Semaine 9 - Monte-Carlo par chaînes de Markov (MCMC)

Exercice 1 – Estimation de π par une méthode de Monte Carlo

Les méthodes de Monte Carlo permettent de calculer de manière approchée des intégrales (ou d'estimer des espérances). Pour comprendre comment elles fonctionnent, nous allons les appliquer à un exemple très simple (pour lequel une méthode de Monte Carlo n'est clairement pas la méthode la plus efficace). Ces méthodes de Monte Carlo se révèlent en fait particulièrement utiles et performantes en grande dimension quand le calcul analytique ou les techniques numériques classiques ne sont plus possibles.

Dans cet exercice, on cherche à estimer π par une méthode de Monte Carlo. Dans $\mathbb{R}^2$, considérons le cercle de rayon 1 et le carré $[-1,1]^2$.

Q 1.1 Donner les aires du cercle et du carré.

L'aire du cercle peut être calculée par l'intégrale suivante :

$$\pi = \int_{-1}^{1}\int_{-1}^{1} \mathbb{I}(\sqrt{x^2+y^2} \leq 1)dxdy$$

où $\mathbb{I}(\sqrt{x^2+y^2} \leq 1)$ est la fonction indicatrice qui vaut 1 si le point de coordonnées (x,y) est à une distance inférieure ou égale à 1 de l'origine.

Q 1.2 Soit la loi uniforme $\mathcal{U}([-1,1])$ définie sur l'intervalle $[-1,1]$. Quelle est la fonction de densité $u(\cdot) : [-1,1] \mapsto \mathbb{R}$ de cette distribution de probabilité ? De même, quelle est la fonction de densité de la loi uniforme $\mathcal{U}([-1,1]^2)$ définie sur le carré $[-1,1] \times [-1,1]$?

Q 1.3 En observant que $\mathbb{I}(\sqrt{x^2+y^2} \leq 1) = \mathbb{I}(x^2+y^2 \leq 1)$, montrer que cette aire peut se réécrire comme une espérance en introduisant la densité de loi uniforme $\mathcal{U}([-1,1]^2)$:

$$\pi = 4 \times \mathbb{E}_{X,Y}\big(\mathbb{I}(X^2+Y^2 \leq 1)\big)$$

où X, Y sont deux variables aléatoires de loi uniforme $\mathcal{U}([-1,1])$.

Q 1.4 La valeur π peut donc être estimée par une méthode de Monte Carlo. Expliquer le principe de l'algorithme. Quelle loi justifie cette approche ?

Un des problèmes quand on applique une méthode de Monte Carlo est de savoir quelle valeur choisir pour N, le nombre de tirages aléatoires. La technique suivante, fondée sur une estimation de l'erreur, peut nous aider à choisir approximativement N.
Nous cherchons à estimer pour un certain niveau de confiance l'erreur qui peut être faite par une procédure de Monte Carlo. Notons ε la variable aléatoire définie par :

$$\varepsilon = 4 \times \mathbb{I}(X^2+Y^2 \leq 1) - \pi$$

où $X \sim U([-1,1]), Y \sim U([-1,1])$.

Q 1.5 Quelle interprétation peut-on donner à ε ? Donner l'expression de l'erreur $\overline{\varepsilon}_N$ (en tant que variable aléatoire) faite par la méthode de Monte Carlo précédente et son espérance.

Q 1.6 En supposant connue la variance σ^2 de ε, vers quelle loi tend $\overline{\varepsilon}_N$? Que peut-on dire quand σ^2 n'est pas connue ?

Q 1.7 En exploitant la loi suivie par $\overline{\varepsilon}_N$, déterminez l'intervalle $[-\varepsilon^*, \varepsilon^*]$ pour lequel la probabilité que $\overline{\varepsilon}_N$ soit inférieur à $-\varepsilon^*$ ($P(\overline{\varepsilon}_N \leq -\varepsilon^*)$) ou bien supérieur à ε^* ($P(\overline{\varepsilon}_N \geq \varepsilon^*)$) est égale à un nombre α donné.

Q 1.8 Déduisez-en la valeur de N minimale pour que ε^* soit inférieur à une valeur ε_α fixée (par exemple $\varepsilon_\alpha = 10^{-3}$). L'interprétation de tous ces calculs est la suivante :

1. dans la question précédente, on a choisi un risque d'erreur de 2α, c'est-à-dire que l'on a 2α chances de générer une chaîne de Markov pour laquelle l'erreur entre notre estimation de π et la véritable valeur de π est supérieure à ε^*. En pratique, on se fixe donc un α très petit, souvent de l'ordre de 10^{-2}, de manière à avoir presque la certitude que notre chaîne génère une erreur d'estimation inférieure à ε^*.
2. dans la question actuelle, on fixe en outre l'erreur maximale $\varepsilon^* = \varepsilon_\alpha$ que l'on est prêt à accepter et l'on calcule donc la valeur de N qui nous permet d'assurer cela. Ici, on va donc choisir une valeur de ε_α très petite (elle est souvent plus petite que la valeur de α).
3. en combinant ces deux propriétés, on en déduit que, pour la valeur de N calculée, on a une chance de 2α d'obtenir une chaîne dont l'erreur d'estimation est supérieure à ε_α.

