Markov Chain Monte Carlo (MCMC) and Hybrid Monte Carlo (HMC)

Alexandre Défossez

January 23, 2019

Slides available at `https://ai.honu.io`

Motivation

- For a random variable X we want to compute $\mathbb{E}(f(X))$.
- We take $(X_i)_{i\in\mathbb{N}}$ i.i.d with $X_i \sim X$ and compute

$$\frac{1}{n}\sum_{i}^{N} f(X_i) \underset{N\to\infty}{\overset{(p.s)}{\to}} \mathbb{E}(f(X))$$

- Convergence and speed of convergence can be obtained from Central Limit Theorem.
- We want to generate $(x_1, \ldots, x_N)$ according to the distribution of X.
- Let P be the density of X.

Monte Carlo Markov Chains methods

- Only requires knowing P up to multiplicative constant (no need to compute normalizing constant).
- We take (X_i) a Markov chain defines as:

$$\begin{aligned}\mathbb{P}(X_{n+1} = y | X_n = x, X_{n-1}, \dots X_0) &= \mathbb{P}(X_{n+1} = y | X_n = x) \\ &= S(x, y).\end{aligned}$$

- S has the right properties, there exist a stationary law π e.g. $X_n \sim \pi \rightarrow X_{n+1} \sim \pi$.
- We want S so that P is stationary.
- If $\forall x, y, S(x, y) > 0$ then π is unique and X_n always converges to π.

Metropolis Hastings

- Given distribution T, Metropolis Hasting rule $S(x, y)$:
  1. Sample $z \sim T(x, \cdot)$.
  2. With probability $\alpha = \min\left(1, \frac{P(z)T(z,x)}{P(x)T(x,z)}\right)$ accept $y = z$ otherwise, reject and take $y = x$.
- $T(x, \cdot)$ is usually a gaussian centered in x.
- If T is symetrical, it cancels out when computing α.

Limitations de MCMC

- No result of convergence. Can take a long time to reach stationary distribution (mixing time).
- Successive (X_i) are not independent. We need to wait long enough to be able to sample a new independant X.
- Taking T gaussian means exploring the space as $\sqrt{n}$ where n is the number of iterations we run.

Hamiltonian mechanics

- $q(t), p(t) \in \mathbb{R}^d$.
- $H(q,p) = K(p) + U(q)$ with $K(p) = \sum_i p_i^2/(2m_i)$.
- We take q, p verifying

$$\dot{q} = \frac{\partial H}{\partial p}$$

$$\dot{p} = -\frac{\partial H}{\partial q}$$

  where $\dot{f} := \frac{\mathrm{d}f}{\mathrm{d}t}$.
- $H(q(t), p(t))$ is constant. The *flow* of (q,p) preserves volumes.

Ensemble canonique

- In statistical physics, the probability of a specific state under a fixed temperature T is given by:

$$P(q,p) = \frac{1}{Z}\exp(-H(q,p)/T)$$

- Evolution of q, p preserves probability.
- For our specific K,

$$P(q,p) = \frac{1}{Z}\exp(-U(q)/T)\exp(-K(p)/T)$$

so that q and p are independent. Marginal density of q is

$$P(q) \propto \exp(-U(q)/T).$$

Discretisation

- We discretize time using the Leap Frog algorithm:

$$p_{n+1/2} = p_n - \frac{\tau}{2}\frac{\partial U}{\partial q}(q_n)$$
$$q_{n+1} = q_n + \tau\frac{p_{n+1/2}}{m}$$
$$p_{n+1} = p_{n+1/2} - \frac{\tau}{2}\frac{\partial U}{\partial q}(q_{n+1})$$

- H_n will have bounded oscillation for τ small enough $=>$ stable simulation of ODE.
- This method preserves the volume (q, p) and is reversible.

