstacle à la convergence du recuit simulé est la profondeur des minima locaux à éviter.

Définition 3.5. Soit f une fonction à valeurs réelles sur un espace E fini et $E_{\min}$ l'ensemble des minimiseurs de f. Pour tout état $k \in E \setminus E_{\min}$, nous disons que k *communique* avec $E_{\min}$ à hauteur $h > 0$ si il existe une séquence $k_0, \ldots, k_\ell \in E$ d'états voisins reliant k à $E_{\min}$,

$$k_0 = k, \quad k_\ell \in E_{\min} \quad \text{et} \quad k_{j-1} \sim k_j, \ \forall j \in \{1, \ldots, \ell\},$$

tels que

$$\forall j \in \{0, \ldots, \ell\}, \ f(k_j) \leqslant f(k) + h.$$

La *hauteur de communication* h^* de f est la plus petite hauteur à laquelle tout élément de $E \setminus E_{\min}$ communique avec $E_{\min}$.

Un résultat théorique important de convergence du recuit simulé a été démontré par Bruce Hajek en 1988 dans [**?**] et peut être énoncé de la façon suivante.

Théorème 3.6. *Soit f une fonction réelle sur un espace E fini de hauteur de communication $h^* > 0$. L'algorithme du recuit simulé converge pour le schéma de température $(T_n)_{n \geqslant 1}$ si et seulement si*

$$\lim_{n \to +\infty} T_n = 0 \quad \textit{et} \quad \sum_{n \geqslant 1} \exp\left(-\frac{h^*}{T_n}\right) = +\infty.$$

Démonstration. Admis. □

Le schéma de température qui semble le plus naturel à la vue du théorème de Hajek est une décroissance logarithmique de la forme

$$\forall n \geqslant 1, \ T_n = \frac{h}{\ln(n)}$$

avec $h \geqslant h^*$. Il est possible de montrer que le recuit simulé converge pour ce schéma de température sans faire appel au théorème de Hajek (voir le chapitre 6 de [**?**]). Cette convergence est d'autant plus rapide que h est proche de h^*. Cependant, la plupart du temps, la valeur de h^* n'est pas connue et le résultat du théorème ne donne aucune indication sur le temps d'atteinte d'un minimiseur avec une précision donnée. Dans les cas concrets, le paramètre h doit être ajusté expérimentalement.

La fonction logarithme est transcendante et prend donc du temps de calcul à chaque évaluation. Comme la valeur de $\ln(n)$ varie peu quand n est incrémenté, il est peu efficace de la recalculer à chaque pas de l'algorithme. De plus, comme nous le mentionnions ci-dessus, puisque la température T_n change à chaque itération, la chaîne de Markov produite est non homogène, ce qui rend son étude plus difficile. Pour contourner ces difficultés, une idée simple et couramment utilisée en pratique consiste à maintenir la température constante sur des paliers de longueur exponentiellement croissante. Cette variante s'appelle le *recuit simulé par étapes* et elle correspond au schéma de température défini par

$$\forall N \geqslant 1, \ \forall n \in \left\{e^{a(N-1)}, \ldots, e^{aN}\right\}, \ T_n = \frac{b}{N}. \tag{3.1}$$

où $a, b > 0$ sont des paramètres à calibrer. Pour un tel schéma de température, la chaîne de Markov est homogène sur des intervalles de temps de plus en plus longs. Nous allons montrer

que ces intervalles sont suffisamment grands pour que la chaîne atteigne son équilibre, à savoir la mesure de Gibbs π^T de fonction d'énergie f à la température T du palier. Les températures T_n étant décroissantes vers 0, la mesure π^{T_n} se concentre vers la loi uniforme sur $E_{\min}$ et cela assurera la convergence du recuit simulé par étapes.

3.1.3 Convergence du recuit simulé par étapes

Soit un espace E de cardinal K fini, nous commençons par établir deux inégalités utiles sur la distance en variation totale entre des mesures de probabilité ν et μ sur E. Dans ce qui suit, la mesure de probabilité μ est supposée charger tous les états, *i.e.* $\mu_k = \mu(k) > 0$ pour tout $k \in E$.

Le premier résultat que nous considérons concerne la divergence de Kullback-Leibler entre ν et μ définie par

$$\mathscr{K}(\nu,\mu) = \sum_{k\in E} \nu_k \ln\left(\frac{\nu_k}{\mu_k}\right).$$

En utilisant l'inégalité $3(t-1)^2 \leqslant 2(2+t)(t\ln(t)+1-t)$ valable pour tout $t \geqslant 0$, nous déduisons de l'inégalité de Cauchy-Schwarz que

$$\begin{aligned}
3\|\nu-\mu\|_{VT}^2 &= \frac{1}{4}\left(\sum_{k\in E}\sqrt{3}|\nu_k-\mu_k|\right)^2 = \frac{1}{4}\left(\sum_{k\in E}\mu_k\sqrt{3}\left|\frac{\nu_k}{\mu_k}-1\right|\right)^2 \\
&\leqslant \frac{1}{4}\left(\sum_{k\in E}\mu_k\sqrt{2\left(2+\frac{\nu_k}{\mu_k}\right)\left(\frac{\nu_k}{\mu_k}\ln\left(\frac{\nu_k}{\mu_k}\right)+1-\frac{\nu_k}{\mu_k}\right)}\right)^2 \\
&\leqslant \frac{1}{4}\left(\sum_{k\in E}2(2\mu_k+\nu_k)\right)\left(\sum_{k\in E}\nu_k\ln\left(\frac{\nu_k}{\mu_k}\right)+\mu_k-\nu_k\right) \\
&= \frac{3}{2}\mathscr{K}(\nu,\mu).
\end{aligned}$$

Nous avons ainsi obtenu

$$\|\nu-\mu\|_{VT}^2 \leqslant \frac{1}{2}\mathscr{K}(\nu,\mu). \tag{3.2}$$

La seconde inégalité est liée à un noyau de Metropolis-Hastings P qui admet μ comme unique mesure de probabilité invariante. Si $(X_n)_{n\in\mathbb{N}}$ est une chaîne de Markov de loi initiale ν et de matrice de transition P, alors nous savons que, pour tout $n \in \mathbb{N}$, la loi de X_n est donnée par

$$\forall k \in E,\ \nu P^n(k) = \sum_{\ell\in E}\nu_l (P^n)_{\ell,k}.$$

Nous pouvons ordonner les K valeurs propres de la matrice P par ordre décroissant avec leur multiplicité,

$$\lambda_1 = 1 > \lambda_2 \geqslant \cdots \geqslant \lambda_K > -1.$$

Ainsi, un raisonnement similaire à ce que nous avons fait pour obtenir (2.13) nous conduit à

$$\|\nu P^n-\mu\|_{VT}^2 \leqslant \frac{1}{4}\max\left\{\lambda_2^{2n},\lambda_K^{2n}\right\}\chi^2(\nu,\mu). \tag{3.3}$$

Pour établir une vitesse de convergence, nous faisons une récurrence sur les étapes du recuit. Soient $a,b>0$, nous considérons le schéma de température (3.1). Pour tout $N \in \mathbb{N}$,

$R_N = \lceil e^{aN} \rceil$ désigne le plus petit entier supérieur ou égal à e^{aN} et, pour $N \geqslant 1$, nous notons $r_N = R_N - R_{N-1}$ la longueur du palier où la température est constante égale à b/N. Remarquons que la suite des entiers R_N devient strictement croissante et les paliers ne sont plus vides dès que N dépasse $-\ln(e^a - 1)/a$, nous nous placerons implicitement dans ce cas dans la suite quitte à décaler l'énumération des paliers si la valeur de a est trop faible.

Notre objectif est de minimiser une fonction $f : E \to \mathbb{R}$ de hauteur de communication $h^* > 0$ pour laquelle nous introduisons les notations suivantes,

$$m = \min_{k \in E} f(k), \quad F_- = \min_{\substack{k \in E \text{ t.q.} \\ f(k) \neq m}} \{f(k) - m\} \quad \text{et} \quad F_+ = \max_{k \in E} \{f(k) - m\}.$$

Pour tout entier $N \geqslant 1$, nous notons également π_N la mesure de Gibbs sur E de fonction d'énergie f à la température b/N,

$$\forall k \in E,\ \pi_{N,k} = \frac{1}{Z_{f,N}} \exp(-N(f(k) - m)/b) \quad \text{où} \quad Z_{f,N} = \sum_{\ell \in E} \exp(-N(f(\ell) - m)/b).$$

Dans la suite, nous supposons qu'il existe $M > 0$ tel que la fonction de partition soit bornée,

$$\sup_{N \geqslant 1} Z_{f,N} \leqslant M.$$

Pour un espace E de cardinal K fini comme cela est le cas ici, il est toujours possible de prendre $M = K$ bien que toute valeur inférieure donne de meilleures constantes dans la suite.

Nous considérons maintenant la chaîne de Markov $(X_n)_{n \in \mathbb{N}}$ du recuit simulé par étapes donné par le schéma de température (3.1) et par la fonction f. Pour tout $N \geqslant 1$, la loi des variables $X_{R_{N-1}}, \ldots, X_{R_N - 1}$ est celle d'une chaîne de Markov homogène donnée par un algorithme de Metropolis-Hastings de noyau P_N réversible par rapport à π_N et de valeurs propres notées $\lambda_{N,1}, \ldots, \lambda_{N,K}$. Par la propriété 2.35 de contraction markovienne, nous savons

$$\forall j \in \{R_{N-1}, \ldots, R_N - 1\},\ \|\mu_j - \pi_N\|_{VT} = \|\mu_{R_{N-1}} P_N^{j - R_{N-1}} - \pi_N\|_{VT} \leqslant \|\mu_{R_{N-1}} - \pi_N\|_{VT}$$

où, pour tout $n \in \mathbb{N}$, μ_n désigne la loi de X_n. Nous considérons ainsi la suite $(u_N)_{N \geqslant 1}$ des carrés des écarts en variation totale entre la loi initiale et la probabilité invariante sur un palier donné,

$$\forall N \geqslant 1,\ u_N = \|\mu_{R_{N-1}} - \pi_N\|_{VT}^2.$$

Fixons un palier $N \geqslant 1$ tel que $r_N > 0$, notre objectif est d'établir une relation de récurence entre deux paliers consécutifs,

$$\begin{aligned} u_{N+1} = \|\mu_{R_N} - \pi_{N+1}\|_{VT}^2 &= \|\mu_{R_{N-1}} P_N^{r_N} - \pi_{N+1}\|_{VT}^2 \\ &\leqslant 2\|\mu_{R_{N-1}} P_N^{r_N} - \pi_N\|_{VT}^2 + 2\|\pi_N - \pi_{N+1}\|_{VT}^2 \\ &\leqslant \frac{1}{2} \max\left\{\lambda_{N,2}^{2r_N}, \lambda_{N,K}^{2r_N}\right\} \chi^2(\mu_{R_{N-1}}, \pi_N) + 2\|\pi_N - \pi_{N+1}\|_{VT}^2 \end{aligned}$$

où la dernière inégalité se déduit de (3.3). La distance du χ^2 est telle que

$$\chi^2(\mu_{R_{N-1}}, \pi_N) = \sum_{k \in E} \frac{(\mu_{R_{N-1},k} - \pi_{N,k})^2}{\pi_{N,k}} \leqslant M e^{NF_+/b} \left(\sum_{k \in E} |\mu_{R_{N-1},k} - \pi_{N,k}| \right)^2 = 4M e^{NF_+/b} u_N.$$

Le contrôle des valeurs propres de P_N demande de faire une analyse spectrale du noyau de Metropolis-Hastings qui dépasse le cadre de ce cours. Nous admettrons le résultat suivant dont la preuve peut être trouvée dans la section 6.III.2 de [**?**].

