Cours
M2MVA

The Metropolis-Hastings algorithm:

This algorithm manage to sample from "almost any" probability distribution (up to the burn in period) in a very simple way. Moreover, it only requires to know the target π up to a multiplicative constant (usually the normalisation constant). In addition, you will see that the possibilities that the MH offers make it possible to think of the currently employed MCMC methods as special cases of this algorithm.

Let now π be our target distribution which we assume to have a pdf w.r.t the Lebesgue measure. We only know π up to a multiplicative constant. We also assume that we are able to sample from what is called the proposal density $q(x, \cdot)$ (w.r.t. the Lebesgue measure).

1) The MH algorithm:

Start with an initial value X_0.

Given X_n, a "candidate" Y_{n+1} is generated according to $q(X_n, \cdot)$. It is then accepted with probability $\alpha(X_n, Y_{n+1})$ given by:

$$\alpha(X_n, Y_{n+1}) = \begin{cases} \min\left(1 ; \dfrac{\pi(Y_{n+1})\, q(Y_{n+1}, X_n)}{\pi(X_n)\, q(X_n, Y_{n+1})}\right) & \text{if denom} > 0 \\ 1 & \text{otherwise.} \end{cases}$$

If Y_{n+1} is accepted then $X_{n+1} = Y_{n+1}$ otherwise $X_{n+1} = X_n$.

This produces a Markov chain $(X_0, \dots, X_n, \dots)$

This is a "reject-type" algorithm. However using all simulations and not only those accepted.

Pseudo-code: $\begin{cases} X_0 \text{ given} \\ n = 0 \end{cases}$ then Repeat

$U_{n+1} \sim \mathcal{U}([0,1])$
$Y_{n+1} \sim q(X_n, \cdot)$
if $U_{n+1} < \alpha(X_n, Y_n)$ ← this enables to manage the acceptation or reject.
then $X_{n+1} = Y_{n+1}$
otherwise $X_{n+1} = x_n$
$n = n+1$

2) Transition kernel:

- Intuitively: Let P be the transition kernel. 2 cases appear:

→ if we accept y which has been proposed : $p(x,y) = q(x,y)\,\alpha(x,y)$ with $y \neq x$. Indeed, the "Probability" to go from x to y is equaled to the probability of simulating y from x : $q(x,y)$ times its proba to be accepted $\alpha(x,y)$

→ If we have rejected, it means that we stay at x and this probability is $P(x,\{x\}) = \int q(x,y)\,(1-\alpha(x,y))\,dy$ ie we sum over all y which have been proposed and rejected.

this writes: $\forall A \in \mathcal{X}$

$$P(x,A) = \int_A q(x,y)\,\alpha(x,y)\,dy + \delta_x(A) \int_{\mathcal{X}} (1-\alpha(x,y))\,q(x,y)\,dy$$

Rq: this kernel depends on π through α which only depends on $\frac{\pi(y)}{\pi(x)}$. This explains why we only need π up to a multiplicative constant

- Formally: Now we can get this expression by a calculation:

Let f be a Borel function continuous on $\mathcal{X}$.

$$\mathbb{E}[f(X_{n+1}) \mid \mathcal{F}_n] = \mathbb{E}_{(U,Y)}\left[f(Y)\,\mathbb{1}_{\{U \leq \alpha(X_n,Y)\}} + f(X_n)\,\mathbb{1}_{\{U > \alpha(X_n,Y)\}}\right]$$

$$\overset{Y \perp\!\!\!\perp U}{=} \mathbb{E}_Y\left[f(Y)\,\alpha(X_n,Y) + f(X_n)\,(1-\alpha(X_n,Y))\right]$$

$$= \int f(y)\,\alpha(x,y)\,q(x,y)\,dy + \delta_{X_n}(f) \int (1-\alpha(x,y))\,q(x,y)\,dy$$

$$= \int f(y)\,P(x,y)\,dy \quad \text{where } P \text{ is given above.}$$

3) Invariant measure:

Prop: $\pi P = \pi$

Proof: We will use the detailed balanced relation. We need to prove that:

" $\pi(x)\,P(x,y) = \pi(y)\,P(y,x)$ " or in a more formal way:

$\forall A, B \in \mathcal{X}$, show that $\int_{A\times B} \pi(dx)\,P(x,dy) = \int_{A\times B} \pi(dy)\,P(y,dx)$