Hybrid Monte Carlo

- We take T given by Hamiltonian mechanics simulated with Leap Frog for N iterations.
- We obtain a new state (q^*, p^*). For T to be symetrical, we take $p^* = -p(N)$.
- We accept (q^*, p^*) with probability

$$\min\left[1, \exp\left(-H(q^*, p^*) + H(q, p)\right)\right]$$

  otherwise we keep (p, q).
- Before starting again, we sample $p \sim \mathcal{N}(0, m)$.

HMC properties

- Resampling p allows to explore all q. We would otherwise stay in an area where U is bounded (as $U \leq H$ and H almost constant with Leap-Frog).
- HMC leaves the canonical distribution of q, p invariant.
- HMC explore all the q-space. Unless solutions to the Hamiltonian equations are τN periodic.
- Solution: add randomness to τ and N.

HMC

We want to sample from $\mathcal{N}(0, \Sigma)$. Let's take $\Sigma = \begin{pmatrix} 1 & 0.95 \\ 0.95 & 1 \end{pmatrix}$, $N = 25$, $\tau = 0.25$.

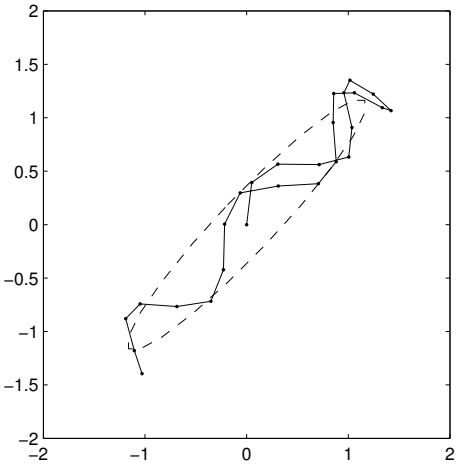


Figure: Values of q during Leap Frog

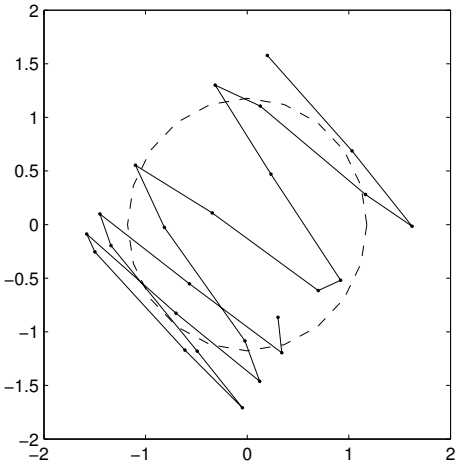


Figure: Values of p during Leap Frog

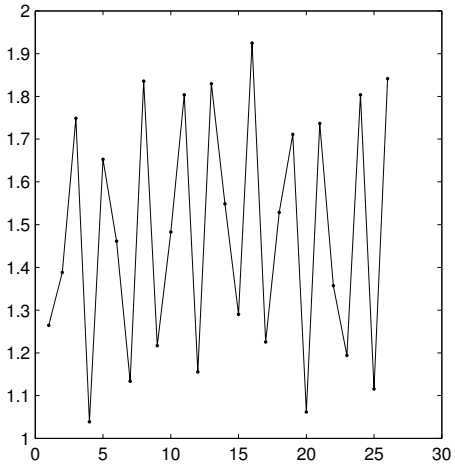


Figure: Values of H during Leap Frog

MCMC

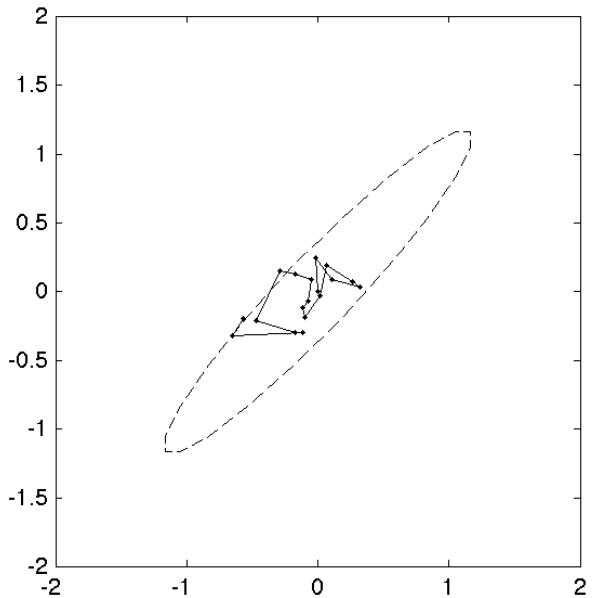


Figure: Values of q with 25 steps of MCMC.

- For MCMC we take $q_{n+1} \sim \mathcal{N}(q_n, \tau^2)$.
- Random walk: after N steps, moved by $\sim \sqrt{N}\tau$.
- For HMC: consistent direction, moved by $\sim \tau N$ while preserving the target distribution.

Comparison of HMC and MCMC

$q \sim \mathcal{N}(0, \Sigma)$ with

$$\Sigma = \begin{pmatrix} 1 & 0.98 \\ 0.98 & 1 \end{pmatrix} \quad N = 20 \quad \tau = 0.18$$