Exercice 2 – Estimation de π par MCMC

Pour appliquer une méthode de Monte Carlo, il est nécessaire de savoir échantillonner selon une certaine loi. Dans l'exercice précédent, on échantillonnait directement selon une loi uniforme. Dans certains problèmes, l'échantillonnage direct peut ne pas être applicable, soit parce qu'on utilise des lois plus complexes pour lesquelles un échantillonnage est difficile voire impossible, soit parce qu'on est en très grandes dimensions. En cours, vous avez vu la méthode d'échantillonnage par rejet. Elle peut se révéler efficace en petites dimensions. Cependant, en très grandes dimensions, il est nécessaire d'échantillonner à l'aide d'une chaîne de Markov en utilisant l'idée suivante.
Soit $S = S_1 \times \cdots \times S_k$ l'ensemble dans lequel on souhaite échantillonner selon la distribution de probabilité $\mathcal{P}$. Supposons qu'on ait une chaîne de Markov dont les états sont S et qui admette une distribution stationnaire. Alors, en partant d'un état initial quelconque, la distribution de l'état tend vers $\mathcal{P}$ après un grand nombre de transitions. Donc, quand on décide de s'arrêter, l'état courant correspond à un tirage selon la distribution voulue $\mathcal{P}$. Une méthode de Monte Carlo où on échantillonne grâce à une chaîne de Markov est appelée méthode de Monte Carlo par chaîne de Markov (MCMC).
Nous allons illustrer le principe d'une méthode MCMC pour estimer π.

Q 2.1 Quel est l'ensemble des états de la chaîne de Markov que nous pourrions utiliser ?

Bien qu'en TME, on fera l'hypothèse que cela ne pose aucun problème technique de travailler avec des chaînes de Markov avec un espace continu d'états, en TD, pour comprendre l'algorithme MCMC avec les outils que nous connaissons, nous allons discrétiser les états pour éviter les problèmes techniques dus à l'espace continu d'états. En prenant une discrétisation suffisamment fine, on pourrait "simuler" le cas de l'espace continu.
On découpe le carré $[-1, 1]^2$ en une grille composée de cases de côté ε (on suppose que $2/\varepsilon \in \mathbb{N}$). Cette grille constituera l'espace d'états de notre chaîne de Markov. Un état est caractérisé par les coordonnées du centre de la case. L'idée est d'échantillonner une case de la grille de manière uniforme et de tester si la case est à l'intérieur du cercle ou non. Ce test peut se faire en vérifiant que le centre de la case tirée est à distance inférieure de 1 de l'origine.
Depuis un état donné (case), une transition possible correspond à un déplacement aléatoire $(\delta_x, \delta_y) \in [-m, m]^2$ où $m \in [0, 1]$. La transition est acceptée si $(x + \delta_x, y + \delta_y) \in [-1, 1]^2$ et on détermine alors la case d'arrivée qui constitue l'état suivant. Si la transition n'est pas acceptée, l'état ne change pas.

Q 2.2 En partant d'une case centrale (où toutes les transitions seraient acceptées), quelle condition sur m garantit une distribution uniforme sur les cases accessibles ?

Par la suite, on suppose que ce choix de m est fait.

Q 2.3 En partant d'une case centrale, quelle distribution de probabilité définit-on sur les cases accessibles ?

Q 2.4 Quelle est la distribution de probabilité de transition quand on est proche du bord du carré $[-1, 1]^2$?

Q 2.5 La chaîne de Markov ainsi obtenue est-elle irréductible ? Déduisez-en que la chaîne admet une distribution de probabilité stationnaire.

Une succession de telles transitions amène à un tirage aléatoire d'une case dans le carré $[-1,1]^2$. Intuitivement, si on se déplace aléatoirement pendant suffisamment longtemps, on a une probabilité égale de se retrouver dans n'importe quelle case quelle que soit la case de départ.

Q 2.6 Montrez que la distribution uniforme sur les états est la distribution stationnaire de cette chaîne de Markov.

Replaçons-nous dans le cas général où l'ensemble d'états est l'ensemble des points du carré $[-1,1]^2$. Une des difficultés des méthodes MCMC est de savoir quand on est proche de la distribution stationnaire (c'est-à-dire combien de transitions faut-il faire ?). Dans ce problème simple, il est possible de donner une borne inférieure au nombre de transitions nécessaires avant convergence vers la distribution uniforme.

Q 2.7 Quel est la distance moyenne d'un déplacement (en supposant qu'il est accepté) ?

