PROBABILITIES: MODELS AND APPLICATIONS. 3. MARKOV CHAINS MONTE CARLO METHODS

O. GOUBET

Abstract : Monte Carlo methods are computational algorithms that can be used to compute numerical approximations of problems that are deterministic, for instance to compte an integral as $\int_0^1 f(x)dx$. Monte Carlo methods are easy to implement on a computer. Here we consider well-known Monte Carlo algorithms that are based on Markov chains.

1. Doeblin condition

Here S is a finite set. S can be very large. Consider X_n a Markov chain on S whose transition matrix is P.

Proposition 1.1 (Doeblin condition)**.** *We say that a Markov chain satisfies the Doeblin condition if there exists a probability measure γ, a constant $c \in (0,1)$ and an integer l such that $P^l_{i,j} \geq c\gamma_j$. Then there exists a invariant probability measure π (i.e. a probability measure such that $\pi = \pi P$).*

Proof. Let us introduce the distance $d(\mu, \tilde{\mu}) = \sum_j |\mu_j - \mu_{\tilde{j}}|$. Then

$$\begin{aligned} d(\mu P^l, \tilde{\mu} P^l) = \sum_j |\sum_i (\mu_i - \mu_{\tilde{i}}) P^l_{i,j}| = \\ \sum_j |\sum_i (\mu_i - \mu_{\tilde{i}})(P^l_{i,j} - c\gamma_j)| \leq \\ (\sum_i |\mu_i - \mu_{\tilde{i}}|) \sum_j (P^l_{i,j} - c\gamma_j) = \\ (1-c) d(\mu, \tilde{\mu}). \end{aligned}$$

Then the map $\mu \mapsto \mu P^l$ is a strict contraction in a complete metric space and ha a unique fixed point. Then this fixed point is also a fixed point for $\mu \mapsto \mu P$. □

The Doeblin condition is instrumental into the proof of the simulated annealing. We will not develop this point in these lectures.

2. Metropolis-Hastings algorithm

2.1. **Setting the problem.** Let π be a probability measure on a finite set S. The idea to approximate π is to draw a Markov chain such that $\mathbb{P}(X_n = j) \to \pi_j$.

Date: November 2020.

Assume that we know π up to a multiplicative constant. Indeed we know $f_i = \kappa\pi_i$ but we do not know neither π nor κ.

2.2. **The transition matrix and the Markov chain.** Consider Q a stochastic matrix such that $q_{i,j} > 0$ if and only if $q_{j,i} > 0$. If $q_{i,j} > 0$ we say that $i \sim j$ are neighbors. The matrix Q is only here to define a way to visit the state space. This matrix has nothing to do with the target measure π.

Let us define

$$\alpha_{i,j} = \min(1, \frac{f_j q_{j,i}}{f_i q_{i,j}}) == \min(1, \frac{\pi_j q_{j,i}}{\pi_i q_{i,j}}).$$

Let us define another stochastic matrix P as

$$p_{i,j} = q_{i,j}\alpha_{i,j} \text{ if } j \neq i,$$
$$p_{i,i} = 1 - \sum_{j\neq i} p_{i,j} = q_{i,i} + \sum_{k\neq i} q_{i,k}(1-\alpha_{i,k}).$$

Lemma 2.1. *The (unknown) probability measure π is a invariant probability for P.*

Proof. We have that if $\alpha_{j,i} = 1$ and $\alpha_{i,j} = \frac{f_j q_{j,i}}{f_i q_{i,j}}$ that

$$\pi_i p_{i,j} = \kappa f_i q_{i,j} \frac{f_j q_{j,i}}{f_i q_{i,j}} = \pi_j p_{j,i}.$$

The other case is symmetric. Let us recall

Definition 2.2. *A Markov chain with transition matrix P is* **reversible** *if there exists a probability distribution π such that $\pi_i p_{i,j} = \pi_j p_{j,i}$.*

We know that then π is an invariant measure. □

It remains to define a Markov chain whose transition matrix is P. Assume $X_n = i$. For any neighbor j of i we set $X_{n+1} = j$ with probability $\alpha_{i,j}$ and $X_{n+1} = i$ with probability $1-\alpha_{i,j}$. Let us observe that if $\alpha_{i,j} = 1$ we always move from i to j.

