Metropolis Sampling

Arsène Pérard-Gayot

May 23, 2016

Introduction

The Metropolis-Hastings Algorithm

- Introduced in 1953 by Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller.
- Initially designed for the Boltzmann distribution, and was later generalized and formalized by W.K. Hastings in 1970.
- Allows to sample from probability distributions that are only known point-wise—and this, even if it is up to a constant.
- The theory behind it is related to Markov chains, which will be introduced in this lecture.

Background

Notation and Reminders

- $\mathcal{X}$: set of states,
- $\mathcal{B}(\mathcal{X})$: σ-algebra over $\mathcal{X}$,
  - $\mathcal{X} \in \mathcal{B}(\mathcal{X})$,
  - $\mathcal{B}(\mathcal{X})$ is stable under complementation,
  - $\mathcal{B}(\mathcal{X})$ is stable under countable union.
  - **Informally:** *"σ-algebras have the properties you would expect for performing algebra on sets."*
- μ is a measure over $\mathcal{B}(\mathcal{X})$ iff:
  - $\mu(\varnothing) = 0$,
  - $\forall B \in \mathcal{B}(\mathcal{X}),\ \ \mu(B) \geq 0$,
  - For all countable collections of disjoint sets $\{E_i\}_{i=1}^{\infty}$, $\mu\left(\sum_{k=1}^{\infty} E_k\right) = \sum_{k=1}^{\infty} \mu(E_k)$.
  - **Informally:** *"Measure functions have the properties you would expect for measuring sets."*

Background

Transition Kernel

A *transition kernel* is a function K defined on $\mathcal{X} \times \mathcal{B}(X)$ s.t.

- $\forall x \in \mathcal{X}$, $K(x, \cdot)$ is a probability measure,
- $\forall A \in \mathcal{B}(\mathcal{X})$, $K(\cdot, A)$ is measurable.

Informally: *"$K(x, A)$ is the probability of ending in the set of states A from a state x."*

Background

Example

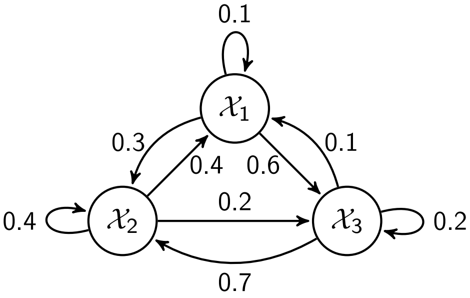
If $\mathcal{X} = \{\mathcal{X}_1, ..., \mathcal{X}_k\}$, the transition kernel is the following matrix:

$$K = \begin{pmatrix} P(X_n = \mathcal{X}_1 | X_{n-1} = \mathcal{X}_1) & \cdots & P(X_n = \mathcal{X}_k | X_{n-1} = \mathcal{X}_1) \\ \vdots & \ddots & \vdots \\ P(X_n = \mathcal{X}_1 | X_{n-1} = \mathcal{X}_k) & \cdots & P(X_n = \mathcal{X}_k | X_{n-1} = \mathcal{X}_k) \end{pmatrix}$$

Note that each row sums up to 1 since $\forall x, \sum_y P(y|x) = 1$.

Background

Example



$$K = \begin{pmatrix} 0.1 & 0.3 & 0.6 \\ 0.4 & 0.4 & 0.2 \\ 0.1 & 0.7 & 0.2 \end{pmatrix}$$

Background

Example

If $\mathcal{X}$ is continuous, we have:

$$P(X \in A|x) = \int_A K(x, y)\,\mathrm{d}y$$

Background

Homogeneous Markov Chain

An homogeneous Markov chain is a sequence (X_n) of random variables s.t.

$$\forall k,\ P(X_{k+1} \in A | x_0, x_1, ..., x_k) = P(X_{k+1} \in A | x_k) = \int_A K(x_k, \mathrm{d}x)$$

Informally: *"Each state of the chain only depends on the previous one."*

This definition implies that the construction of the chain is determined by an initial state x_0, and a transition kernel.