Q 2.8 À partir d'un point (x_0, y_0), donnez une borne inférieure en espérance sur le nombre de transitions avant convergence vers la distribution stationnaire. *Indication :* considérez la distance minimale à parcourir pour garantir que tous les points soient atteignables à partir d'un état courant (x_0, y_0).

Exercice 3 – Décodage par la méthode de Métropolis-Hastings

Dans cet exercice, nous allons appliquer une méthode de type MCMC, l'algorithme de Métropolis-Hastings, au problème de décodage d'un texte codé par substitution. Nous supposons la langue du texte connue. Nous supposons également que nous avons à notre disposition une modélisation de cette langue sous forme de bigramme : plus formellement, cette langue s'écrit avec l'alphabet fini Λ. Par exemple, en français, Λ contient les lettres minuscules et majuscules, les lettres accentuées, les signes de ponctuation, les chiffres, etc... Le modèle bigramme est donné par μ et M où μ est une distribution de probabilité sur Λ et M est une matrice stochastique qui donne pour chaque lettre de Λ la probabilité de la lettre suivante. Ce modèle peut facilement être estimé à partir d'un grand corpus de texte. Une fonction d'encodage (ou de décodage) par substitution est une fonction bijective τ de Λ dans Λ. Si T' est un texte, le texte encodée $T = \tau(T')$ est obtenu en remplaçant chaque lettre c de T' par $\tau(c)$.
Le problème que nous souhaitons résoudre ici est, étant donné un texte encodé $T = (c_1, c_2, \ldots, c_{|T|})$ (où $c_i \in \Lambda$, $\forall i$), de retrouver le texte initial décodé.

Q 3.1 Comment peut-on mesurer la vraisemblance d'une fonction d'encodage en utilisant le modèle bigramme ?

Une méthode de type Monte Carlo pour résoudre ce problème serait de tirer au hasard une fonction d'encodage avec une probabilité proportionnelle à sa vraisemblance (c'est-à-dire $\mathcal{P}(\tau) = L(\tau(T), \mu, M)/(\sum_{\tau'} L(\tau'(T), \mu, M))$) et de répéter un grand nombre de fois cette opération en gardant le texte décodé le plus vraisemblable.

Q 3.2 Combien y-a-t-il de fonctions d'encodage ? Est-ce qu'une méthode de Monte Carlo est réalisable ici ?

Comme il n'est pas possible d'échantillonner directement une fonction d'encodage selon la loi $\mathcal{P}$, on souhaite recourir à un échantillonnage par chaîne de Markov. Cette méthode ne nécessite de connaître les probabilités de tirage qu'à un facteur de normalisation près, ce qui est le cas ici.

Q 3.3 Définir une chaîne de Markov (sans donner les distributions de probabilité) qui nous permettrait de réaliser cette échantillonnage.

Étant donné la matrice stochastique A de transition de la chaîne de Markov, la méthode MCMC de Métropolis-Hastings se définit comme suit :
Répéter N fois les étapes suivantes à partir d'un état initial τ_0 choisi de manière quelconque :

- Calculer τ à partir de τ_t en échangeant deux lettres c_1, c_2 dans l'encodage, c'est-à-dire $\tau(c) = \tau_t(c)$ pour $c \neq c_1$ et $c \neq c_2$, $\tau(c_1) = \tau_t(c_2)$ et $\tau(c_2) = \tau_t(c_1)$.
- Accepter la transition de τ_t vers τ avec la probabilité $\alpha(\tau_t, \tau) = \min(1, \frac{L(\tau(T),\mu,M)A(\tau_t,\tau)}{L(\tau_t(T),\mu,M)A(\tau,\tau_t)})$ et $\tau_{t+1} = \tau$, sinon, $\tau_{t+1} = \tau_t$.

où $A(\tau, \tau')$ est la probabilité de la transition vers τ' depuis τ.
Après avoir itéré un nombre suffisamment grand de fois, la fonction de décodage τ_N correspond à un tirage aléatoire selon $\mathcal{P}$. Pour obtenir d'autres tirages selon $\mathcal{P}$, on répète ces opérations en gardant les τ_{N+kh} pour un entier $k > 0$ fixé et $h \in \mathbb{N}^*$. Le paramètre k permet d'espacer les tirages pour éviter les auto-corrélations dans l'échantillonnage.

Q 3.4 Montrer que le log de la probabilité d'acceptation s'écrit finalement :

$$\log \alpha(\tau_t, \tau) = \min(0, \mu(\tau(c_1)) + \sum_{i=1}^{|T|} \log M(\tau(c_{i-1}), \tau(c_i)) - \mu(\tau_t(c_1)) - \sum_{i=1}^{|T|} \log M(\tau_t(c_{i-1}), \tau_t(c_i)))$$

Q 3.5 Montrer que la distribution de probabilité $\mathcal{P}$ est bien la distribution stationnaire de la chaîne de Markov.