We first write $P(x,A)$ as : $P(x,A) = P_1(x,A) + P_2(x,A)$ where

$P_2(x,A) = c(x)\,\delta_x(A)$ and $P_1(x,A) = \alpha(x,y)\,q(x,y)\,\mathbb{1}_{\{y \in A\}}$

We will show the previous inequality for both P_j $(j=1,2)$ which by linearity will conclude the proof. For clarity we will work with densities w.r.t. Lebesg

- Let us start with P_1 :

By definition of the acceptation ratio α, we have

$$\pi(x)\,\alpha(x,y)\,q(x,y) = \pi(x)\left[\frac{\pi(y)}{q(x,y)}\,\frac{q(y,x)}{\pi(x)} \wedge 1\right] q(x,y)$$

$$= \left[\pi(y)\,q(y,x) \wedge \pi(x)\,q(x,y)\right]$$

$$= \pi(y)\,\alpha(y,x)\,q(y,x)$$

therefore : $\int_{A\times B} \pi(dx)\,P_1(x,dy) = \int_{A\times B} \pi(x)\,\alpha(x,y)\,q(x,y)\,dx\,dy$

$$= \int_{A\times B} \pi(y)\,\alpha(y,x)\,q(y,x)\,dx\,dy$$

$$= \int_{A\times B} \pi(dy)\,P_1(y,dx)$$

- For P_2 we will use the convention that $\int_\emptyset = 0$

$$\int_{A\times B} \pi(dx)\,P_2(x,dy) = \int_{A\times B} \pi(dx)\,c(x)\,\delta_x(y)\,dy$$

$$= \int_{A\cap B} \pi(dx)\,c(x) = \int_{A\cap B} \pi(dy)\,c(y)$$

$$= \int_{A\times B} \pi(dy)\,c(y)\,\delta_y(dx) = \int_{A\times B} \pi(dy)\,P_2(y,dx)$$

4) Examples : what choice for q ?

2) Independente sampler :

This is the most simple case since the proposal does not depend on the current state of the chain : $q(x,y) = q(y)$. Of course, the dependency will appear through the acceptation ration : $\alpha(x,y) = \left(\frac{\pi(y)}{q(y)}\,\frac{q(x)}{\pi(x)} \wedge 1\right)$

(: This looks like a generalisation of the acceptation reject method.

The convergence properties of the generated chain are obviously depending on q

Prop: $(X_n)_n$ is irreductible and aperiodic iif q is positive (>0) on the support of π

This also enable to get the stronger convergence property = the geometric ergodicity - under a weak condition. Here I give you the theorem which you will be able to prove with the tools introduced in the following classes.

Let us first define the Total variation norm:

$\| P(x,.) - \pi \|_{TV} = \sup_{A \in \mathcal{X}} \left| \int_A (P(x,y) - f(y))\, dy \right|$ where P is used for both the kernel and its corresponding probability measure.

then:

thm: the ncnc algorithm given by the Independents sampler (IS) M-H algorithm produces a uniformaly ergodic Markov chain if there exist a constant $M > 0$ s.t: $\pi(x) \leq M\, q(x) \quad \forall x \in \text{supp}\,\pi$ (M)

In this case: $\| P^n(x,.) - \pi(.) \|_{TV} \leq \left(1 - \frac{1}{M}\right)^n$

Moreover, if $\forall M > 0 \quad \exists A \in \mathcal{X}$ s.t $\pi(A) \neq 0$ and (M) is not satisfied on A, then (X_n) is not geometrically ergodic.

This provides a simple condition to prove the convergence property.

One can also wonder: how efficient is this sampler? Its efficiency may be measured by its mean acceptance ratio: $\mathbb{E}[\alpha(X,Y)]$

lemma: If the chain is stationary, then if q satisfies (M) then

$$\mathbb{E}[\alpha(X,Y)] \geq \frac{1}{M}$$

Proof:
$$\mathbb{E}[\alpha(X,Y)] = \int \left(1, \frac{\pi(y)}{q(y)} \frac{q(x)}{\pi(x)}\right) \pi(x)\, q(y)\, dx\, dy$$

$$= \int \left(\pi(x)\, q(y) \wedge \pi(y)\, q(x)\right) dx\, dy$$

$$\overset{q(y) \geq \frac{1}{M}\pi(y)}{\geq} \frac{1}{M} \int \left[\pi(x)\, \pi(y) \wedge \pi(y)\, \pi(x)\right] dx\, dy = \frac{1}{M}$$