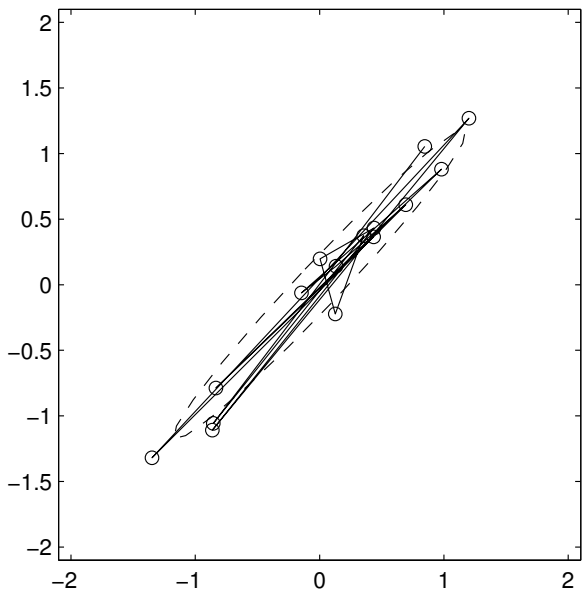


Figure: Sampled values with HMC

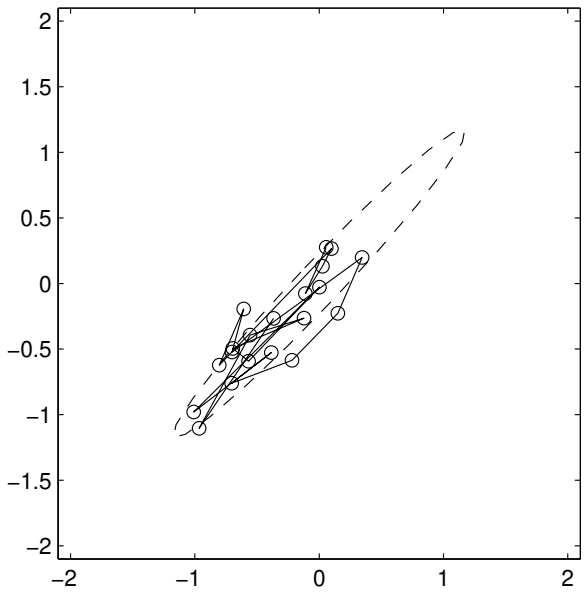


Figure: sampled values with MCMC

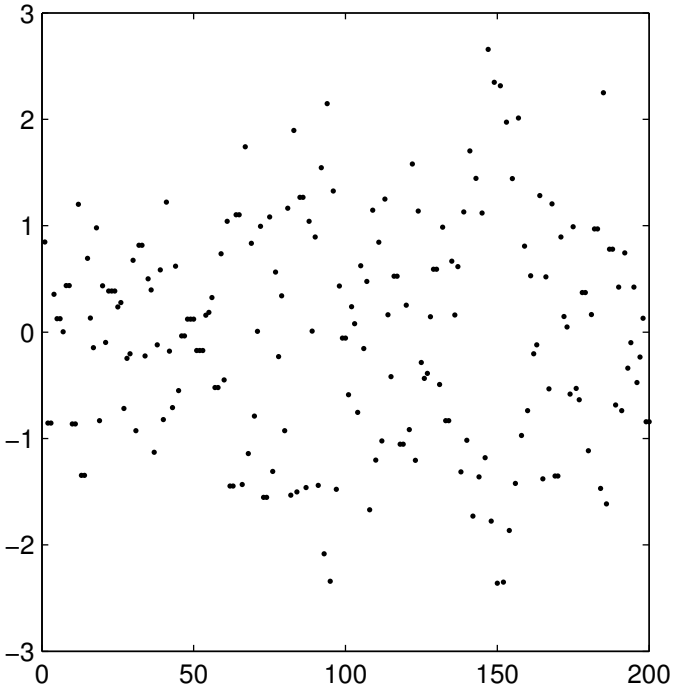


Figure: First coordinate for HMC

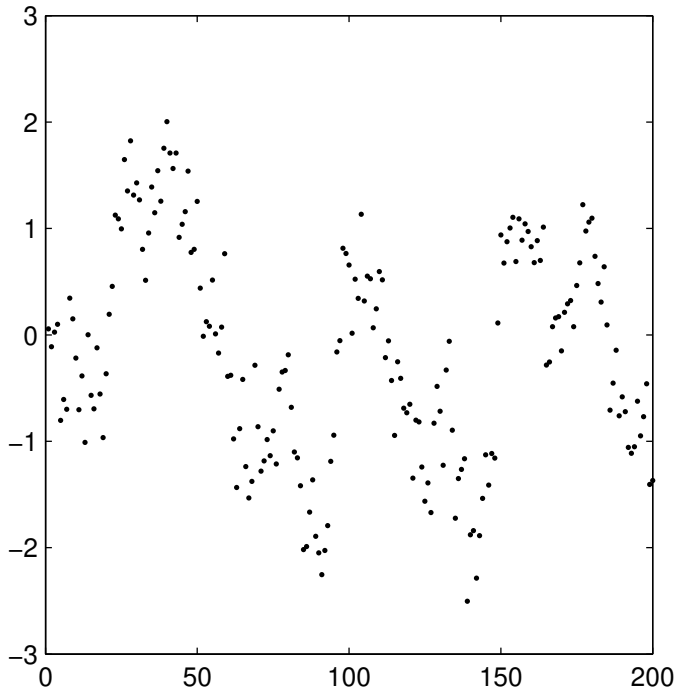


Figure: First coordinate for MCMC

Choosing N and τ

- Let σ_1^2 (resp σ_2^2) be the smallest (resp largest) eigenvalue of Σ.
- For MCMC, to have sufficiant acceptance rate, τ needs to be of the same order as σ_1. de λ_2 donc $N \sim \left(\frac{\lambda_2}{\lambda_1}\right)^2$.
- To have independent consecutive values, we need to move by at least σ_2 after N iterations so that

$$N \sim \frac{\lambda_2^2}{\lambda_1^2}.$$

- Same as the condition number in optimization !
- For HMC, we just need

$$N \sim \frac{\lambda_2}{\lambda_1}.$$

Statistical test

- Kolmogorov test on the first coordinate (statistical test comparing the empirical CDF with the true one).
- HMC succeed with a p-value similar to Gaussian value generated by MATLAB while MCMC fails.