Théorème 3.7. *Soit $T > 0$, nous notons P_T la matrice donnée par le noyau de transition de la chaîne de Markov homogène produite par l'algorithme de Metropolis-Hastings pour la mesure de Gibbs π^T sur E de cardinal K fini, à température T et de fonction d'énergie f de hauteur de communication $h^* > 0$. Alors, il existe une constante $C > 0$ telle que*

$$\forall j \in \{2,\dots,K\},\ \lambda_{T,j}^2 \leqslant 1 - C\exp\left(-\frac{h^*}{T}\right)$$

où $\lambda_{T,2},\dots,\lambda_{T,K}$ désignent les valeurs propres de P_T strictement inférieures à 1.

Démonstration. Admis. □

À l'aide de ce théorème, nous déduisons l'inégalité suivante qui établit une relation entre u_N et u_{N+1},

$$u_{N+1} \leqslant 2Me^{NF_+/b}\left(1 - Ce^{-Nh^*/b}\right)^{2r_N} u_N + 2\|\pi_N - \pi_{N+1}\|_{VT}^2. \tag{3.4}$$

Ce résultat permet de donner l'ordre de grandeur de la vitesse de convergence. En effet, si r_N est assez grand pour négliger le premier terme, cette borne sur u_{N+1} se limite au second terme. Grâce à (3.2) et de simples calculs à partir des hypothèses, nous savons qu'il existe une constante $C' > 0$ telle que

$$2\|\pi_N - \pi_{N+1}\|_{VT}^2 \leqslant \mathscr{K}(\pi_N, \pi_{N+1}) \leqslant \frac{C'e^{-NF_-/b}}{b}.$$

Par construction, $N \geqslant \ln(n)/a$ et nous en déduisons une vitesse de convergence polynomiale,

$$\|\mu_n - \pi^{ab/\ln(n)}\|_{VT} \lesssim \frac{C'n^{-F_-/(2ab)}}{b}.$$

Cette heuristique se formalise en faisant une étude fine de la récurrence donnée par (3.4),

$$u_{N+1} \leqslant \beta_N u_N + \delta_N$$

avec, pour tout $N \geqslant 1$,

$$\beta_N = 2Me^{NF_+/b}\left(1 - Ce^{-Nh^*/b}\right)^{2r_N} \quad \text{et} \quad \delta_N = \frac{C'e^{-NF_-/b}}{b}.$$

En itérant cette relation, nous obtenons

$$u_{N+1} \leqslant B_N\left(u_1 + \sum_{j=1}^{N}\frac{\delta_j}{B_j}\right)$$

où, pour tout $j \geqslant 1$, $B_j = \prod_{\ell=1}^{j}\beta_\ell$.

Nous ne détaillerons pas ici l'ensemble des calculs pour étudier cette récurrence, il s'agit d'un exercice exigeant qui peut être réalisé par un lecteur rigoureux. Nous précisons cependant que la première étape consiste à contrôler B_N. Pour cela, nous savons qu'il existe $\rho_a > 0$ tel que $r_N \geqslant \rho_a e^{aN}$. Si b est assez petit pour satisfaire $h^* > b\ln(C)$, alors nous obtenons

$$\begin{aligned} B_N &= \prod_{j=1}^{N} \exp\left(2r_j \ln(1-Ce^{-jh^*/b}) + \frac{jF_+}{b} + \ln(2M)\right) \\ &\leqslant \prod_{j=1}^{N} \exp\left(-2C\rho_a e^{j(a-h^*/b)} + \frac{jF_+}{b} + \ln(2M)\right) \\ &= \exp\left(-2C\rho_a \sum_{j=1}^{N} e^{j(a-h^*/b)} + \frac{N(N+1)F_+}{2b} + N\ln(2M)\right) \end{aligned}$$

où nous avons utilisé l'inégalité $\ln(1-x) \leqslant -x$ valable pour tout $0 < x < 1$. Une condition nécessaire pour assurer la convergence de B_N vers 0 est que les constantes $a, b > 0$ du schéma de température vérifient

$$a - h^*/b > 0.$$

Cette condition est équivalente à $ab > h^*$ et, en faisant jouer au produit ab le rôle de h dans le schéma de température à décroissance logarithmique, nous retrouvons la condition $h > h^*$ donnée par le théorème de Hajek.

Sous certaines conditions, l'étude de la récurrence (3.4) conduit à la vitesse de convergence suivante pour le recuit simulé par étapes,

$$\|\mu_n - \pi^{ab/\ln(n)}\|_{VT} \leqslant C_1 \ln\ln(n)\, n^{-C_2 F_-/(ab)}$$

où $C_1, C_2 > 0$ sont des constantes.

3.1.4 Application : problème du voyageur de commerce

Afin d'illustrer l'algorithme du recuit simulé, nous proposons de l'appliquer au problème du voyageur de commerce qui est un des problèmes les plus célèbres de l'optimisation combinatoire. Il s'agit d'un problème NP-complet qui fut initialement décrit en 1832 dans un écrit relatif aux choix d'itinéraires efficaces entre plusieurs villes pour un commis voyageur.

Pour $d > 1$, nous identifions d villes aux éléments de $\{1,\ldots,d\}$ et nous notons $\delta(k,\ell)$ la distance entre les villes $k, \ell \in \{1,\ldots,d\}$. Le problème consiste à trouver un trajet partant de la ville 1 et revenant à la ville 1 de longueur minimale pour que chaque ville soit visitée une seule fois. Autrement dit, un trajet peut être vu comme une permutation circulaire $\sigma = (\sigma_1,\ldots,\sigma_d)$ de $\{1,\ldots,d\}$ telle que $\sigma_1 = 1$ et nous cherchons celle qui minimise la fonction

$$f(\sigma) = \sum_{k=1}^{d} \delta(\sigma_k, \sigma_{k+1}) \quad \text{où} \quad \sigma_{d+1} = \sigma_1.$$

Bien que l'énoncé de ce problème soit simple, il est très difficile de trouver un tel trajet de longueur minimale car il existe $(d-1)!$ permutations circulaires possibles. Même pour des valeurs de d raisonnables, il n'est pas possible de calculer f pour chaque trajet, *e.g.* pour 11 villes, il y a 3628800 trajets possibles et pour 26 villes, ce nombre dépasse 10^{25}.

Le recuit simulé permet d'obtenir une solution approchée du problème du voyageur de commerce. Dans la mise en place de l'algorithme, la structure de graphe connexe régulier

de l'espace d'états E issue de notre choix de voisinage est cruciale et a des conséquences importantes en pratique. Pour illustrer cela, nous proposons deux structures différentes sur l'espace E des permutations circulaires de $\{1,\dots,d\}$ qui fixent 1 au travers de la définition de deux marches aléatoires symétriques sur E. Étant donné un trajet $\sigma = (\sigma_1,\dots,\sigma_d) \in E$, nous introduisons les noyaux de transition Q_1 et Q_2 définis comme suit :

- pour $2 \leqslant k < \ell \leqslant d$, nous notons $\sigma^{(k,\ell)}$ le trajet obtenu en permutant l'ordre de visite des villes σ_k et σ_ℓ,

$$\forall j \in \{1,\dots,d\},\ \sigma_j^{(k,\ell)} = \begin{cases} \sigma_\ell & \text{si } j = k, \\ \sigma_k & \text{si } j = \ell, \\ \sigma_j & \text{sinon.} \end{cases}$$

  Nous posons alors

$$\forall 2 \leqslant k < \ell \leqslant d,\ Q_1(\sigma,\sigma^{(k,\ell)}) = \frac{2}{(d-1)(d-2)}.$$

- pour $2 \leqslant k < \ell \leqslant d$, nous notons $\sigma^{[k,\ell]}$ le trajet obtenu en inversant l'ordre de visite des villes entre σ_k et σ_ℓ,

$$\forall j \in \{1,\dots,d\},\ \sigma_j^{[k,\ell]} = \begin{cases} \sigma_{k+\ell-j} & \text{si } k \leqslant j \leqslant \ell, \\ \sigma_j & \text{sinon.} \end{cases}$$

  Nous posons alors

$$\forall 2 \leqslant k < \ell \leqslant d,\ Q_2(\sigma,\sigma^{[k,\ell]}) = \frac{2}{(d-1)(d-2)}.$$

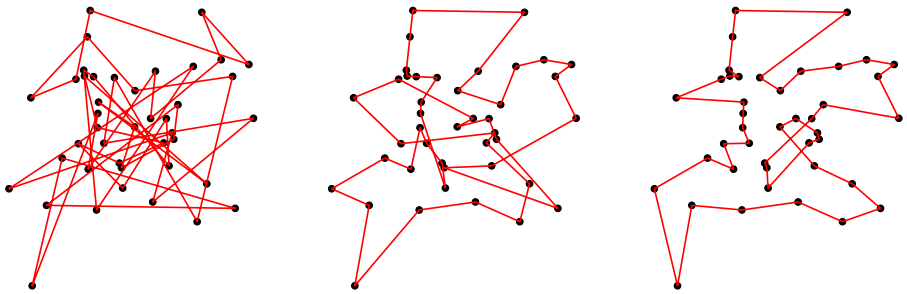

FIGURE 3.1 – Exemples d'itérations du recuit simulé pour résoudre le problème du voyageur à 42 villes tirées au hasard dans le plan. Le noyau Q_1 est utilisé avec le schéma de température par étapes pour $a = 0.01$ et $b = 100$. En partant d'un état initial chaotique (à gauche), la longueur du chemin est réduite après 1000 itérations (au centre) et le trajet tend à converger après 10000 itérations vers une solution (presque) optimale (à droite).