Ex: Simulation of $\pi = \mathcal{N}(0,1)$ using $q = \mathcal{N}(0, \sigma^2)$

$$\alpha(x,y) = \inf\left[1, \exp\left(-\frac{1}{2}y^2 + \frac{1}{2}x^2 + \frac{1}{2\sigma^2}y^2 - \frac{1}{2\sigma^2}x^2\right)\right]$$

$$= \inf\left[1, \exp\left(-\frac{1}{2}(y^2 - x^2)(1 - \sigma^{-2})\right)\right]$$

→ if $\sigma^{-2} > 1$: we have a proposal with a heavier tail than the target distribution. We will propose many elements far from 0 : ie $|Y_{n+1}| > |X_n|$. But

$1 - \frac{1}{\sigma^2} > 0 \Rightarrow \exp\left(-\frac{1}{2}(Y_{n+1}^2 - X_n^2)(1 - \sigma^{-2})\right) < 1$ and even more if $Y_{n+1}^2 >> X_n^2$. These elements are therefore very likely to be rejected. The elements Y_{n+1} s.t. $Y_{n+1}^2 \leq X_n^2$ will be accepted with probability 1. Therefore, to compensate the heavy tail proposal, many elements will be rejected.

→ On the other hand, if $\sigma^{-2} < 1$, all the candidates s.t. $|Y_{n+1}| > |X_n|$ will be accepted with probability 1 but not all the elements in the complementary set. This compensates the fact that many elements will be proposed close to 0.

Demo: Cible : $\pi : \mathcal{N}(0,\mu)|_{[0,1]}$ avec $\mu = 0,01$

General Independence sampler → draw a distribution with small variance than large one.

b) Symmetric Random Walk M-H (SRWMH)

Assume that $\mathcal{X}$ is a vector space (typically $\mathbb{R}^d$) on which we can define a random walk with transition density $q(x,y) = q(|x-y|)$ with q even function.

In this case, the acceptation ratio reduces to $\alpha(x,y) = \inf\left(1, \frac{\pi(y)}{\pi(x)}\right)$

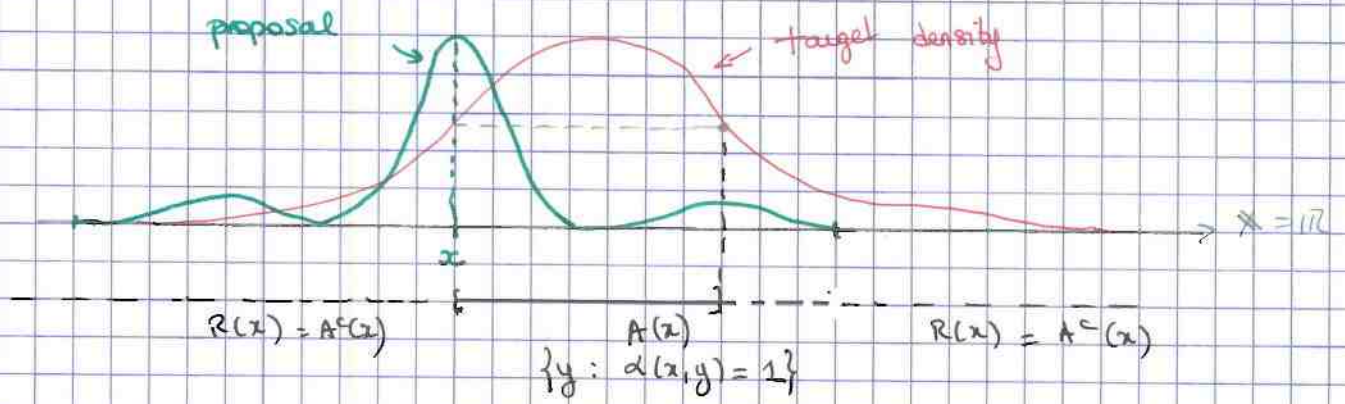
This has the following interpretation:

With the usual X_n, we accept systematically Y_{n+1} such that the "likelihood ratio" $\frac{\pi(\cdot)}{q(X_n, \cdot)}$ is greater than the opposite move ($Y_{n+1} \to X_n$)

Here, the condition becomes : π increases ⇒ accept with probability 1.