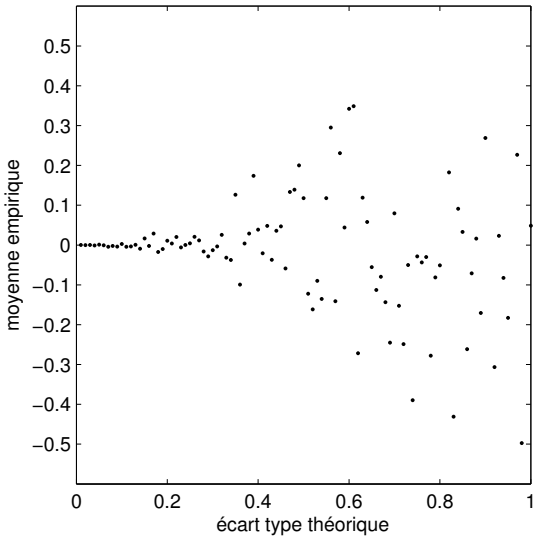
MCMC	HMC	Gaussienne
4e-4	0.4955	0.54

Figure: p-values from the Kolmogorov test applied on the first coordinate

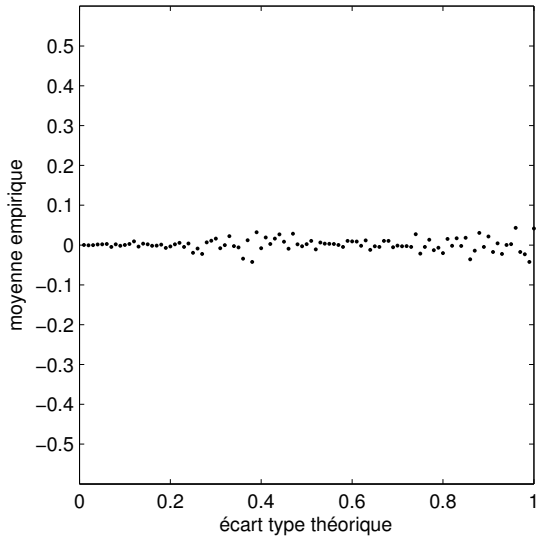
Exploration in higher dimension

- For more complex distributions, it is not always possible to choose appropriate τ and N for MCMC.
- $\Sigma = \mathrm{diag}(0.01, 0.02, \ldots, 1)^2$.
- For HMC we take $\tau = 0.013 \pm 20\%$, $N = 150$ (remember the periodicity problem!).
- For MCMC we take $\tau = 0.022 \pm 20\%$, $N = 150$.
- We run both for 1000 steps, with a 1000 steps of warmup for MCMC.