Pour comparer les choix de noyau et de schéma de température, nous considérons le même problème du voyageur à 42 villes tirées au hasard dans le plan. Dans chaque cas, le trajet final

est représenté ainsi que la longueur du chemin, la probabilité d'acceptation et la température au cours des itérations. La figure 3.2 a été obtenue avec le noyau Q_1 et la figure 3.3 avec Q_2. La longueur du chemin décroît plus rapidement avec Q_2 qui semble mieux adapté pour le problème du voyageur puisque il faut attendre plus de 4000 itérations avec le noyau Q_1 pour s'approcher de la convergence alors que 2000 environ semblent suffire avec Q_2. La figure 3.4 est également obtenue avec le noyau Q_2 mais pour un schéma de température continu à décroissance logarithmique. Les propriétés sont sensiblement les mêmes que dans le cas du recuit simulé par étapes mais la complexité en temps est plus élevée à cause des nombreux appels à la fonction logarithme. Il est cependant remarquable que la longueur fluctue plus dans le cas du schéma de température continu car l'absence de palier ne laisse pas le temps à l'algorithme de se stabiliser, ce qui peut aussi être un avantage du point de vue de l'exploration de l'espace d'états.

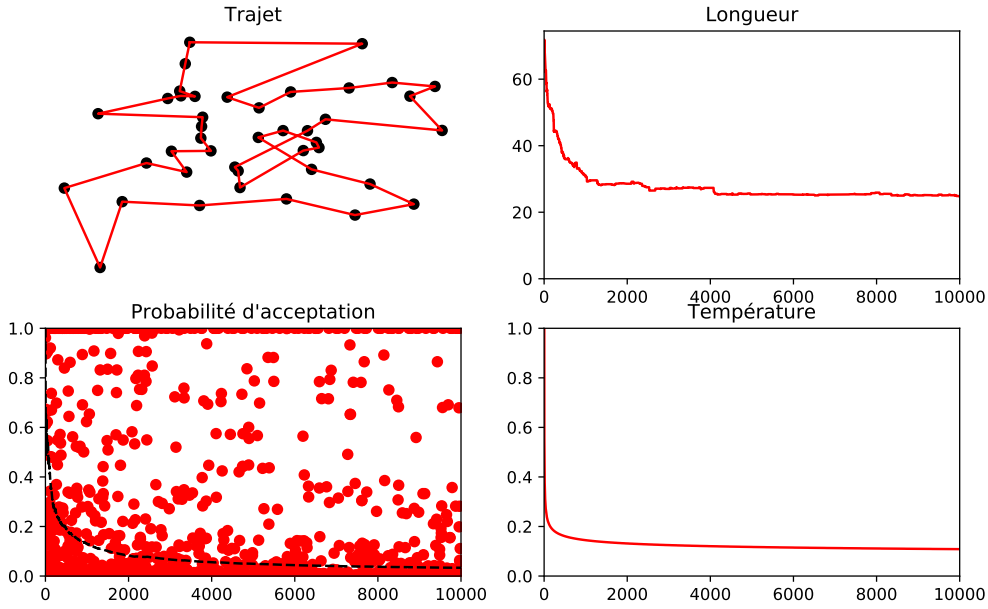

FIGURE 3.2 – Évolution du recuit simulé pour le problème du voyageur à 42 villes avec le noyau Q_1 et un schéma de température par étapes ($a = 0.01$ et $b = 100$).

3.2 Descente de gradient stochastique

Avec le recuit simulé, nous avons pu générer des réalisations selon une loi adaptée à la fonction f à minimiser. En particulier, cette méthode ne nécessite pas d'avoir des propriétés particulières sur f. Bien que cela puisse être un avantage dans le cas d'une fonction quelconque, il peut aussi être bénéfique d'exploiter des propriétés locales de f pour diriger la marche aléatoire, *e.g.* nous pourrions descendre d'autant plus « vite » que l'amplitude entre la valeur courante de la fonction et la valeur candidate est importante. Dans un cadre déterministe, la régularité de la fonction f est une caractéristique importante pour la mise en place de méthodes d'optimisation et nous proposons dans cette section d'étudier comment cette approche peut être exploitée du point de vue des algorithmes stochastiques.

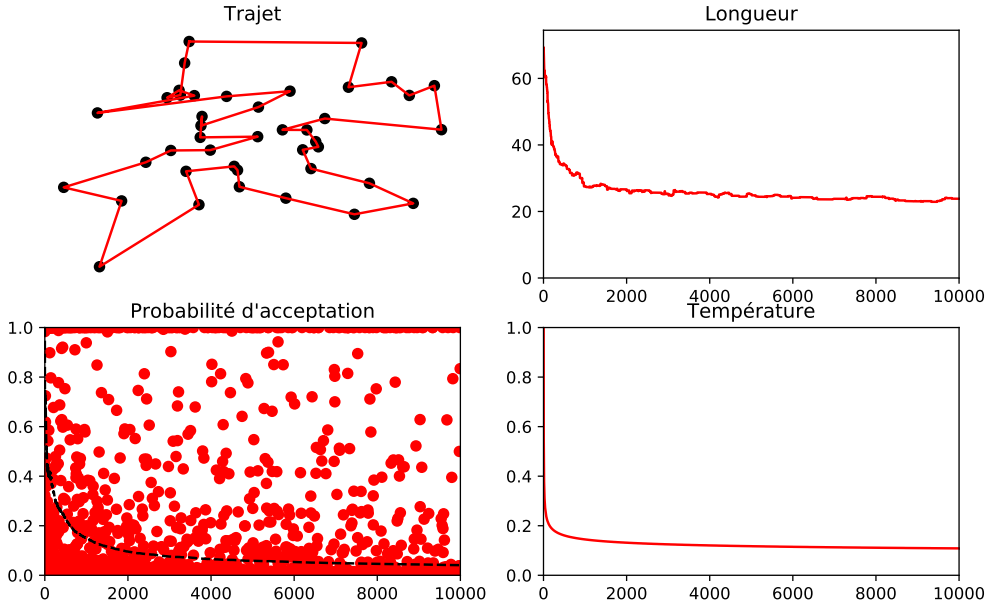

FIGURE 3.3 – Évolution du recuit simulé pour le problème du voyageur à 42 villes avec le noyau Q_2 et un schéma de température par étapes ($a = 0.01$ et $b = 100$).

Pour une fonction régulière, la recherche d'un extremum se ramène à celle d'un zéro de son gradient. Ce problème est plus général que la seule optimisation stochastique et les premiers résultats dans le domaine remontent aux années 1950 grâce aux outils développés par Herbert Ellis Robbins dans le papier séminal [**?**] co-écrit avec Sutton Monro. Une approche alternative introduite à la même époque par Jack Kiefer et Jacob Wolfowitz dans [**?**] permet de répondre sensiblement aux même questions d'optimisation lorsque le gradient est approché par différences finies. Nous ne développerons pas ici la méthode de Kiefer-Wolfowitz et nous présenterons celle de Robbins-Monro et ses liens avec la théorie des martingales.

Dans la suite de cette section, pour tout entier $d > 0$, nous utilisons des notations génériques pour désigner le produit scalaire et la norme canoniques sur $\mathbb{R}^d$,

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d,\ \langle \theta_1, \theta_2 \rangle = \theta_1' \theta_2 \quad \text{et} \quad \|\theta_1\| = \sqrt{\langle \theta_1, \theta_1 \rangle}.$$

3.2.1 Éléments de théorie des martingales

Nous donnons ici les définitions principales de la théorie des martingales nécessaires pour présenter quelques résultats de convergence importants qui seront utiles dans la suite. Le lecteur intéressé par une présentation plus complète de cette théorie pourra se reporter à [**?**].

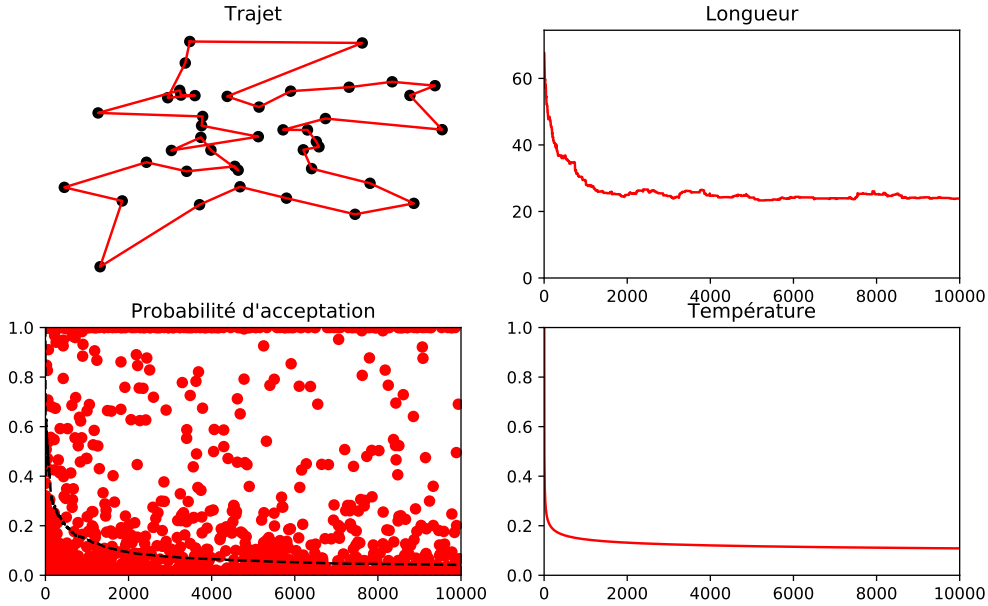

FIGURE 3.4 – Évolution du recuit simulé pour le problème du voyageur à 42 villes avec le noyau Q_2 et un schéma de température continu à décroissance logarithmique ($h = 1.0$).

Définition 3.8. Une suite croissante de tribus $(\mathscr{F}_n)_{n\in\mathbb{N}}$, *i.e.* telle que $\mathscr{F}_n \subset \mathscr{F}_{n+1}$ pour tout $n \in \mathbb{N}$, est appelée un *filtration*.

- Une suite de variables aléatoires $(M_n)_{n\in\mathbb{N}}$ est dite *adaptée* à la filtration $(\mathscr{F}_n)_{n\in\mathbb{N}}$ si, pour tout $n \in \mathbb{N}$, la variable M_n est $\mathscr{F}_n$-mesurable.
- Une suite de variables aléatoires $(X_n)_{n\in\mathbb{N}}$ est dite *prévisible* pour la filtration $(\mathscr{F}_n)_{n\in\mathbb{N}}$ si, pour tout $n \in \mathbb{N}$, la variable X_{n+1} est $\mathscr{F}_n$-mesurable.