Let $Y_{n,i}$ be a sequence of independent random variables such that $\mathbb{P}(Y_{n,i} = j) = q_{i,j}$. Let V_n a sequence of i.i.d random variables whose law is the uniform law in $[0,1]$. Then set

$$X_{n+1} = Y_{n,X_n} \mathbb{1}_{V_{n+1} \leq \alpha_{X_n, Y_{n,X_n}}} + X_n \mathbb{1}_{V_{n+1} > \alpha_{X_n, Y_{n,X_n}}}.$$

Proposition 2.3. *X_n is a Markov chain whose transition matrix is P.*

Proof. Admitted □

Proposition 2.4. *Assume $\pi > 0$. If Q satisfies the Doeblin condition then P satisfies also the Doeblin condition.*

Proof.
Let us introduce

$$0 < a = \min\{\frac{f_j q_{j,i}}{f_i q_{i,j}} \text{ for } i \sim j\} \leq 1.$$

For $i \neq j$ we have $q_{i,j} \geq p_{i,j} \geq a q_{i,j}$. Therefore

$$p_{i,i} = 1 - \sum_j p_{i,j} \geq 1 - \sum_j q_{i,j} = q_{i,i} \geq a q_{i,i}.$$

We then can prove recursively that $P^l \geq a^l Q^l$. Therefore P satisfies the Doeblin condition with constant $ca^l > 0$. □

Theorem 2.5. *Here S is finite. If X_n is irreducible and aperiodic then P satisfies the Doeblin condition (and vice-versa).*

2.3. **The algorithm.**

(1) Initialize X_0 in S.

(2) Given $X_n = i$ in S move to X_{n+1} as follows
— draw a sample j from Q, i.e. chose j in S with probability $q_{i,j}$.
— Accept the move $i \mapsto j$ with probability $\alpha_{i,j}$, i.e draw u a uniform distributed random variable on $[0, 1]$ and accept the move if and only if $u < \alpha_{i,j}$. If not $X_{n+1} = i$.

In practise this algorithm is often used with Q being symmetric. Then computing $\alpha_{i,j}$ is simpler.

3. An example : the Ising model

The state space S is $\{-1, 1\}^{N \times N}$ the space of all square matrices whose entries are either -1 or 1. For $B \in S$ the number $b_{i,j}$ is called the *spin* at site (i, j). We say that two sites are adjacent if they are neighbours on the grid and we note this as $(i, j) \sim (i_1, j_1)$. A site has at most four neighbours. For any B in S its probability π_B is a function of its *energy* defined as

$$\mathcal{H}(B) = - \sum_{(i,j)\sim(i_1,j_1)} b_{i,j} b_{i_1,j_1}.$$

The probability of a state B (using Boltzman's theory) favors the states with low energy and is given by

$$\pi_B = \frac{1}{Z_\beta} \exp(-\beta \mathcal{H}(B)),$$

where $\beta > 0$ is a constant and Z_β is the normalizing constant that is very hard to compute.

For the Metropolis-Hastings algorithm, we need to

(1) Choose a symmetric candidate Q. For instance move from B to B' according to the uniform distribution on all states that differ from B in on site, namely $q_{B,B'} = N^{-2}$ if they differ at exactly one site, 0 otherwise.

(2) Compute α which is given by $\frac{\pi_B}{\pi_{B'}} = \exp(-\beta(\mathcal{H}(B) - \mathcal{H}(B')))$.

In fact $\mathcal{H}(B) - \mathcal{H}(B')$ contains only a finite number of nonzero terms.

