Simulation methods: Markov Chains Monte Carlo

Abass SAGNA,
abass.sagna@ensiie.fr

Maître de Conférences à l'ENSIIE,
Laboratoire de Mathématiques et Modélisation d'Evry
Université d'Evry Val-d'Essonne, UMR CNRS 8071

http://www.math-evry.cnrs.fr/members/asagna/

March 5, 2019

Plan

Plan

Plan

$\rightsquigarrow$ Let $(X_n)_{n\geq 0}$ be a Markov chain on a countable state space E with initial distribution μ: $\mu(i) = \mathbb{P}(X_0 = i)$, $i \in E$, and transition matrix $P = (P(i,j))_{i,j\in E}$:

$$\mathbb{P}(X_{i_0}, \ldots, X_{i_n}) = \mu(i_0) \prod_{k=1}^{n} P(i_{k-1}, i_k)$$

$\rightsquigarrow$ Each row $P(i,\bullet) = \{P(i,j),\ j \in E\}$ of the transition matrix is a probability on E.

$\rightsquigarrow$ We can simulate a *sample path of size n* of the Markov chain using a sequential inversion algorithm described below:

1. Generate $X_0 \sim \mu$.
2. For k from 1 to n, generate $X_k \sim P(X_{k-1}, \bullet)$.
3. Return $(X_0, \ldots, X_n)$.

Example (Simple random walk on $\mathbb{Z}$). Consider a mobile which moves up on $\mathbb{Z}$ with probability $p \in (0,1)$ and moves down with proba. $q = 1 - p$:

$$P(i,j) = \begin{cases} p & \text{if } j = i+1 \\ q & \text{if } j = i-1 \\ 0 & \text{otherwise.} \end{cases}$$

Generate a sample path of size n of the associated Markov chain $(X_k)_{k \geq 0}$ with initial distribution $\mu = \delta_0$.

Plan

Definition. Let $(X_n)_{n\geq 0}$ be a Markov chain on E with initial distribution μ and transition matrix P. $(X_n)_{n\geq 0}$ is a revertible Markov chain if for any $n \geq 1$, $(X_0, \ldots, X_n) \stackrel{d}{=} (X_n, \ldots, X_0)$: mean, for any i_0, ..., $i_n \in E$,

$$\mathbb{P}(X_0 = i_0, \ldots, X_n = i_n) = \mathbb{P}(X_0 = i_n, \ldots, X_n = i_0).$$

$\rightsquigarrow$ If $n = 1$, $i, j \in E$, we have in particular

$$\mathbb{P}(X_0 = i, X_1 = j) = \mathbb{P}(X_0 = j, X_1 = i) \iff \mu(i)P(i,j) = \mu(j)P(j,i). \quad (1)$$

Equation (1) is called *detailed balance equation*.

$\rightsquigarrow$ Summing over j gives

$$(\mu P)(i) = \sum_{j\in E} \mu(j)P(j,i) = \mu(i) \sum_{j\in E} P(i,j) = \mu(i),$$

meaning that μ is stationary.

$\rightsquigarrow$ If the detailed balance equation holds then the Markov chain is reversible. In fact

$$\begin{aligned}
\mathbb{P}(X_0 = i_0, \ldots, X_n = i_n) &= \mu(i_0)P(i_0, i_1)\ldots P(i_{n-1}, i_n) \\
&= P(i_1, i_0)\mu(1)\ldots P(i_{n-1}, i_n) \\
&= \ldots\ldots \\
&= P(i_1, i_0)\ldots P(i_n, i_{n-1})\mu(i_n) \\
&= \mathbb{P}(X_0 = i_n, \ldots, X_n = i_0).
\end{aligned}$$

Proposition. A Markov on E with stationary distribution π is reversible if and only if for any $i, j \in E$

$$\pi(i)P(i,j) = \pi(j)P(j,i) \tag{2}$$

Example (Random walks on weighted graph). Consider a complete graph for which every undirected edge between vertices i and j has a weight $\varpi_{ij} = \varpi_{ji}$. Let

$$\varpi_i = \sum_k \varpi_{ik} \qquad \text{and} \qquad \varpi = \sum_{ik} \varpi_{ik}.$$