La notion de filtration peut paraître abstraite au premier regard mais elle est nécessaire pour définir le concept de *martingale*. En pratique, une filtration peut se comprendre comme une accumulation de connaissances au cours du temps et la tribu $\mathscr{F}_n$ désigne toutes les événements qui ont pu avoir lieu entre l'origine et le temps $n \in \mathbb{N}$. Pour une suite de variables aléatoires $(X_n)_{n\in\mathbb{N}}$, nous disposons d'une filtration naturelle définie par les tribus engendrées par les premières variables,

$$\forall n \in \mathbb{N},\ \mathscr{F}_n = \sigma(X_0, \ldots, X_n). \tag{3.5}$$

Par construction, la suite $(X_n)_{n\in\mathbb{N}}$ est adaptée à cette filtration et la tribu $\mathscr{F}_n$ contient toute l'histoire de la suite jusqu'au temps $n \in \mathbb{N}$.

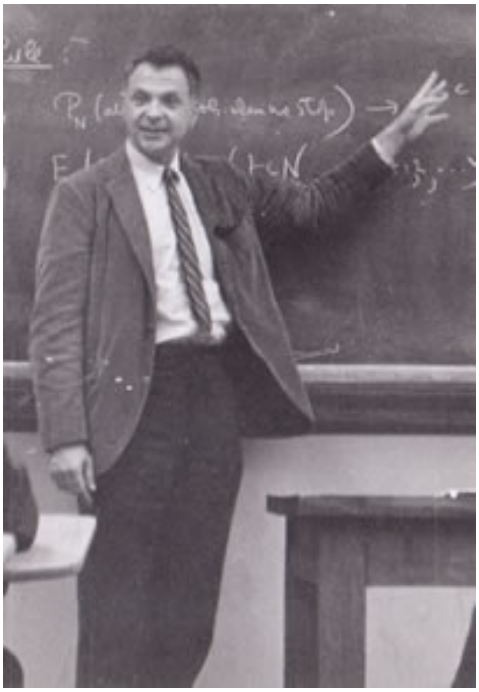

FIGURE 3.5 – Herbert Ellis Robbins (1915-2001)

Définition 3.9. Soit $(M_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires réelles adaptée à une filtration $(\mathscr{F}_n)_{n\in\mathbb{N}}$. Si les variables M_n sont intégrables, *i.e.* $\mathbb{E}[|M_n|] < +\infty$ pour tout $n \in \mathbb{N}$, alors

- $(M_n)_{n\in\mathbb{N}}$ est une *martingale* si

$$\forall n \in \mathbb{N},\ \mathbb{E}[M_{n+1} \mid \mathscr{F}_n] = M_n.$$

- $(M_n)_{n\in\mathbb{N}}$ est une *sur-martingale* si

$$\forall n \in \mathbb{N},\ \mathbb{E}[M_{n+1} \mid \mathscr{F}_n] \leqslant M_n.$$

- $(M_n)_{n\in\mathbb{N}}$ est une *sous-martingale* si

$$\forall n \in \mathbb{N},\ \mathbb{E}[M_{n+1} \mid \mathscr{F}_n] \geqslant M_n.$$

Une conséquence immédiate de cette définition est que, si $(M_n)_{n\in\mathbb{N}}$ est une sur-martingale, alors $(-M_n)_{n\in\mathbb{N}}$ est une sous-martingale. De plus, si $(M_n)_{n\in\mathbb{N}}$ est une martingale adaptée à la filtration $(\mathscr{F}_n)_{n\in\mathbb{N}}$, les *incréments* $(\Delta M_n)_{n\geqslant 1}$ sont donnés par

$$\forall n \geqslant 1,\ \Delta M_n = M_n - M_{n-1}.$$

Par définition, des incréments de martingale sont centrés,

$$\forall n \in \mathbb{N},\ \mathbb{E}[\Delta M_{n+1} \mid \mathscr{F}_n] = \mathbb{E}[M_{n+1} \mid \mathscr{F}_n] - M_n = 0,$$

et orthogonaux deux à deux car

$$\begin{aligned}\forall 0 \leqslant k < \ell,\ \mathbb{E}[\Delta M_{k+1}\Delta M_{\ell+1}] &= \mathbb{E}\left[\Delta M_{k+1}\mathbb{E}[M_{\ell+1} - M_\ell \mid \mathscr{F}_{k+1}]\right] \\ &= \mathbb{E}\left[\Delta M_{k+1}(M_{k+1} - M_{k+1})\right] \\ &= 0.\end{aligned}$$

Le principe d'une martingale $(M_n)_{n\in\mathbb{N}}$ est que, au temps $n \in \mathbb{N}$, toute l'information utile pour prédire la valeur M_{n+1} à venir est contenue dans la valeur courante M_n. Le plus ancien résultat de convergence en théorie des martingales est dû à Joseph Leo Doob (1910-2004).

Théorème 3.10 (Doob)**.** *Soit $(X_n)_{n\in\mathbb{N}}$ une sous-martingale telle que*

$$\sup_{n\in\mathbb{N}} E\left[X_n^+\right] < +\infty$$

où $x^+ = \max\{0, x\}$. Alors il existe une variable aléatoire X_∞ intégrable telle que

$$X_n \xrightarrow[n\to+\infty]{p.s.} X_\infty.$$

Démonstration. Admis. □

La théorie des martingales fournit des outils puissants pour l'étude des algorithmes stochastiques. En particulier, le résultat suivant joue un rôle clé pour la méthode de Robbins-Monro et a été publié en 1971 par Herbert Robbins et David Siegmund dans [**?**].

Théorème 3.11. *Soient quatre suites de variables aléatoires intégrables $(U_n)_{n\in\mathbb{N}}$, $(V_n)_{n\in\mathbb{N}}$, $(\alpha_n)_{n\in\mathbb{N}}$ et $(\beta_n)_{n\in\mathbb{N}}$ adaptées à une filtration $(\mathscr{F}_n)_{n\in\mathbb{N}}$ et à valeurs dans $\mathbb{R}_+$ telles que*

$$A = \sum_{n\in\mathbb{N}} \alpha_n < +\infty,\ p.s. \quad et \quad B = \sum_{n\in\mathbb{N}} \beta_n < +\infty,\ p.s.$$

Si, pour tout $n \in \mathbb{N}$,

$$\mathbb{E}[V_{n+1} \mid \mathscr{F}_n] \leqslant V_n(1+\alpha_n) + \beta_n - U_n$$

alors

$$V_n \xrightarrow[n\to+\infty]{p.s.} V_\infty \quad et \quad \sum_{n\in\mathbb{N}} U_n < +\infty,\ p.s.$$

où V_∞ est une variable aléatoire intégrable.

Démonstration. Nous commençons par définir une suite $(A_n)_{n\in\mathbb{N}}$ adaptée à $(\mathscr{F}_n)_{n\in\mathbb{N}}$ par

$$\forall n \in \mathbb{N},\ A_n = \prod_{k=0}^{n}(1+\alpha_k) \leqslant \exp\left(\sum_{k=1}^{n}\ln(1+\alpha_k)\right) \leqslant e^A,\ p.s.$$

Soit $n \geqslant 1$, par construction, nous avons $(1+\alpha_n) = A_n/A_{n-1}$ et donc

$$\mathbb{E}\left[\frac{V_{n+1}}{A_n} \mid \mathscr{F}_n\right] \leqslant \frac{V_n}{A_{n-1}} + \frac{\beta_n - U_n}{A_n}.$$

Considérons la suite $(X_n)_{n\in\mathbb{N}}$ adaptée à $(\mathscr{F}_n)_{n\in\mathbb{N}}$ définie par $X_0 = V_0$ et

$$\forall n \geqslant 1,\ X_n = \frac{V_n}{A_{n-1}} - \sum_{k=0}^{n-1} \frac{\beta_k - U_k}{A_k}.$$

La suite $(X_n)_{n\in\mathbb{N}}$ est une sur-martingale car l'inégalité précédente conduit à

$$\forall n \in \mathbb{N},\ \mathbb{E}[X_{n+1} \mid \mathscr{F}_n] \leqslant \frac{V_n}{A_{n-1}} + \frac{\beta_n - U_n}{A_n} - \sum_{k=0}^{n} \frac{\beta_k - U_k}{A_k} = X_n.$$

Soit $a > 0$, nous introduisons la variable aléatoire

$$T_a = \inf\left\{ n \in \mathbb{N} \text{ tel que } \sum_{k=0}^{n} \frac{\beta_k}{A_k} > a \right\}$$

avec $T_a = +\infty$ si l'ensemble est vide. Nous appelons T_a un *temps d'arrêt* car, pour tout $n \in \mathbb{N}$, l'événement $\{T_a = n\}$ ne dépend que des variables $\alpha_0, \ldots, \alpha_n$ et $\beta_0, \ldots, \beta_n$ et il est donc $\mathscr{F}_n$-mesurable. De même, la variable $X_{T_a}\mathbf{1}_{T_a \leqslant n}$ est $\mathscr{F}_n$-mesurable et comme $(X_n)_{n\in\mathbb{N}}$ est une sur-martingale, nous en déduisons

$$\begin{aligned}
\mathbb{E}[X_{(n+1)\wedge T_a} \mid \mathscr{F}_n] &= \mathbb{E}[X_{T_a}\mathbf{1}_{T_a \leqslant n} \mid \mathscr{F}_n] + \mathbb{E}[X_{n+1}\mathbf{1}_{T_a > n} \mid \mathscr{F}_n] \\
&= X_{T_a}\mathbf{1}_{T_a \leqslant n} + \mathbb{E}[X_{n+1} \mid \mathscr{F}_n]\mathbf{1}_{T_a > n} \\
&\leqslant X_{T_a}\mathbf{1}_{T_a \leqslant n} + X_n\mathbf{1}_{T_a > n} \\
&= X_{n\wedge T_a}.
\end{aligned}$$

où $n \wedge T_a = \min\{n, T_a\}$. Ainsi, $(X_{n\wedge T_a})_{n\in\mathbb{N}}$ est également une sur-martingale et elle satisfait

$$\forall n \in \mathbb{N},\ X_{n\wedge T_a} \geqslant -\sum_{k=0}^{n\wedge T_a - 1} \frac{\beta_k}{A_k} \geqslant -a.$$

Le théorème 3.10 s'applique et nous en déduisons que $X_n\mathbf{1}_{T_a=+\infty}$ converge presque sûrement vers une variable aléatoire intégrable. L'événement $\{T_a = +\infty\}$ est tel que

$$T_a = +\infty \iff \sum_{k\geqslant 0} \frac{\beta_k}{A_k} \leqslant a$$

et puisque nous avons par hypothèse que

$$\sum_{k\geqslant 0} \frac{\beta_k}{A_k} \leqslant B < +\infty,\ p.s.$$

alors, en faisant tendre a vers l'infini, nous obtenons l'existence d'une variable aléatoire X_∞ intégrable telle que

$$X_n \xrightarrow[n\to+\infty]{p.s.} X_\infty.$$

En tant que différence de termes positifs, ce résultat implique la convergence presque sûre des variables

$$\frac{V_n}{A_{n-1}} + \sum_{k=0}^{n-1} \frac{U_k}{A_k} \quad \text{et} \quad \sum_{k=0}^{n-1} \frac{\beta_k}{A_k}$$

quand n tend vers l'infini. Le résultat annoncé découle alors de la convergence presque sûre de A_n qui est immédiate à établir sous les hypothèses du théorème. □

3.2.2 Méthode de Robbins-Monro

Le problème que nous considérons ici est celui de la recherche des solutions $\theta \in \mathbb{R}^d$ d'une équation du type $h(\theta) = 0$ pour une fonction $h : \mathbb{R}^d \to \mathbb{R}^d$ de la forme

$$h(\theta) = \mathbb{E}[H(X, \theta)] \tag{3.6}$$

où $H(\cdot, \theta)$ est une fonction connue d'une certaine variable aléatoire X. Ce cadre est intéressant en pratique lorsque la fonction h ne peut pas être calculée explicitement (ou quand l'évaluation de H est beaucoup moins coûteuse que celle de h) mais simplement estimée par des observations $H(X_1, \theta), \ldots, H(X_n, \theta)$. Cette question est liée à l'optimisation d'une fonction $f : \mathbb{R}^d \to \mathbb{R}$ différentiable car elle permet d'identifier les zéros de $h = \nabla f$. Nous reviendrons sur ce cas particulier en fin de section.