4. Simulated Annealing Algorithm

4.1. **A minimization problem.** Let us assume that we are given a function $\mathcal{H} : S \to \mathbb{R}$ with S finite. We want to compute the states/points where the *energy* $\mathcal{H}$ achieves its minimum. We assume here that the minimum is zero (up to a translation).

Definition 4.1 (Gibbs measure)**.** *The Gibbs measure associated to the energy at temperature θ is the probability measure defined for j in S*

$$\mu_\theta(j) = \frac{1}{Z_T}\exp(-\frac{\mathcal{H}(j)}{\theta}),$$

where Z_θ is the normalization constant also called partition function.

When $\theta \to 0$ the Gibbs measure concentrates on the points where $\mathcal{H}$ achieves its minimum. Indeed if $\mathcal{M}$ is the set of minimizers then $\mu_\theta(x) \to \frac{1}{\#\mathcal{M}}$ if $x \in \mathcal{M}$, 0 if not.

4.2. **The algorithm.**

— Generate a Markov chain which is an irreducible, aperiodic, with invariant probability μ_θ, like in the Metropolis-Hastings algorithm.
— Let it evolves to become closer to the invariant measure.
— Decrease (slowly) the temperature.

Proposition 4.2. *Let X_n be a Markov chain which is irreducible, aperiodic with invariant measure μ_θ. Then for any initial distribution*

$$\lim_{\theta\to 0}\lim_{n\to\infty}\mathbb{P}(X_n = j) = \frac{1}{\#\mathcal{M}},$$

if $x \in \mathcal{M}$, 0 if not.

In practise we vary θ as a function of n also, i.e. where θ_n is a decreasing sequence that converges towards 0, but not too fast. We have the following theorem (see [4])

Theorem 4.3. *There exists h_0 such that for any $h > h_0$ the simulated annealing algorithm associated to the Metropolis-Halsting algorithm scheme with temperature $\theta_n = \frac{h}{\log n}$, whose transition matrix is irreducible and aperiodic converges to the uniform measure on $\mathcal{M}$.*

4.3. **Example : the traveling salesman problem.** A salesman has to visit m cities located at $V_1, ...V_m$. He plans to minimize its displacements, i.e. to find out a permutation $\sigma \in S = \Sigma_m$ that minimizes

$$\mathcal{H}(\sigma) = \sum_{j=1}^{m} \operatorname{dist}(V_{\sigma(m)}, V_{\sigma(m+1)}),$$

with $m + 1 = 1$.

We say that $\sigma \sim \tilde{\sigma}$ if there exists a transposition τ such that $\sigma = \tau\tilde{\sigma}$.

— Choose an irreducible aperiodic transition probability matrix Q on S such that $p_{\sigma,\tilde{\sigma}} > 0$ if and only if $\sigma \sim \tilde{\sigma}$.
— Initialize X_0
— Repeat the Metropolis-Hastings algorithm scheme to construct X_n changing $\theta_n = \frac{c}{\log n}$ at each step.

Références

[1] P. Brémaud *Markov chains. Gibbs fields, Monte Carlo simulation, and queues.* Springer texts in applied mathematics. Springer, 1998

[2] D. Chafai and C. Malrieu *Recueil de modèles aléatoires* Mathématiques et applications 78, Springer Berlin Heidelberg New-York 2016.

[3] J-F. Delmas and B. Jourdain, *Modèles aléatoires. Applications aux sciences de l'ingénieur et du vivant.* Mathématiques et applications 57, Springer Berlin Heidelberg New-York 2006.

[4] B. Hajek *Cooling schedules for optimal annealing,* Math. Op. Research, 13(2), 311-329, 1988.

[5] S. Méléard, *Modèles aléatoires en écologie et évolution,* Mathématiques et Applications, Springer-Verlag Berlin Heidelberg (2016)

(Olivier Goubet) Laboratoire Paul Painlevé CNRS UMR 8524, et équipe projet INRIA PARADYSE, Université de Lille, 59 655 Villeneuve d'Ascq cedex.
Email address: olivier.goubet@univ-lille.fr