$\rightsquigarrow$ A walker moves from i to j with a probability $P(i,j) = p_{ij}$ proportional to w_{ij}, so that $p_{ij} = \varpi_{ij}/\varpi_i$. The probability measure π defined by

$$\pi_i = \frac{\varpi_i}{\varpi}$$

is a reversible measure. In fact

$$\pi_i p_{ij} = \frac{\varpi_i}{\varpi} \times \frac{\varpi_{ij}}{\varpi_i} = \frac{\varpi_{ij}}{\varpi} = \frac{\varpi_{ji}}{\varpi} = \pi_j p_{ji}.$$

Plan

$\rightsquigarrow$ To estimate $\mathbb{E}(g(X))$ by MC algorithm, we use an iid sequence $X_1, \ldots, X_n$ of r.v. with the same distribution π as X: with probability 1 or a.s.,

$$\frac{1}{n}\sum_{k=0}^{n} g(X_k) \longrightarrow \mathbb{E}(g(X)) = \int g(x)\pi(dx).$$

$\rightsquigarrow$ Let $(X_n)_{n\geq 0}$ be an irreducible Markov chain with stationary distribution π. If $X_0 \sim \pi$, then,

$$\mathbb{P}(X_1 = j) = \sum_{i\in E} P(i,j)\pi(i) = (\pi P)(j) = \pi(j) \quad \Longrightarrow X_1 \sim \pi,$$

and we show by induction that $X_0, \ldots, X_{n-1}$ have the same distribution π but are (in general) not independent.

$\rightsquigarrow$ The ergodic theorem makes a similar estimate of $\mathbb{E}(g(X))$, $X \sim \pi$:

$$\frac{1}{n}\sum_{k=0}^{n-1} g(X_k) \longrightarrow \mathbb{E}(g(X)) = \sum_{i\in E} g(i)\pi(i) = \int g(x)\pi(dx).$$

$\rightsquigarrow$ Suppose $(X_n)_{n\geq 0}$ is an irreducible Markov chain with initial distribution μ stationary distribution π and we want to compute $\mathbb{E}g(X)$, $X \sim \pi$.

$\rightsquigarrow$ In practice it is often difficult to simulate from π (which is some time not explicit).

$\rightsquigarrow$ The ergodic theorem holds for an initial distribution μ, so that the bias introduced by starting the chain from the distribution μ instead of π disappears asymptotically.

$\rightsquigarrow$ The associated Monte Carlo algorithm is:

1. Generate $X_0 \sim \mu$.
2. For k from 1 to n, generate $X_k \sim P(X_{k-1}, \bullet)$.
3. Return $(g(X_0) + \ldots + g(X_{n-1}))/n$.

MC algorithm for Markov chain: example

Plan

Let $E \subset \mathbb{R}^d$ be a countable set and let π be a probability on E.
$\rightsquigarrow$ Our aim is to approximate

$$\int g(x)\pi(dx) \tag{3}$$

using the Hastings-Metropolis (H-M) algorithm.

$\rightsquigarrow$ The H-M algorithm is useful when it is very difficult to generate an independent sequence $X_1, \ldots, X_N$ of random variables with distribution π in order to use the Monte Carlo method.

$\rightsquigarrow$ The aim of the Hastings-Metropolis algorithm is to build a reversible Markov chain $(X_n)_{n\geq 0}$ with reversible distribution π and to approximate (3) (using the ergodic theorem) by $(g(X_0) + \ldots + g(X_{n-1}))/n$.

$\rightsquigarrow$ Let E be a countable set and let π and Q be resp. a probability and a transition probability on E. We suppose $Q(x, y) = 0 \iff Q(y, x) = 0$.

How to build the Markov chain $(X_n)_{n\geq 0}$ with reversible distribution π?

$\rightsquigarrow$ Choose $x_0 \in E$ such that $\pi(x_0) > 0$ and set $X_0 = x_0$.

$\rightsquigarrow$ $(X_n)_{n\geq 1}$ is built recursively using the rejection method. Suppose in fact that $X_k = x_k$, for $k = 0, \ldots, n$, is built and say how to define X_{n+1}.

