Introduction aux méthodes de Monte Carlo par dynamique Hamiltonienne

Shmuel RAKOTONIRINA-RICQUEBOURG, Amaury DURAND

1er novembre 2017

Table des matières

1 Introduction

Ce rapport présente le travail effectué lors d'un projet du cours d'approfondissements en chaînes de Markov par Eric Moulines dans le cadre du master Mathématiques de l'aléatoires à l'université Paris-Sud. Le contenu présenté ci-dessous repose essentiellement sur [1] et [2].

Les méthodes de Monte Carlo par dynamique Hamiltonienne, plus communément appelées Hamiltonian Monte Carlo (HMC), font partie d'une grande famille de méthodes de simulation : les méthodes de Monte Carlo, et plus précisemment dans la famille des méthodes de Monte Carlo par chaînes de Markov (ou Monte Carlo Markov Chains). Ces méthodes se placent dans le cadre suivant :

Définition 1.1 (Carde général)**.** Soit $(\mathbb{X}, \mathcal{X})$ un espace mesurable. Soit π une loi de probabilité sur cet espace. On suppose que π n'est connue qu'à un facteur de proportionnalité près i.e on connait $\lambda\pi$ où $\lambda \in \mathbb{R}$ une constante. Le but est de pouvoir approcher $\pi f = \mathbb{E}[f(X)]$ avec $X \sim \pi$ et $f \in \mathbb{F}_+(\mathbb{X}, \mathcal{X}) \cup \mathbb{F}_b(\mathbb{X}, \mathcal{X})$.

1.1 Méthodes de Monte Carlo

La méthode de Monte Carlo la plus simple (appelée Monte Carlo naïf) est la suivante : supposons que l'on sait simuler des variables aléatoires de loi π alors en tirant n échantillons $X_1, \cdots, X_n \overset{\text{iid}}{\sim} \pi$, la loi des grands nombre nous indique qu'une bonne approximation de πf est $\frac{1}{n}\sum_{i=1}^{n} f(X_i)$. D'autres méthodes permettent d'atteindre le même but en ne sachant pas simuler de variables aléatoires de loi π. C'est le cas par exemple de l'échantillonnage d'importance qui consiste à simuler des variables i.i.d sous une loi différente de π.

Dans ces deux cas, l'approximation de πf repose sur une simulation de variables aléatoires i.i.d. Cette particularité permet alors de montrer des résultats de convergence notamment grâce à la loi des grands nombre. Les méthodes de Monte Carlo par chaînes de Markov ne reposent pas sur le caractère i.i.d des variables mais sur les propriétés des Chaînes de Markov.

1.2 Monte Carlo Markov Chains (MCMC)

Les méthode MCMC se basent sur les notions de loi invariante, de réversibilité et d'ergodicité.

1.2.1 Loi invariante et réversibilité

On considère P un noyau de Markov sur $\mathbb{X} \times \mathcal{X}$.

Définition 1.2. Une loi de probabilité π sur $(\mathbb{X}, \mathcal{X})$ est dite

- P-invariante si $\pi P = \pi$
- P-réversible si $\forall A, B \in \mathcal{X}$, $\pi \otimes P(A \times B) = \pi \otimes P(B \times A)$

Proposition 1.1. *Soit π une loi de probabilité sur $(\mathbb{X}, \mathcal{X})$ alors*

$$\pi \text{ est } P\text{-reversible} \Rightarrow \pi \text{ est } P\text{-invariante}$$

1.2.2 Ergodicité

Définition 1.3 (Système ergodique)**.** Soit $(\Omega, \mathcal{B}, \mathbb{P})$ un espace de probabilité. Soit $T : (\Omega, \mathcal{B}) \to (\Omega, \mathcal{B})$ une application mesurable.

- On dit que $\mathbb{P}$ est invariante pour T si $\forall A \in \mathcal{B}, \mathbb{P}\left(T^{-1}(A)\right) = \mathbb{P}(A)$. Dans ce cas, on dit que $(\Omega, \mathcal{B}, \mathbb{P}, T)$ est un système dynamique.
- $A \in \mathcal{B}$ est dit invariant pour T si $A = T^{-1}(A)$.
- Si pour tout A invariant pour T on a $\mathbb{P}(A) \in \{0, 1\}$ alors on dit que $(\Omega, \mathcal{B}, \mathbb{P}, T)$ est un système ergodique.

Théorème 1.1. *Soit P un noyau de Markov sur $\mathbb{X} \times \mathcal{X}$ et on se place sur l'espace canonique $(\mathbb{X}^n, \mathcal{X}^{\otimes n})$. On note* $\theta : \begin{array}{ccc} \mathbb{X}^{\mathbb{N}} & \to & \mathbb{X}^{\mathbb{N}} \\ (\omega_t)_{t\in\mathbb{N}} & \mapsto & (\omega_{t+1})_{t\in\mathbb{N}} \end{array}$ *et $\forall k \in \mathbb{N}^*, \theta_k = \theta_{k-1} \circ \theta$ avec $\theta_0 = Id$. On suppose*

1. *P possède une loi invariante π*
2. *$(\mathbb{X}^{\mathbb{N}}, \mathcal{X}^{\otimes\mathbb{N}}, \mathbb{P}_\pi, \theta)$ est ergodique*

Alors pour toute v.a $Y \in L^1(\mathbb{X}^{\mathbb{N}}, \mathcal{X}^{\otimes\mathbb{N}}, \mathbb{P}_\pi)$, pour π-presque tout $x \in \mathbb{X}$,

$$\frac{1}{n}\sum_{k=0}^{n-1} Y \circ \theta_k \xrightarrow[n\to+\infty]{\mathbb{P}_x\text{- }p.s.} \mathbb{E}_\pi[Y]$$