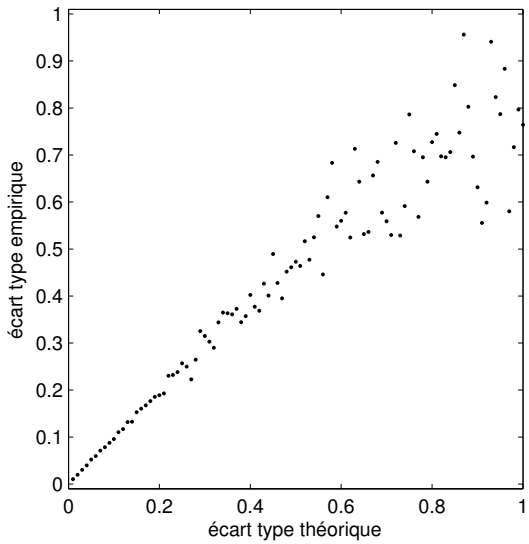
Exploration in higher dimension

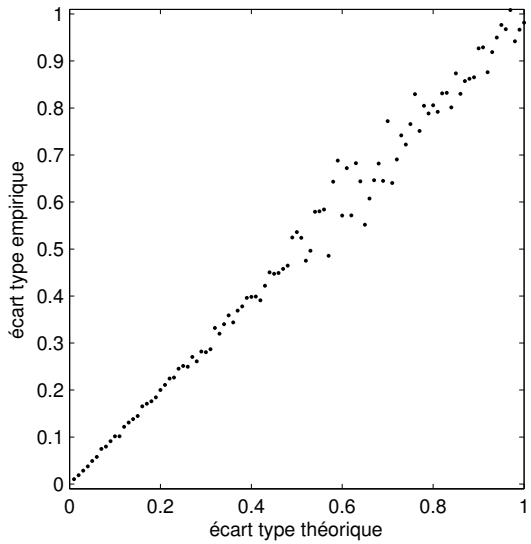


MCMC

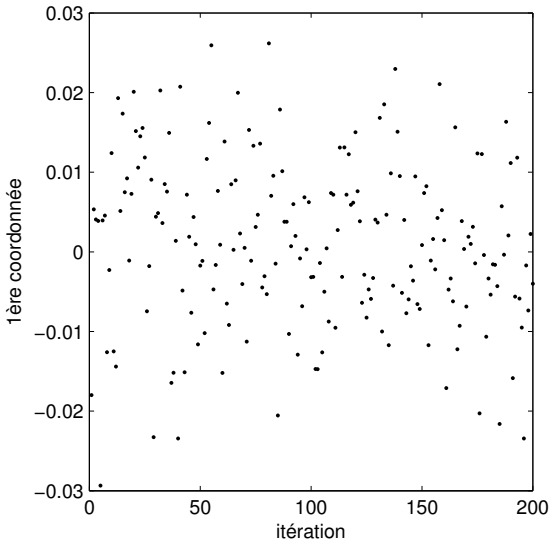