L'*algorithme de Robbins-Monro* est une méthode récursive pour estimer un zéro d'une fonction $h : \mathbb{R}^d \to \mathbb{R}^d$ de la forme (3.6) dont nous avons vu un exemple très simple dans la sous-section 1.4.2 pour l'estimation de la moyenne d'une variable aléatoire réelle X, *i.e.* $h(\theta) = \theta - \mathbb{E}[X]$ et $H(X, \theta) = \theta - X$. À partir d'un état initial $\theta_0 \in \mathbb{R}^d$ arbitraire et d'une suite $(X_n)_{n\in\mathbb{N}}$ de réalisations indépendantes et de même loi que X, nous définissons la suite $(\theta_n)_{n\in\mathbb{N}}$ par récurrence

$$\forall n \in \mathbb{N},\ \theta_{n+1} = \theta_n - \gamma_{n+1} H(X_{n+1}, \theta_n) \tag{3.7}$$

où les termes déterministes de la suite $(\gamma_n)_{n\geqslant 1}$ à valeurs dans $\mathbb{R}_+^*$ sont appelés les *pas*.

ALGORITHME 3.12 – Robbins-Monro

Initialisation :
- une fonction H telle que $h(\theta) = \mathbb{E}[H(X, \theta)] \in \mathbb{R}^d$ pour tout $\theta \in \mathbb{R}^d$
- une suite de pas $(\gamma_n)_{n\geqslant 1}$ à valeurs dans $\mathbb{R}_+^*$
- un état initial $\theta_0 \in \mathbb{R}^d$

À l'étape $n \geqslant 1$:
  Tirer X_n selon la loi de X
  Calculer $\theta_n = \theta_{n-1} - \gamma_n H(X_n, \theta_{n-1})$

Retourner les états $\theta_0, \theta_1, \theta_2, \ldots$

La suite de variables aléatoires $(X_n)_{n\in\mathbb{N}}$ induit une filtration naturelle $(\mathscr{F}_n)_{n\in\mathbb{N}}$ comme dans (3.5) et, par construction, la suite $(\theta_n)_{n\in\mathbb{N}}$ est adaptée à cette filtration. Nous pouvons également remarquer que θ_0 peut être une variable aléatoire $\mathscr{F}_0$-mesurable. La récurrence (3.7) se réécrit alors

$$\forall n \in \mathbb{N},\ \theta_{n+1} = \theta_n - \gamma_{n+1} h(\theta_n) + \gamma_{n+1} \Delta M_{n+1} \tag{3.8}$$

où les $\Delta M_{n+1} = h(\theta_n) - H(X_{n+1}, \theta_n)$ sont des incréments de martingale car ils vérifient

$$\forall n \in \mathbb{N},\ \mathbb{E}[\Delta M_{n+1} \mid \mathscr{F}_n] = h(\theta_n) - \mathbb{E}[H(X_{n+1}, \theta_n) \mid \mathscr{F}_n] = 0.$$

Autrement dit, l'algorithme de Robbins-Monro est une méthode du point fixe perturbée par un incrément de martingale. La forme (3.8) est parfois utilisée comme définition d'une classe d'algorithme plus générale à laquelle un terme de reste peut être intégré tant que celui-ci demeure négligeable par rapport à l'incrément de martingale. Par souci de simplicité, nous ne

considérerons pas de termes de reste dans la suite mais le lecteur intéressé est invité à lire [**?**] pour une étude plus générale de cette famille d'algorithmes stochastiques.

L'algorithme 3.12 ne converge pas systématiquement et nous donnons ci-dessous un cadre théorique dans lequel il est possible d'établir certains résultats théoriques à son sujet. Avant de formuler les hypothèses nécessaires pour cela, nous attirons l'attention sur l'importance des pas $(\gamma_n)_{n\geqslant 1}$ dans la méthode de Robbins-Monro. Nous définissons la suite $(t_n)_{n\in\mathbb{N}}$ strictement croissante par $t_0 = 0$ et la récurrence

$$\forall n \geqslant 1,\ t_n = \sum_{k=0}^{n} \gamma_k.$$

Soit $n \in \mathbb{N}$, nous pouvons considérer θ_n comme la valeur au temps t_n d'une fonction $\theta : \mathbb{R}_+ \to \mathbb{R}^d$ définie par interpolation linéaire,

$$\forall n \in \mathbb{N},\ \forall t \in [t_n, t_{n+1}],\ \theta(t) = \theta_n + (\theta_{n+1} - \theta_n) \times \frac{t - t_n}{t_{n+1} - t_n}.$$

Pour que cette fonction θ soit correctement définie sur tout $\mathbb{R}_+$, il faut supposer que les pas vérifient

$$\sum_{n\geqslant 1} \gamma_n = +\infty. \tag{3.9}$$

La récurrence sans les perturbations est telle que, pour tout $n \in \mathbb{N}$,

$$\theta_{n+1} = \theta_n - \gamma_{n+1} h(\theta_n) \iff \frac{\theta(t_{n+1}) - \theta(t_n)}{t_{n+1} - t_n} = -h(\theta(t_n))$$

et correspond donc à une méthode d'Euler explicite pour l'équation différentielle ordinaire

$$\frac{d\theta(t)}{dt} = -h(\theta(t)).$$

Cette approximation sans bruit de (3.8) peut se justifier si l'accumulation des perturbations $\gamma_1 \Delta M_1 + \cdots + \gamma_n \Delta M_n$ jusqu'au temps $n \geqslant 1$ n'est pas « trop grande », par exemple, au sens de la variance qui vaut, par orthogonalité des incréments,

$$\sum_{k=1}^{n} \gamma_k^2 \mathbb{E}\left[\|\Delta M_k\|^2\right].$$

Si la variance des incréments de martingale reste bornée, nous voyons ici apparaître la condition suivante sur la suite des pas nécessaire pour envisager un résultat de convergence de l'algorithme de Robbins-Monro,

$$\sum_{n\geqslant 1} \gamma_n^2 < +\infty. \tag{3.10}$$

Compte tenu de (3.9) et (3.10), la suite de pas donnée par $\gamma_n = 1/n$ que nous avons considérée dans la sous-section 1.4.2 semble satisfaisante et nous discuterons plus loin d'autres formes de pas.

Le lien avec la théorie des équations différentielles ordinaires que nous avons établi ci-dessus met à notre disposition des techniques dédiées pour étudier la méthode de Robbins-Monro. Cette relation dépasse le cadre de ce cours mais elle permet de comprendre l'introduction des *fonctions de Lyapounov* comme cadre théorique pour obtenir des résultats de convergence.

Définition 3.13. Soit $h : \mathbb{R}^d \to \mathbb{R}^d$, une fonction $V : \mathbb{R}^d \to \mathbb{R}_+$ différentiable est appelée *fonction de Lyapounov* associée à h si elle vérifie les conditions suivantes :

- *Régularité* : il existe $L > 0$ tel que le gradient ∇V de V soit L-Lipschitz,

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d,\ \|\nabla V(\theta_1) - \nabla V(\theta_2)\| \leqslant L\|\theta_1 - \theta_2\|,$$

- *Contrôle* : il existe $C > 0$ tel que

$$\forall \theta \in \mathbb{R}^d,\ \|h(\theta)\|^2 \leqslant C(1 + V(\theta)),$$

- *Condition du gradient* : il existe $\alpha > 0$ tel que

$$\forall \theta \in \mathbb{R}^d,\ \langle \nabla V(\theta), h(\theta)\rangle \geqslant \alpha\|\nabla V(\theta)\|^2.$$

L'existence d'une fonction de Lyapounov peut être difficile à justifier en toute généralité mais elle permet d'obtenir le résultat suivant sur la convergence de l'algorithme de Robbins-Monro.