The algorithm baises the Random Walk by promoting samples which move toward the modes of π

Schematically this can be viewed here:



Although quite simple, this algorithm is very versatile as one has a large range of possible q. In particular, it is appealing as π only shows up in α. However, it has some drawbacks also; in particular:

thm: If π has non-compact support and if q is symmetric, the Hastings chain $(X_n)_n$ is not uniformly ergodic.

However with additional assumptions on π and q, one can achieve the geometric ergodicity.

Ex: $\pi(x) \propto \exp(-\gamma|x|^\beta)$, $\gamma > 0$, $\beta \geq 0$

$$q(x,y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-y)^2}{2\sigma^2}\right)$$

Given X_n, the SRW samples $Y_{n+1} \sim \mathcal{N}(X_n, \sigma^2)$ and gives

$$\alpha(X_n, Y_{n+1}) = \left[1 \wedge \exp\left(\gamma\{|X_n|^\beta - |Y_{n+1}|^\beta\}\right)\right]$$

this implies as previously 2 cases: if $|Y_{n+1}| \leq |X_n|$ then it is automatically accepted; on the other hand if $|Y_{n+1}| > |X_n|$ it may be rejected

It seems that σ^2 does not have a real impact on the results. To see how it actually does, we compute the auto-correlation of the generated chain.

$$C_i = \frac{\sum_{j=1}^{N-i} (X_j - \bar{X})(X_{j+i} - \bar{X})}{N-i} \times \frac{N}{\sum_{j=1}^{N} (X_j - \bar{X})^2} \quad \text{with } \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

Demo: π Exponential density on $(0,1)$ $\gamma = 1$, $\beta = 1$

q: Gaussian RW

$\sigma^2 = 2 \rightarrow$ good C_i

$\sigma^2 = 0.1 \rightarrow$ bad C_i.

A second question is the dependency on X_0.

slides $\pi = \exp$ $\quad q = \mathcal{N}(0, \sigma^2)$ $\quad X_0 = 1$ or $X_0 = 100$ is burn in 200 steps

The auto-correlation measure how "independent" are the first samples w.r.t. the last ones. You expect that it will decrease quickly so that your chain does not depend too long on its initial value.

this is what is illustrated here.

You can also introduce a burn-in and the autocorrelation will help you have a first guess of its value - This plot reflects the stationarity behaviour

Ex2: A heavy tailed Cauchy distribution:

$\pi(x) \propto (1 + |x|)^{-2}$.

$q \sim \mathcal{N}(0, \sigma^2)$ where σ^2 so that $\mathbb{E}[\alpha] \simeq 35\%$.

slide 5 : quite good at first sight but! the maximum value of our sampler is less than 30 even though we have done 10.000 iterations. The Cauchy distribution has a heavy tail and one can expect a maximum value around 100.

Since the cumulative distribution function is available here, one can plot a q-q plot to visually assess the goodness of fit of this output with our target (qq plot : plots the quantiles against each other → goal: to be as close as possible to the diagonal)

slide 6 : 3 independent runs of the algo together with their qq plot.

→ good agreement for small values

→ fail completely for large ones

→ ⊕ erratic behaviour 3 runs → 3 very different shapes.

demo: cas multimodal

c) MALA:

One may wonder how one can better use the target into the proposal. Many solutions were proposed although one appeared to be well adapted to high dimensional problems. It starts from a diffusion equation in continuous time:

$dX(t) = a(t, X(t))\,dt + \sigma^{-2}(t, X(t))\,dB(t)$ where B is a Brownian motion. This equation is given as elements of $L^2(d\mathbb{P})$

In terms of diffusion process π is the stationary distribution under conditions on a and σ^2.

The goal of the MALA algorithm is to go from continuous to discreet schemes. This enables (thanks to an Euler scheme) to propose the following algorithm:

for a C^1 target π, we define $q(x,y)$ by:

$$q(x,y) = \mathcal{N}\left(x + \delta \nabla_x \log \pi\,;\, \delta\, Id\right)$$

This introduces a complex dependency of q w.r.t. π

→ Expliquer sur ex gaussienne $\pi \propto e^{-\frac{x^2}{2}}$

$$\log \pi = -\frac{x^2}{2}$$

$$\nabla \log \pi = -x$$

Rappel vers les valeurs du mode

→ Journalole au multimodal → demo