Background

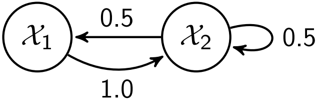
Irreducibility

The Markov chain (X_n) with transition kernel K is ϕ-irreducible iff:

$$\forall A \in \mathcal{B}(\mathcal{X}) \text{ with } \phi(A) > 0, \exists n \;\; s.t. \;\; K^n(x, A) > 0 \;\; \forall x \in \mathcal{X}$$

Informally: *"All states communicate in a finite number of steps."*

Example



$$K = \begin{pmatrix} 0.0 & 1.0 \\ 0.5 & 0.5 \end{pmatrix}$$

Background

Detailed Balance

A Markov chain with transition kernel K statisfies the *detailed balance condition* if there exists a function f s.t.

$$\forall (x, y),\ K(y, x)\, f(y) = K(x, y)\, f(x)$$

Informally: *"Going from state x to state y has the same probability as going from y to x."*

Background

Stationary Distribution

A probability measure π is a stationary distribution for the transition kernel K iff

$$\forall B \in \mathcal{B}(\mathcal{X}),\ \pi(B) = \int K(x, B)\pi(x)\,\mathrm{d}x$$

Informally: *"A transition leaves a stationary distribution unchanged."*

Under the condition of irreducibility, this distribution is unique up to a multiplicative constant.

Background

Theorem

If a Markov chain with transition kernel K statisfies the *detailed balance condition* with the *pdf* π, then π is the stationary distribution of the chain.

Proof: Using the fact that $K(y,x)\,\pi(y) = K(x,y)\,\pi(x)$.

$$
\begin{aligned}
\int_Y K(y,B)\,\pi(y)\,\mathrm{d}y &= \int_Y \int_B K(y,x)\,\pi(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_Y \int_B K(x,y)\,\pi(x)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_B \pi(x) \int_Y K(x,y)\;\mathrm{d}y\,\mathrm{d}x \\
&= \int_B \pi(x)\,\mathrm{d}x = \pi(B)
\end{aligned}
$$

Metropolis Sampling

Problem

- Sampling $X \sim f(x)$

Metropolis Sampling

Problem

- Sampling $X \sim f(x)$
- When f can be inversed analytically, use inversion.

Metropolis Sampling

Problem

- Sampling $X \sim f(x)$
- When f can be inversed analytically, use inversion.
- When f is known up to a constant, use rejection sampling.

Metropolis Sampling

Problem

- Sampling $X \sim f(x)$
- When f can be inversed analytically, use inversion.
- When f is known up to a constant, use rejection sampling.
- When f is only known point-wise and up to a constant, *what can we do*?

Metropolis Sampling

The Metropolis-Hastings algorithm

Idea: Construct an homogeneous Markov chain that converges to the target distribution $f(x)$. Here, g is a function s.t. $g \, \alpha \, f$.

Start from an initial state x_0, and $t = 0$.
loop
  Choose a proposal sample $Y_t \sim q(y|x_t)$.
  Compute $a = min(1, \frac{q(x_t|y_t)g(y_t)}{q(y_t|x_t)g(x_t)})$.
  Sample $U \sim \mathcal{U}(0, 1)$.
  if $u \leq a$ **then**
    $x_{t+1} \longleftarrow y_t$ ▷ Accept
  else
    $x_{t+1} \longleftarrow x_t$ ▷ Reject
  end if
  $t \longleftarrow t + 1$
end loop

Metropolis Sampling

Proposal distribution

- How to design the proposal distribution q?

Metropolis Sampling

Proposal distribution

- How to design the proposal distribution q?
- Freedom in the choice of q as long as it follows some properties to ensure convergence.
- The two following conditions form a sufficient convergence criterion:
  - *Non-zero rejection probability*
    $P\left[f(X_t)q(Y_t|X_t) \leq f(Y_t)q(X_t|Y_t)\right] < 1$
  - *Strong irreducibility*
    $\forall(x, y),\ q(y|x) > 0$
- When these conditions are met, the chain converges to the *stationary distribution* of the chain.

Metropolis Sampling

Convergence

We can prove that:

- The kernel associated with the Markov chain generated by the algorithm statisfies the *detailed balance* with the target function f.
- This implies that f is a stationary distribution of the chain.
- *Under the sufficient convergence conditions, the chain then converges to the distribution* f*.*

Metropolis Sampling

Key Messages

- The Metropolis Hastings algorithm generates a Markov chain which converges to the distribution f.
- There is freedom in the choice of the proposal q as long as the convergence is ensured.
- The target function f needs only be known point-wise and up to a constant.

Practical Example

Sampling a Complex Function

- Sampling from the function $f(x) = (cos(50\,x) + sin(20\,x))^2$.
- Python-powered utterly cool demo.