HMC

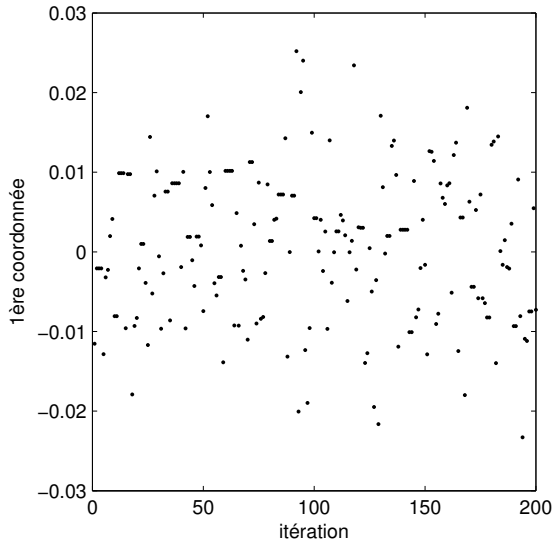




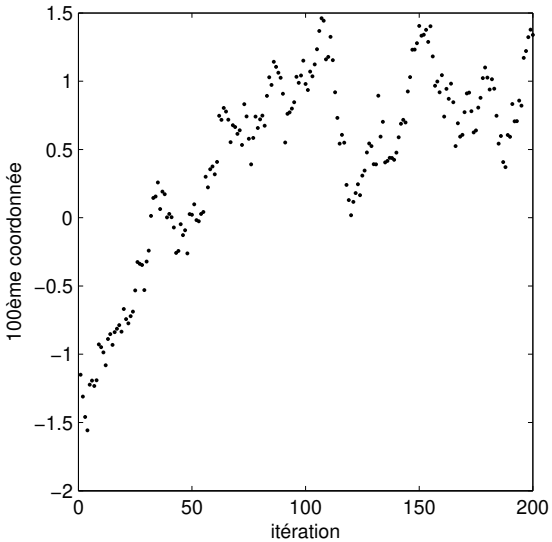
Exploration in higher dimension

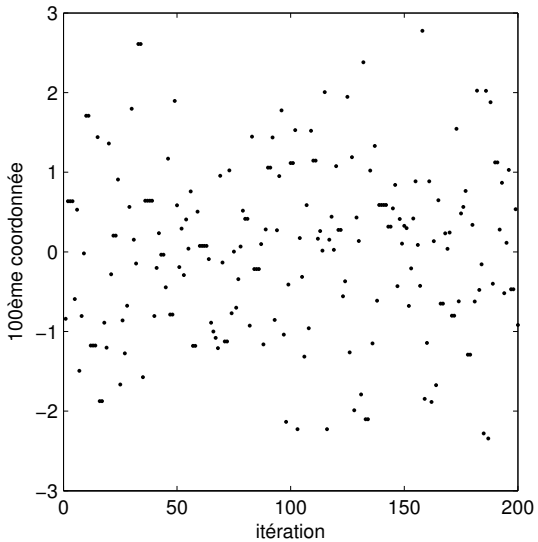