Théorème 3.14. *Soit $h : \mathbb{R}^d \to \mathbb{R}^d$, nous considérons la suite de variables aléatoires $(\theta_n)_{n\in\mathbb{N}}$ donnée par (3.8) pour une suite de pas $(\gamma_n)_{n\geqslant 1}$ strictement positifs tels que*

$$\sum_{n\geqslant 1} \gamma_n = +\infty \quad \text{et} \quad \sum_{n\geqslant 1} \gamma_n^2 < +\infty.$$

Si h admet une fonction de Lyapounov V et qu'il existe $\kappa > 0$ tel que les incréments de martingales vérifient

$$\forall n \in \mathbb{N},\ \mathbb{E}[\|\Delta M_{n+1}\|^2 \mid \mathscr{F}_n] \leqslant \kappa(1 + V(\theta_n)).$$

Alors, il existe une variable aléatoire V_∞ intégrable telle que

$$V(\theta_n) \xrightarrow[n\to+\infty]{p.s.} V_\infty \quad \text{et} \quad \sum_{n\geqslant 0} \gamma_{n+1}\langle \nabla V(\theta_n), h(\theta_n)\rangle < +\infty.$$

Démonstration. Soit $n \in \mathbb{N}$, par intégration de la dérivée de $t \in [0,1] \mapsto V(t\theta_n + (1-t)\theta_{n+1})$, nous avons

$$\begin{aligned}
V(\theta_{n+1}) &= V(\theta_n) + \int_0^1 \langle \nabla V(t\theta_n + (1-t)\theta_{n+1}), \theta_{n+1} - \theta_n\rangle dt \\
&= V(\theta_n) + \langle \nabla V(\theta_n), \theta_{n+1} - \theta_n\rangle \\
&\quad + \int_0^1 \langle \nabla V(t\theta_n + (1-t)\theta_{n+1}) - \nabla V(\theta_n), \theta_{n+1} - \theta_n\rangle dt \\
&= V(\theta_n) - \gamma_{n+1}\langle \nabla V(\theta_n), h(\theta_n)\rangle + \gamma_{n+1}\langle \nabla V(\theta_n), \Delta M_{n+1}\rangle \\
&\quad + \int_0^1 \langle \nabla V(t\theta_n + (1-t)\theta_{n+1}) - \nabla V(\theta_n), \theta_{n+1} - \theta_n\rangle dt.
\end{aligned}$$

Puisque le gradient ∇V est L-Lipschitz, l'intégrale admet la borne supérieure suivante

$$\begin{aligned}
&\int_0^1 \langle \nabla V(t\theta_n + (1-t)\theta_{n+1}) - \nabla V(\theta_n), \theta_{n+1} - \theta_n \rangle dt \\
&\leqslant 2\int_0^1 \|\nabla V(t\theta_n + (1-t)\theta_{n+1}) - \nabla V(\theta_n)\|^2 dt + 2\|\theta_{n+1} - \theta_n\|^2 \\
&\leqslant 2(L^2+1)\|\theta_{n+1} - \theta_n\|^2 \\
&\leqslant 4(L^2+1)\left(\|h(\theta_n)\|^2 + \|\Delta M_{n+1}\|^2\right).
\end{aligned}$$

Comme $\|h(\theta_n)\|^2 \leqslant C(1+V(\theta_n))$, nous déduisons l'inégalité

$$\begin{aligned}
V(\theta_{n+1}) \leqslant (1+C_1\gamma_{n+1}^2)V(\theta_n) - \gamma_{n+1}\langle \nabla V(\theta_n), h(\theta_n)\rangle + \gamma_{n+1}\langle \nabla V(\theta_n), \Delta M_{n+1}\rangle \\
+ 4(L^2+1)\gamma_{n+1}^2(C + \|\Delta M_{n+1}\|^2).
\end{aligned}$$

où $C_1 = 4C(L^2+1)$. En prenant l'espérance conditionnelle à $\mathscr{F}_n$, il vient que

$$\mathbb{E}[V(\theta_{n+1}) \mid \mathscr{F}_n] \leqslant (1+C_2\gamma_{n+1}^2)V(\theta_n) + 4(C+\kappa)(L^2+1)\gamma_{n+1}^2 - \gamma_{n+1}\langle \nabla V(\theta_n), h(\theta_n)\rangle$$

où $C_2 = C' + 4\kappa(L^2+1)$. Le théorème 3.11 s'applique par positivité de $\langle \nabla V(\theta_n), h(\theta_n)\rangle$ et nous obtenons les conclusions annoncées. □

Lorsque la méthode de Robbins-Monro est appliquée au gradient $h = \nabla f$ d'une fonction $f : \mathbb{R}^d \to \mathbb{R}$ différentiable, cela conduit à un algorithme d'optimisation stochastique appelé *descente de gradient stochastique*. En écrivant le gradient de f sous la forme

$$\forall \theta \in \mathbb{R}^d,\ \nabla f(\theta) = \mathbb{E}[\nabla F(X,\theta)]$$

pour une variable aléatoire X et une fonction $F(X,\cdot)$ connue et différentiable en θ, nous obtenons la récurrence

$$\forall n \in \mathbb{N},\ \theta_{n+1} = \theta_n - \gamma_{n+1}\nabla F(X_{n+1}, \theta_n)$$

où $(X_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires indépendantes de même loi que X, $(\gamma_n)_{n\geqslant 1}$ est une suite de pas strictement positifs et $\theta_0 \in \mathbb{R}^d$ est un point de départ arbitraire.

ALGORITHME 3.15 – Descente de gradient stochastique

Initialisation :
- une fonction F telle que $\nabla f(\theta) = \mathbb{E}[\nabla F(X,\theta)] \in \mathbb{R}^d$ pour tout $\theta \in \mathbb{R}^d$
- une suite de pas $(\gamma_n)_{n\geqslant 1}$ à valeurs dans $\mathbb{R}_+^*$
- un état initial $\theta_0 \in \mathbb{R}^d$

À l'étape $n \geqslant 1$:
Tirer X_n selon la loi de X
Calculer $\theta_n = \theta_{n-1} - \gamma_n \nabla F(X_n, \theta_{n-1})$

Retourner les états $\theta_0, \theta_1, \theta_2, \ldots$

Dans le cas de la descente de gradient stochastique, si la fonction $f : \mathbb{R}^d \to \mathbb{R}$ le permet (*i.e.* gradient L-Lipschitz, ...), alors il est possible de considérer la fonction de Lyapounov donnée par

$$\forall \theta \in \mathbb{R}^d,\ V(\theta) = f(\theta) - \inf_{t\in\mathbb{R}^d} f(t).$$

Ce choix n'est pas nécessairement le meilleur possible mais il reste une option par défaut même quand la fonction f n'est pas convexe. En appliquant le théorème 3.14, nous obtenons que $f(\theta_n)$ converge presque sûrement vers une variable aléatoire intégrable et

$$\sum_{n\geqslant 0} \gamma_{n+1}\|\nabla f(\theta_n)\|^2 < +\infty.$$

Puisque la somme des γ_n diverge, nous pouvons en déduire que

$$\liminf_{n\to\infty} \nabla f(\theta_n) = 0, \ p.s.$$

mais pas que $\nabla f(\theta_n)$ converge presque sûrement vers 0, *i.e.* nous ne pouvons pas déduire la convergence presque sûre de θ_n vers un point critique de f. Pour obtenir un tel résultat, il faut faire des hypothèses additionnelles telles que l'unicité du minimiseur dans le cas d'une fonction convexe, par exemple.

3.2.3 Moyennes de Polyak-Ruppert

Bien que cela n'apparaisse pas dans les résultats précédents, il est possible de montrer que la convergence de l'algorithme de Robbins-Monro dépend fortement du choix des pas $(\gamma_n)_{n\geqslant 1}$. Cette dépendance est difficile à illustrer sans entrer dans des détails techniques qui dépassent le cadre de cours mais elle n'est pas désirable d'un point de vue pratique et peut détériorer la vitesse de convergence de la suite $(\theta_n)_{n\in\mathbb{N}}$. Pour contourner cela, Polyak et Juditsky [**?**] et Ruppert [**?**] ont proposé d'utiliser la suite des moyennes $(\bar{\theta}_n)_{n\geqslant 1}$ à la place,

$$\forall n \geqslant 1, \ \bar{\theta}_n = \frac{1}{n}\sum_{k=0}^{n-1} \theta_k.$$

Un atout de cette approche est que, comme pour la descente de gradient, ces moyennes peuvent être calculées en ligne par la récurrence

$$\forall n \geqslant 1, \ \bar{\theta}_{n+1} = \bar{\theta}_n + \frac{1}{n+1}(\theta_n - \bar{\theta}_n)$$

où $\bar{\theta}_1 = \theta_0$.

Comme nous le verrons dans la sous-section suivante, sous certaines hypothèses, la vitesse de convergence de $\bar{\theta}_n$ ne dépend plus du choix des pas $(\gamma_n)_{n\geqslant 1}$ et cette vitesse atteint même parfois la vitesse optimale proportionnelle à $1/n$ donnée par la borne de Cramer-Rao pour d'autres formes de moyennes.

3.2.4 Propriétés non asymptotiques du gradient stochastique

Nous nous plaçons maintenant dans le cadre de la descente de gradient stochastique. Étant donnée une filtration $(\mathscr{F}_n)_{n\in\mathbb{N}}$ (par exemple, la filtration naturelle engendrée par une suite de variables indépendantes de même loi) et une fonction $f : \mathbb{R}^d \to \mathbb{R}$ différentiable à minimiser, nous disposons d'une suite de fonctions aléatoires différentiables $(F_n)_{n\in\mathbb{N}}$ adaptée à la filtration $(\mathscr{F}_n)_{n\in\mathbb{N}}$ telle que

$$\forall n \in \mathbb{N}, \ \forall \theta \in \mathbb{R}^d, \ \mathbb{E}[\nabla F_{n+1}(\theta) \mid \mathscr{F}_n] = \nabla f(\theta), \ p.s.$$

Nous supposerons également ici que la fonction f est convexe, c'est-à-dire qu'elle vérifie

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d,\ \forall \lambda \in [0,1],\ f(\lambda\theta_1 + (1-\lambda)\theta_2) \leqslant \lambda f(\theta_1) + (1-\lambda) f(\theta_2).$$

Cette hypothèse permet d'assurer la convergence de la descente de gradient vers le minimum global de la fonction comme nous l'avons évoqué ci-dessus. Elle est également nécessaire pour obtenir la convergence de la descente de gradient déterministe (voir [**?**]). Dans la suite, il sera plus simple de travailler avec propriété suivante pour manipuler la convexité de f.

Proposition 3.16. *Soit $f : \mathbb{R}^d \to \mathbb{R}$ une fonction différentiable, alors f est convexe si et seulement si*

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d,\ f(\theta_1) \geqslant f(\theta_2) + \langle \nabla f(\theta_2), \theta_1 - \theta_2 \rangle. \tag{3.11}$$

Démonstration. Soit $\theta_1, \theta_2 \in \mathbb{R}^d$, nous introduisons la fonction g définie par

$$\forall \lambda \in [0,1],\ g(\lambda) = f(\lambda\theta_1 + (1-\lambda)\theta_2) - \lambda f(\theta_1) - (1-\lambda) f(\theta_2).$$

Comme f est différentaible, il en va de même pour g et nous obtenons

$$g'(\lambda) = \langle \nabla f(\lambda\theta_1 + (1-\lambda)\theta_2), \theta_1 - \theta_2 \rangle - f(\theta_1) + f(\theta_2).$$

($\Rightarrow$) Si f est convexe, alors $g(\lambda) \leqslant 0$ pour tout $\lambda \in [0,1]$. Puisque $g(0) = 0$, nous déduisons que

$$g'(0) = \lim_{\lambda \to 0} \frac{g(\lambda)}{\lambda} \leqslant 0,$$

ce qui correspond à (3.11).
($\Leftarrow$) Si f satisfait (3.11), alors nous avons

$$f(\theta_1) \geqslant f(\lambda\theta_1 + (1-\lambda)\theta_2) + (1-\lambda)\langle \nabla f(\lambda\theta_1 + (1-\lambda)\theta_2), \theta_1 - \theta_2 \rangle$$

et

$$f(\theta_2) \geqslant f(\lambda\theta_1 + (1-\lambda)\theta_2) - \lambda \langle \nabla f(\lambda\theta_1 + (1-\lambda)\theta_2), \theta_1 - \theta_2 \rangle.$$

Ces deux inégalités nous conduisent à

$$\lambda f(\theta_1) + (1-\lambda) f(\theta_2) = f(\lambda\theta_1 + (1-\lambda)\theta_2),$$

ce qui montre que f est convexe. □

En pratique, il est fréquent que la recherche du minimiseur de f ne se fasse pas dans $\mathbb{R}^d$ entier mais sur une boule $\mathscr{B}_r$ de rayon $r > 0$ potentiellement très grand,

$$\mathscr{B}_r = \left\{ \theta \in \mathbb{R}^d \text{ tel que } \|\theta\| \leqslant r \right\}.$$

Dans ce cas, il est possible de contraindre la descente de gradient stochastique à rester dans $\mathscr{B}_r$ en projetant chaque itération à l'aide de la projection orthogonale Π_r dans la boule de rayon $r > 0$. Cela conduit à définir une variante de l'algorithme 3.15 appelée *descente de gradient stochastique projeté* et donnée par la récurrence

$$\forall n \in \mathbb{N}, \theta_{n+1} = \Pi_r\left(\theta_n - \gamma_{n+1} \nabla F_{n+1}(\theta_n)\right) \tag{3.12}$$

où $\theta_0 \in \mathscr{B}_r$ est un état initial arbitraire.