- Generate two independent r.v. $Y_n \sim Q(x_n, \bullet)$ and $U_n \sim \mathcal{U}(]0,1[)$, both independent from $(X_k)_{k\leq n}$ and set

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

  with the convention $h(x,y) = 1$ if $\pi(x)Q(x,y) = 0$.
- If $U_n \leq h(X_n, Y_n)$, set $X_{n+1} = Y_n$.
- If $U_n > h(X_n, Y_n)$, set $X_{n+1} = X_n$.

Proposition. The process $(X_n)_{n\geq 0}$ is a reversible Markov process with reversible distribution π and transition matrix P defined as:

$$P(x,y) = \begin{cases} Q(x,y)h(x,y) & \text{if } x \neq y \\ 1 - \sum_{y \neq x} P(x,y) & \text{if } x = y. \end{cases}$$

Proof. By construction, $(X_n)_{n\geq 0}$ is a Markov chain. Let $x \neq y$. We have

$$\begin{aligned} \mathbb{P}(X_{n+1} = y | X_n = x) &= \mathbb{P}(X_{n+1} = y, U_n \leq h(X_n, Y_n) | X_n = x) \\ &\quad + \mathbb{P}(X_{n+1} = y, U_n > h(X_n, Y_n) | X_n = x) \\ &= \mathbb{P}(X_{n+1} = y, U_n \leq h(X_n, Y_n) | X_n = x) \\ &= \mathbb{P}(Y_n = y, U_n \leq h(x,y) | X_n = x) \\ &= \mathbb{P}(Y_n = y, U_n \leq h(x,y)) \\ &= \mathbb{P}(Y_n = y)\mathbb{P}(U_n \leq h(x,y)) \\ &= Q(x,y)h(x,y). \end{aligned}$$

If $x = y$, we have

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

On the other hand, if $x \neq y$,

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

which shows that π is a reversible distribution and end the proof.

$\rightsquigarrow$ A particular, but interesting example of distribution π is the Gibbs distribution.

Plan

Definition. Let E be a finite set, $\beta > 0$, and $V : E \mapsto \mathbb{R}$ be a function. The Gibbs distribution associated to V and β is defined by

$$\pi_\beta(x) = \frac{\exp(-\beta V(x))}{Z(\beta)}, \qquad x \in E,$$

where $Z(\beta) = \sum_{x \in E} \exp(-\beta V(x))$ is the normalizing constant.

$\rightsquigarrow$ Gibbs distribution is solution of the optimization problem $\max_{\pi \in \mathcal{A}} H(\pi)$ where the entropy H and the set of constrains $\mathcal{A}$ are defined as

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

and $$\mathcal{A} = \left\{ \pi \geq 0 : \sum_{x \in E} \pi(x) = 1, \quad \sum_{x \in E} V(x)\pi(x) = c \right\}.$$

Remark. If Q is symmetric,

$$h(x, y) = \min(1, \pi(y)/\pi(x)) = \exp\left(-\beta (V(y) - V(x))^+ \right).$$

and this will simplify the algorithm.

The Ising model. Let us consider the integer lattice $\Lambda = \{0, \ldots, N-1\}^2$ equipped with the horizontal and vertical neighbor relation $x \sim y$. Any spin $x = (i,j) \in \Lambda$ has one the two types $s(i,j) \in \{-1,1\}$, so that the state space of the system is $E = \{s : \Lambda \mapsto \{-1,1\}\} = \{-1,1\}^{\Lambda}$.

$\rightsquigarrow$ Let $S = (s(x))_{x\in\Lambda}$ be a possible spins configuration. We consider the Gibbs distribution on E which associates to any configuration S the proba.

$$\pi_\beta(S) = \frac{1}{Z(\beta)} \exp\Big(-\beta H(S)\Big), \quad H(S) = -\sum_{x,y\in E: x\sim y} s(x)s(y).$$

$\rightsquigarrow$ In this model, the spins interact with their neighbors and the energy associated to configuration S is $H(S)$. The term $T = 1/\beta$ is the temperature.