Remarque 1.1. Considérons $(X_k)_{k\in\mathbb{N}}$ la chaîne de markov canonique de noyau P et $f \in \mathbb{F}_+(\mathbb{X}, \mathcal{X}) \cup \mathbb{F}_b(\mathbb{X}, \mathcal{X})$ alors en prenant $Y = f(X_0)$, on a $Y \circ \theta_k = f(X_k)$ et $\mathbb{E}_\pi[Y] = \mathbb{E}_\pi[f(X_0)] = \pi f$ donc le résultat du théorème 1.1 se réécrit : pour π-presque tout $x \in \mathbb{X}$,

$$\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \xrightarrow[n\to+\infty]{\mathbb{P}_x\text{- p.s.}} \pi f$$

Ce qui montre que l'on peut avoir une bonne approximation de πf en construisant une chaîne de Markov. De plus il est intéressant de constater que la convergence est $\mathbb{P}_x$ presque sûre pour π-presque tout $x \in \mathbb{X}$ ce qui signifie que quelque soit le point de départ, on estsûr d'avoir une bonne approximation de πf si on attend suffisamment longtemps.

Les algorithmes MCMC sont des méthodes permettant de construire la chaîne de Markov $(X_k)_{k\in\mathbb{N}}$ afin d'approcher πf en calculant $\frac{1}{n}\sum_{k=0}^{n-1} f(X_k)$.

1.2.3 Algorithme de Metropolis (Random Walk Metropolis)

Avant de présenter la méthode HMC, nous définissons ici l'algorithme de Métropolis que nous utiliserons comme base de comparaison. On se place dans le cas où $\mathbb{X} = \mathbb{R}^d$, $\mathcal{X} = \mathcal{B}(\mathbb{R}^d)$ et μ est la mesure de Lebesgue.

On suppose que π a une densité h_π par rapport à une mesure μ. On considère de plus un loi Q sur $(\mathbb{X}, \mathcal{X})$ de densité q par rapport à μ telle que $\forall x \in \mathbb{X}, q(x) = q(-x)$. La construction de la chaîne de Markov $(X_k)_{k\in\mathbb{N}}$ se fait par les étapes suivantes :

Algorithme 1.1 : Random Walk Metropolis

Initialization :
$X_0 = x \in \mathbb{X}$ arbitrary

repeat
Propose a motion $Y_{k+1} = X_k + U_{k+1}$ with $(U_k)_{k\in\mathbb{N}} \overset{\text{iid}}{\sim} Q$ and $(U_k)_{k\in\mathbb{N}} \perp\!\!\!\perp (X_k)_{k\in\mathbb{N}}$
Compute $\alpha_{k+1} = \alpha(X_k, Y_{k+1})$ where $\alpha(x, y) = 1 \wedge \frac{h_\pi(y)}{h_\pi(x)}$
Set $X_{k+1} = \begin{cases} Y_{k+1} & \text{with probability } \alpha_{k+1} \\ X_k & \text{with probability } 1 - \alpha_{k+1} \end{cases}$

until *some condition*;

Une explication intuitive de cet algorithme est de voir que l'on cherche à visiter l'espace $\mathbb{X}$ sans pour autant aller dans des régions où h_π est faible (et donc des régions de probabilité faible) car ce sont les régions de forte probabilité qui donnent des informations sur la loi π. Ainsi si le mouvement proposé est tel que $h_\pi(Y_{k+1}) \leq h_\pi(X_k)$ on accepte le mouvement avec probabilité 1. Dans le cas contraire on considère qu'il n'est pas très intéressant de bouger. Néanmoins on autorise quand même un mouvement avec une probabilité de $\frac{h_\pi(Y_{k+1})}{h_\pi(X_k)}$ qui sera d'autant plus faible que la position proposée est dans une région de probabilité faible. Remarquons enfin que le fait de faire un rapport permet de ne connaître π qu'à une constante de proportionnalité près.

2 Hamiltonan Monte Carlo

On se place dans le cas où $\mathbb{X} = \mathbb{R}^d$, $\mathcal{X} = \mathcal{B}(\mathbb{R}^d)$ et μ est la mesure de Lebesgue.

2.1 Intuition

De même que pour l'algorithme de Métropolis, on veut construire une chaîne de Markov qui visite majoritairement les régions de forte probabilité. L'idée de la méthode Hamiltonian Monte Carlo est de considérer que les états $x \in \mathbb{R}^d$ représentent des positions, d'introduire une variable de vitesse $v \in \mathbb{R}^d$, une énergie potentielle $U(x)$ et une énergie cinétique $K(v)$. On considère alors que h_π est de la forme

$$\forall x \in \mathbb{R}^d, h_\pi(x) \propto \exp\left(-\frac{U(x)}{T}\right) \tag{1}$$

$T \in \mathbb{R}^*$ s'appelle la température. Encore une fois on ne connaît h_π qu'à une constante de proportionnalité près.

En dimension 1, le problème a une interprétation physique très simple : générer la chaîne de Markov revient à simuler le mouvement d'un objet sur une rampe sans frottement. Comme l'énergie potentielle est proportionnelle à là hauteur, les creux de la rampe représentent les régions de faible énergie potentielle et donc de forte probabilité par (1). Le rôle de la vitesse v et de l'énergie cinétique $K(v)$ est de modéliser de fait que l'on peut remonter une pente après un creux (si l'énergie cinétique est assez grande).

Références

[1] Neal Radford M. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 54 :113–162, 2010.

[2] Douc Randal ; Moulines Eric ; Priouret Pierre ; Soulier Philippe. *Markov Chains*. 2017.