ALGORITHME 3.17 – Descente de gradient stochastique projeté (cas i.i.d.)
Initialisation :
- une fonction F telle que $\nabla f(\theta) = \mathbb{E}[\nabla F(X,\theta)] \in \mathbb{R}^d$ pour tout $\theta \in \mathbb{R}^d$
- une suite de pas $(\gamma_n)_{n\geqslant 1}$ à valeurs dans $\mathbb{R}_+^*$ et un état initial $\theta_0 \in \mathbb{R}^d$

À l'étape $n \geqslant 1$:
Tirer X_n selon la loi de X
Calculer $\theta_n = \Pi_r(\theta_{n-1} - \gamma_n \nabla F(X_n, \theta_n))$
Retourner les états $\theta_0, \theta_1, \theta_2, \ldots$

Théorème 3.18. *Soit $r > 0$, nous considérons une fonction $f : \mathbb{R}^d \to \mathbb{R}$ différentiable et convexe qui atteint son minimum global en un point $\theta^* \in \mathscr{B}_r$. Si les fonctions $(F_n)_{n\in\mathbb{N}}$ vérifient*

$$\exists B > 0,\ \forall n \in \mathbb{N},\ \sup_{\theta\in\mathscr{B}_r} \|\nabla F_n(\theta)\| \leqslant B,\ p.s.$$

alors la descente de gradient stochastique projecté $(\theta_n)_{n\in\mathbb{N}}$ definie par (3.12) avec

$$\forall n \geqslant 1,\ \gamma_n = \frac{r\sqrt{2}}{B\sqrt{n}}$$

est telle que

$$\forall n \geqslant 1,\ \mathbb{E}\left[f(\bar{\theta}_n) - f(\theta^*)\right] \leqslant \frac{2\sqrt{2}Br}{\sqrt{n}}$$

où

$$\forall n \geqslant 1,\ \bar{\theta}_n = \frac{1}{n}\sum_{k=0}^{n-1} \theta_k.$$

Démonstration. Soit $n \in \mathbb{N}$, par définition de la descente de gradient stochastique projeté, nous avons

$$\begin{aligned}\|\theta_{n+1} - \theta^*\|^2 &= \|\Pi_r(\theta_n - \gamma_{n+1}\nabla F_{n+1}(\theta_n)) - \theta^*\|^2 \\ &\leqslant \|(\theta_n - \theta^*) - \gamma_{n+1}\nabla F_{n+1}(\theta_n)\|^2 \\ &= \|\theta_n - \theta^*\|^2 + \gamma_{n+1}^2\|\nabla F_{n+1}(\theta_n)\|^2 - 2\gamma_{n+1}\langle \nabla F_{n+1}(\theta_n), \theta_n - \theta^*\rangle.\end{aligned}$$

La borne supérieure sur le gradient ∇F_{n+1} conduit à

$$\|\theta_{n+1} - \theta^*\|^2 \leqslant \|\theta_n - \theta^*\|^2 + \gamma_{n+1}^2 B^2 - 2\gamma_{n+1}\langle \nabla F_{n+1}(\theta_n), \theta_n - \theta^*\rangle.$$

En prenant l'espérance conditionnelle par rapport à la tribu $\mathscr{F}_n$, la proposition 3.16 nous permet de déduire que

$$\begin{aligned}\mathbb{E}\left[\|\theta_{n+1} - \theta^*\|^2 \mid \mathscr{F}_n\right] &\leqslant \|\theta_n - \theta^*\|^2 + \gamma_{n+1}^2 B^2 - 2\gamma_{n+1}\langle \nabla f(\theta_n), \theta_n - \theta^*\rangle \\ &\leqslant \|\theta_n - \theta^*\|^2 + \gamma_{n+1}^2 B^2 - 2\gamma_{n+1}(f(\theta_n) - f(\theta^*)).\end{aligned}$$

Pour l'espérance globale, cette inégalité nous donne

$$\mathbb{E}\left[\|\theta_{n+1} - \theta^*\|^2\right] \leqslant \mathbb{E}\left[\|\theta_n - \theta^*\|^2\right] + \gamma_{n+1}^2 B^2 - 2\gamma_{n+1}\mathbb{E}\left[f(\theta_n) - f(\theta^*)\right],$$

ce qui est équivalent à

$$\mathbb{E}\left[f(\theta_n) - f(\theta^*)\right] \leqslant \frac{\gamma_{n+1}B^2}{2} + \frac{1}{2\gamma_{n+1}}\left(\mathbb{E}\left[\|\theta_n - \theta^*\|^2\right] - \mathbb{E}\left[\|\theta_{n+1} - \theta^*\|^2\right]\right).$$

Pour $n \geqslant 1$, la somme de ces inégalités se télescope,

$$\begin{aligned}\sum_{k=0}^{n-1} \mathbb{E}\left[f(\theta_k) - f(\theta^*)\right] \leqslant \frac{B^2}{2}\sum_{k=0}^{n-1}\gamma_{k+1} + \frac{\mathbb{E}\left[\|\theta_0 - \theta^*\|^2\right]}{2\gamma_1} - \frac{\mathbb{E}\left[\|\theta_n - \theta^*\|^2\right]}{2\gamma_n} \\ + \frac{1}{2}\sum_{k=1}^{n-1}\left(\frac{1}{\gamma_{k+1}} - \frac{1}{\gamma_k}\right)\mathbb{E}\left[\|\theta_k - \theta^*\|^2\right],\end{aligned}$$

et la décroissance des pas $(\gamma_n)_{n\geqslant 1}$ avec le fait que $\|\theta_k - \theta^*\|^2 \leqslant 4r^2$ impliquent

$$\sum_{k=0}^{n-1} \mathbb{E}\left[f(\theta_k) - f(\theta^*)\right] \leqslant \frac{B^2}{2}\sum_{k=0}^{n-1}\gamma_{k+1} + \frac{2r^2}{\gamma_1} + \frac{2r^2}{\gamma_n} - \frac{2r^2}{\gamma_1} = \frac{B^2}{2}\sum_{k=0}^{n-1}\gamma_{k+1} + \frac{2r^2}{\gamma_n}.$$

Pour $\gamma_n = \dfrac{r\sqrt{2}}{B\sqrt{n}}$, nous déduisons de la convexité de f et de l'inégalité de Jensen que

$$\begin{aligned}\mathbb{E}\left[f(\bar{\theta}_n) - f(\theta^*)\right] &\leqslant \frac{1}{n}\sum_{k=0}^{n-1}\mathbb{E}\left[f(\theta_{k-1}) - f(\theta^*)\right] \\ &\leqslant \frac{Br}{n\sqrt{2}}\sum_{k=0}^{n-1}\frac{1}{\sqrt{k+1}} + \frac{\sqrt{2}Br}{\sqrt{n}}\end{aligned}$$

ce qui conduit au résultat annoncé avec la borne élémentaire suivante

$$\sum_{k=0}^{n-1}\frac{1}{\sqrt{k+1}} \leqslant \int_0^n \frac{dx}{\sqrt{x}} = 2\sqrt{n}.$$

□

Sous la seule hypothèse de convexité, la vitesse en $1/\sqrt{n}$ que nous obtenons est optimale au sens du minimax. Cependant, comme dans le cas déterministe, la vitesse de convergence de la descente de gradient stochastique projeté peut être améliorée en supposant la forte convexité de la fonction f, *i.e.* l'existence de $\mu > 0$ tel que la fonction ϕ définie par

$$\forall \theta \in \mathbb{R}^d,\ \phi(\theta) = f(\theta) - \frac{\mu}{2}\|\theta\|^2 \tag{3.13}$$

soit convexe.

Théorème 3.19. *Soit $r > 0$, nous considérons une fonction $f : \mathbb{R}^d \to \mathbb{R}$ différentiable et μ-fortement convexe qui atteint son minimum global en un point $\theta^* \in \mathscr{B}_r$. Si les fonctions $(F_n)_{n\in\mathbb{N}}$ vérifient*

$$\exists B > 0,\ \forall n \in \mathbb{N},\ \sup_{\theta\in\mathscr{B}_r} \|\nabla F_n(\theta)\| \leqslant B,\ p.s.$$

alors la descente de gradient stochastique projeté $(\theta_n)_{n\in\mathbb{N}}$ definie par (3.12) est telle que

- *si $\gamma_n = \dfrac{1}{\mu n}$ pour tout $n \geqslant 1$, alors*

$$\forall n \geqslant 1,\ \mathbb{E}\left[f(\bar{\theta}_n) - f(\theta^*)\right] \leqslant \frac{B^2(1+\ln(n))}{2\mu n} \quad où \quad \bar{\theta}_n = \frac{1}{n}\sum_{k=0}^{n-1}\theta_k.$$

- *si $\gamma_n = \dfrac{2}{\mu(n+1)}$ pour tout $n \geqslant 1$, alors*

$$\forall n \geqslant 1,\ \mathbb{E}\left[f(\tilde{\theta}_n) - f(\theta^*)\right] \leqslant \frac{2B^2}{\mu(n+1)} \quad où \quad \tilde{\theta}_n = \frac{2}{n(n+1)}\sum_{k=0}^{n-1}(k+1)\theta_k.$$