$\rightsquigarrow$ Remark that if $N = 40$, then $\text{card}(E) = 2^{40\times 40} \approx 10^{481}$, so that it is not possible to enumerate all the possible configurations S in order to compute $\pi_\beta(S)$. We use the H-M algorithm to simulate the Ising model.

$\rightsquigarrow$ Denote S_x the configuration obtained by changing the sign of the spin in $x \in \Lambda$ of the configuration S.

$\rightsquigarrow$ The matrix Q describes the evolution on the set of configurations:

$$\forall x \in \Lambda, \quad Q(S, S_x) = \frac{1}{\text{card}(\Lambda)}.$$

$\rightsquigarrow$ Then, at each step of the procedure, we choose a lattice $x \in \Lambda$ from a uniform distribution on Λ and change the sign of its spin.

$\rightsquigarrow$ The corresponding variation of energy is

$$\begin{aligned} \Delta H(S, S_x) := H(S_x) - H(S) &= -\sum_{y \sim x} s(-x)s(y) - \Big(-\sum_{y \sim x} s(x)s(y) \Big) \\ &= 2s(x) \sum_{y \sim x} s(y); \end{aligned}$$

$\rightsquigarrow$ Then, since $Q(S, S_x) = Q(S_x, S)$,

$$h(S, S_x) = \frac{\pi_\beta(S_x)Q(S_x, S)}{\pi_\beta(S)Q(S, S_x)} = \exp\big(-\beta\,\Delta H(S, S_x)\big).$$

$\rightsquigarrow$ *Description of the Ising algorithm*. The Ising algorithm is described from the following steps:

1. Choose an initial configuration S.
2. Repeat M times (with M large enough):
   - Simulate independently $V = x \sim \mathcal{U}(\Lambda)$ and $U \sim \mathcal{U}(]0, 1[)$,
   - If $U \leq \exp\big(-\beta\,\Delta H(S, S_x)\big)$, then replace S by S_x, otherwise, let S unchanged.
3. Return S.

Plan

The Simulated is used to find the global minimum of a function $V : E \mapsto \mathbb{R}$, where E is a finite (it may be extended to optimization with continuous control parameters) set but very large to allow a systematic search on all the domain E.

$\rightsquigarrow$ One of the most popular problems effectively solved by the Simulated annealing algorithm is the famous *travelling salesman problem*.

$\rightsquigarrow$ It consists of finding the shortest cyclical itinerary for a travelling salesman who must visit N given cities in turn.

$\rightsquigarrow$ To describe the Simulated annealing algorithm, we need to define a inhomogeneous Markov chain.

Definition. A Markov chain $(X_n)_{n\geq 0}$ with state space E is inhomogeneous if there is a sequence $(P_n)_{n\geq 1}$ of transition probabilities on E such that for any $n \geq 0$ and for any $x_k \in E$, $k = 0, \ldots, n$,

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

$\rightsquigarrow$ If $(X_n)_{n\geq 0}$ is an inhomogeneous Markov chain, then,

$$\mathbb{P}(X_{n+1} = y | X_0 = x) = (P_1 \cdots P_n)(x, y).$$

$\rightsquigarrow$ If $X_0 \sim \nu_0$, the the distribution ν_n of X_n is

$$\nu_n = \nu_0 P_1 \cdots P_n.$$

In fact, it follows from *Bayes* formula that

$$\nu_n(y) = \sum_{x\in E} \mathbb{P}(X_n = y | X_{n-1} = x)\nu_{n-1}(x) = \sum_{x\in E} P_n(x, y)\nu_{n-1}(x) = (\nu_{n-1}P_n)(y).$$

$\rightsquigarrow$ The Simulated annealing is defined from the following algorithm:

1. Fix a symmetric transition matrix Q satisfying the *Doeblin* condition: there exists $n \in \mathbb{N}$, $\alpha \in]0,1[$ and a probability π on E s.t. $\forall x, y \in E$,

$$Q^n(x,y) \geq \alpha \pi(y).$$

2. Choose a nondecreasing sequence $(\beta_n)_{n \geq 1}$ converging toward $+\infty$.
3. Build the transition matrix P_n associated to Q and β_n using the Hastings-Metropolis algorithm:

$$P_n(x,y) = \begin{cases} Q(x,y) \exp\left(-\beta_n (V(y) - V(x))^+ \right) & \text{if } x \neq y \\ 1 - \sum_{y \neq x} P_n(x,y) & \text{if } x = y. \end{cases}$$

$\rightsquigarrow$ The probability π_{β_n} defined by

$$\pi_{\beta_n}(x) = \frac{1}{Z(\beta_n)} \exp(-\beta_n V(x))$$

is the stationary distribution associated with the homogeneous Markov chain with transition matrix P_n.