Démonstration. Soit $n \in \mathbb{N}$, comme dans le cas convexe, nous obtenons

$$\mathbb{E}\left[\|\theta_{n+1} - \theta^*\|^2 \mid \mathscr{F}_n\right] \leqslant \|\theta_n - \theta^*\|^2 + \gamma_{n+1}^2 B^2 - 2\gamma_{n+1}\langle \nabla f(\theta_n), \theta_n - \theta^*\rangle.$$

Soit la fonction convexe ϕ donnée par (3.13), nous avons $\nabla\phi(\theta) = \nabla f(\theta) - \mu\theta$ pour tout $\theta \in \mathbb{R}^d$ et la proposition 3.16 implique

$$\begin{aligned}\langle \nabla f(\theta_n), \theta_n - \theta^*\rangle &\geqslant \phi(\theta_n) - \phi(\theta^*) - \mu\langle \theta_n, \theta^* - \theta_n\rangle \\ &= f(\theta_n) - f(\theta^*) + \frac{\mu}{2}\|\theta_n - \theta^*\|^2.\end{aligned}$$

Ainsi, nous avons la relation

$$\mathbb{E}\left[\|\theta_{n+1} - \theta^*\|^2 \mid \mathscr{F}_n\right] \leqslant (1 - \mu\gamma_{n+1})\|\theta_n - \theta^*\|^2 + \gamma_{n+1}^2 B^2 - 2\gamma_{n+1}(f(\theta_n) - f(\theta^*)).$$

En prenant l'espérance de chaque côté de cette inégalite, nous déduisons

$$\mathbb{E}\left[\|\theta_{n+1} - \theta^*\|^2\right] \leqslant (1 - \mu\gamma_{n+1})\mathbb{E}\left[\|\theta_n - \theta^*\|^2\right] + \gamma_{n+1}^2 B^2 - 2\gamma_{n+1}\mathbb{E}\left[f(\theta_n) - f(\theta^*)\right]$$

ce qui est équivalent à

$$\mathbb{E}\left[f(\theta_n) - f(\theta^*)\right] \leqslant \frac{\gamma_{n+1}B^2}{2} + \frac{1 - \mu\gamma_{n+1}}{2\gamma_{n+1}}\mathbb{E}\left[\|\theta_n - \theta^*\|^2\right] - \frac{1}{2\gamma_{n+1}}\mathbb{E}\left[\|\theta_{n+1} - \theta^*\|^2\right] \quad (3.14)$$

Prenons les pas $\gamma_n = 1/(\mu n)$ pour tout $n \geqslant 1$, l'inégalité (3.14) devient

$$\mathbb{E}\left[f(\theta_n) - f(\theta^*)\right] \leqslant \frac{B^2}{2\mu(n+1)} + \frac{\mu n}{2}\mathbb{E}\left[\|\theta_n - \theta^*\|^2\right] - \frac{\mu(n+1)}{2}\mathbb{E}\left[\|\theta_{n+1} - \theta^*\|^2\right].$$

Pour $n \geqslant 1$, la somme de ces inégalités se télescope,

$$\sum_{k=0}^{n-1} \mathbb{E}[f(\theta_k) - f(\theta^*)] \leqslant \frac{B^2}{2\mu} \sum_{k=0}^{n-1} \frac{1}{k+1} - \frac{\mu n}{2} \mathbb{E}\left[\|\theta_n - \theta^*\|^2\right] \leqslant \frac{B^2}{2\mu} \sum_{k=0}^{n-1} \frac{1}{k+1}.$$

La convexité de f et l'inégalité de Jensen impliquent alors

$$\mathbb{E}\left[f(\bar{\theta}_n) - f(\theta^*)\right] \leqslant \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{E}[f(\theta_k) - f(\theta^*)] \leqslant \frac{B^2}{2\mu n} \sum_{k=0}^{n-1} \frac{1}{k+1}$$

ce qui donne le résultat annoncé puisque nous savons que

$$\sum_{k=0}^{n-1} \frac{1}{k+1} \leqslant 1 + \int_1^n \frac{dx}{x} \leqslant 1 + \ln(n).$$

Nous considérons maintenant les pas $\gamma_n = 2/(\mu(n+1))$ pour tout $n \geqslant 1$, l'inégalité (3.14) nous conduit à

$$\mathbb{E}[f(\theta_n) - f(\theta^*)] \leqslant \frac{B^2}{\mu(n+2)} + \frac{\mu n}{4} \mathbb{E}\left[\|\theta_n - \theta^*\|^2\right] - \frac{\mu(n+2)}{4} \mathbb{E}\left[\|\theta_{n+1} - \theta^*\|^2\right]$$

Comme ci-dessus, nous en déduisons

$$\sum_{k=0}^{n-1} (k+1)\mathbb{E}[f(\theta_k) - f(\theta^*)] \leqslant \frac{B^2}{\mu} \sum_{k=0}^{n-1} \frac{k+1}{k+2} - \frac{\mu n(n+1)}{4} \mathbb{E}\left[\|\theta_n - \theta^*\|^2\right] \leqslant \frac{B^2 n}{\mu}$$

et le même argument de convexité de f donne

$$\mathbb{E}\left[f(\tilde{\theta}_n) - f(\theta^*)\right] \leqslant \frac{2}{n(n+1)} \sum_{k=0}^{n-1} (k+1)\mathbb{E}[f(\theta_k) - f(\theta^*)] \leqslant \frac{2B^2}{\mu(n+1)}.$$

□

3.2.5 Application : régression logistique

La *régression logistique* est un modèle statistique largement utilisé en apprentissage automatique pour expliquer une variable binaire $Y \in \{0,1\}$ à partir d'un vecteur $X \in \mathbb{R}^d$ de variables quantitatives. L'étape de modélisation consiste à considérer que Y sachant $X = x$ suit une loi de Bernoulli de paramètre $p(x)$ de la forme

$$\forall x \in \mathbb{R}^d,\ p(x) = \frac{\exp(x'\beta)}{1 + \exp(x'\beta)}$$

où les coefficients $\beta \in \mathbb{R}^d$ sont les paramètres du modèle. Étant données des observations $(x_1, y_1), \ldots, (x_N, y_N) \in \mathbb{R}^d \times \{0,1\}$, l'approche usuelle pour estimer β consiste à maximiser la vraisemblance

$$\forall \beta \in \mathbb{R}^d,\ L_N(\beta) = \prod_{k=1}^{N} \mathbb{P}(Y = y_k \mid X = x_k) = \prod_{k=1}^{N} p(x_k)^{y_k} (1 - p(x_k))^{1-y_k}.$$

De façon équivalente, nous chercherons à minimiser la fonctions suivante,

$$\begin{aligned}\forall \beta \in \mathbb{R}^d,\ -\frac{1}{N}\ln\left(L_N(\beta)\right) &= -\frac{1}{N}\sum_{k=1}^{N} y_k \ln(p(x_k)) + (1-y_k)\ln(1-p(x_k)) \\ &= \frac{1}{N}\sum_{k=1}^{N} \ln\left(1+\exp\left((1-2y_k)x_k'\beta\right)\right) \\ &= \mathbb{E}\left[F(x_K, y_K, \beta)\right]\end{aligned}$$

où K suit la loi uniforme sur $\{1,\dots,N\}$ et la fonction F est donnée par

$$\forall (x,y) \in \mathbb{R}^d \times \{0,1\},\ \forall \beta \in \mathbb{R}^d,\ F(x,y,\beta) = \ln\left(1+\exp\left((1-2y)x'\beta\right)\right).$$

Il n'existe pas de forme explicite pour la solution de ce problème de minimisation. Cependant, un simple exercice de calcul montre que F est convexe par rapport à β, ce qui nous donne l'existence d'un minimiseur global.

Pour résoudre ce problème d'optimisation, il est possible d'utiliser une descente de gradient stochastique. Le gradient de F se calcule simplement,

$$\forall (x,y) \in \mathbb{R}^d \times \{0,1\},\ \forall \beta \in \mathbb{R}^d,\ \nabla_\beta F(x,y,\beta) = \frac{(1-2y)x}{1+\exp\left(-(1-2y)x'\beta\right)}$$

et, pour une fonction $\sigma : \mathbb{N} \to \{1,\dots,N\}$ à préciser, nous pouvons définir la suite $(\beta_n)_{n\in\mathbb{N}}$ par récurrence

$$\forall n \in \mathbb{N},\ \beta_{n+1} = \beta_n - \gamma_{n+1}\nabla_\beta F(x_{\sigma(n)}, y_{\sigma(n)}, \beta_n)$$

avec une point de départ $\beta_0 \in \mathbb{R}^d$ arbitraire et des pas donnés par

$$\forall n \geqslant 1,\ \gamma_n = \gamma \lceil n/N \rceil^{-\alpha}$$

avec $\gamma > 0$ et $\alpha \in [0,1]$.

Plusieurs choix sont envisageables pour la fonction σ. Une méthode simple consiste à parcourir séquentiellement les observations en revenant au début une fois la dernière observation atteinte,

$$\forall n \in \mathbb{N},\ \sigma(n) = n + 1 - N\lfloor n/N \rfloor.$$

Si l'ordre des données a une structure particulière, cette façon de faire peut introduire un biais. Pour éviter ce problème en pratique, il est préférable de considérer des permutations aléatoires $(\sigma_m)_{m\in\mathbb{N}}$ de $\{1,\dots,N\}$ et de prendre

$$\forall n \in \mathbb{N},\ \sigma(n) = \sigma_{\lfloor n/N \rfloor}(n + 1 - N\lfloor n/N \rfloor).$$

Crédits

• Figure 1.6 : John von Neumann, from period while at Los Alamos National Laboratory, taken from a Los Alamos publication (*Los Alamos : Beginning of an era*, 1943–1945, Los Alamos Scientific Laboratory, 1986).
Domaine public (Terms and Conditions of Use).

• Figure 1.10 : Stanislaw Ulam (vers 1945).
Domaine public (Terms and Conditions of Use).

• Figure 2.4 : Photo of mathematician Andrey Markov.
Domaine public.

• Figure 2.8 : Portrait of American computer scientists Nicholas Metropolis (1915–1999) (seated) and James Henry Richardson (1918–1996) at Los Alamos National Laboratory, Los Alamos, New Mexico, November 1953. (Photo by Loomis Dean/The LIFE Picture Collection/Getty Images).
Rights managed, Getty Images content licence agreement.

• Figure 2.9 : Wilfred Keith Hastings.
Published in Victoria Times Colonist from May 21 to May 22, 2016.

• Figure 2.10 : Frontispiece of *The Scientific Papers of J. Willard Gibbs*, in two volumes, eds. H. A. Bumstead and R. G. Van Name, (London and New York : Longmans, Green, and Co., 1906).
Domaine public.

• Figure 3.5 : Statistician Herbert Robbins visiting Purdue in 1966.
Fair use under the *Copyright law of the United States*.