$\rightsquigarrow$ If β_n goes to $+\infty$, π_{β_n} will be concentrated on points close to those realizing the minimum of V.

Example. The travelling salesman problem. We consider that:

- The travelling salesman must visit each of N cities in turn: by ending from the first visited city.
- Cities are numbered from 1 to N and each city i has coordinates (x_i, y_i).
- The set of configurations is the set S_N of possible permutations σ in $\{1, \ldots, N\}$, with cardinality $N!$

- A permutation gives an order in which the cities are visited.
- Our aim is to find the shortest cyclical itinerary for the travelling salesman, i.e., finding $\varpi = (\varpi(1), \ldots, \varpi(N)) \in S_N$ which minimizes the function

$$\begin{aligned} \sigma \in S_N \mapsto H(\sigma) &= \sum_{i=1}^{N} \text{dist}\Big(\big(x_{\sigma(i)}, y_{\sigma(i)}\big), \big(x_{\sigma(i+1)}, y_{\sigma(i+1)}\big) \Big) \\ &= \sum_{i=1}^{N} \sqrt{\big(x_{\sigma(i)} - x_{\sigma(i+1)}\big)^2 + \big(y_{\sigma(i)} - y_{\sigma(i+1)}\big)^2}, \end{aligned}$$

  with the convention that $\big(x_{\sigma(N+1)}, y_{\sigma(N+1)}\big) = \big(x_{\sigma(1)}, y_{\sigma(1)}\big)$.
- We use the following neighborhood relation in S_N: $\varpi \sim \sigma$ if their is $i < k$ such that

$$\varpi = (\sigma(1), \ldots, \sigma(i-1), \sigma(k), \sigma(i+1), \ldots, \sigma(k-1), \sigma(i), \sigma(k+1), \ldots, \sigma(N)).$$

- Then, from a permutation σ, we generate a neighbor by choosing a couple (i, k) uniformly on the set $\{1, \ldots, N\} \times \{1, \ldots, N\}$ of cardinal N^2: if $i = k$ we let σ unchanged, otherwise, we interchange the positions of cities i and k.
- The neighborhood reference transition matrix is then

$$Q(\sigma, \varpi) = \begin{cases} 1/N & \text{if } \sigma = \varpi \\ 2/N^2 & \text{if } \sigma \sim \varpi \text{ and } \sigma \neq \varpi \\ 0 & \text{if } \sigma \not\sim \varpi. \end{cases}$$

The simulated annealing algorithme to find $\varpi = \arg\min_{\sigma \in S_N} H(\sigma)$.

$\rightsquigarrow$ We choose $\beta_n = c \ln(n+1)$, $c > 0$.

- Choose $\sigma_0 \in S_N$.
- Repeat a large number of times
  - Simulate independently $\varpi \sim Q(\sigma_n, \cdot)$ and $U \sim \mathcal{U}(]0,1[)$.
  - If $U \leq \exp(-\beta_n(H(\varpi) - H(\sigma_n))$, set $\sigma_{n+1} = \varpi$, otherwise, set $\sigma_{n+1} = \sigma_n$.
- Return σ_{n+1}.

Plan

Some references

1. Annie Millet. *Méthodes de Monte-Carlo*: https://samos.univ-paris1.fr/archives/ftp/cours/millet/montecarlo.pdf
2. Matthias Winkel. *Simulation*: http://www.stats.ox.ac.uk/ winkel/ASim11.pdf
3. Otten, R.H.J.M., and van Ginneken, L.P.P.P. 1989,The Annealing Algorithm(Boston: Kluwer).