MCMC

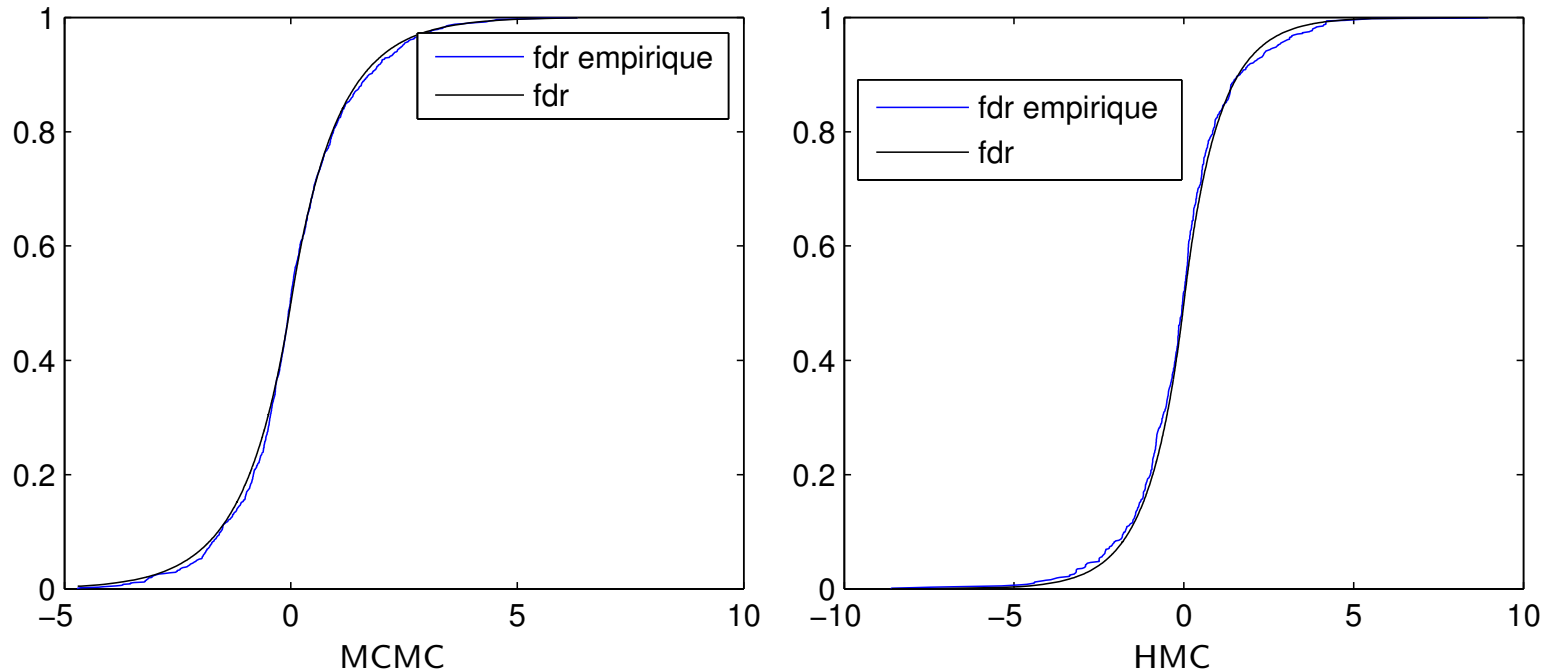


HMC

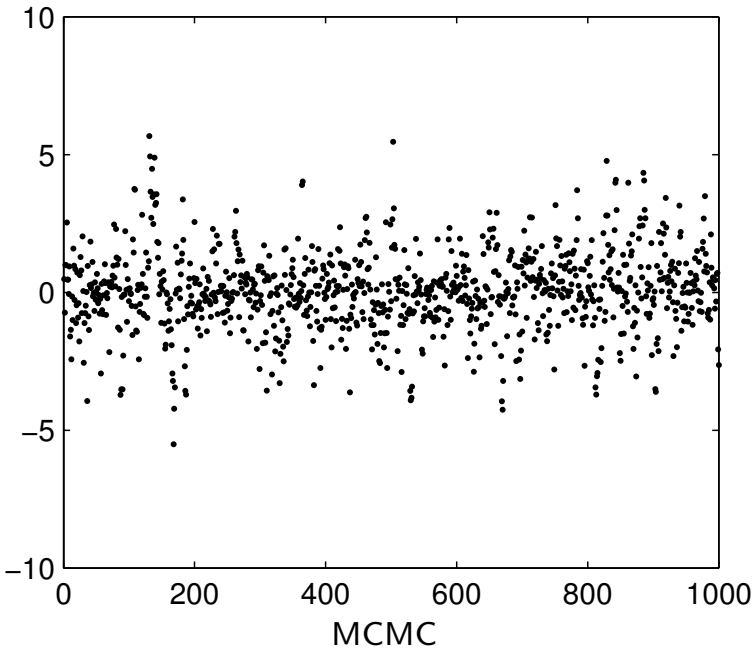


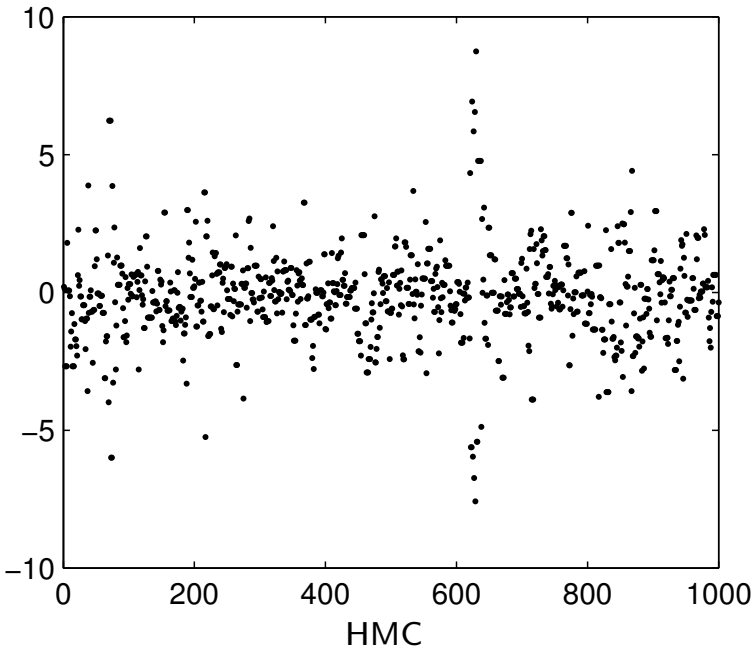


Loi $P(x) = \exp(-|x|)$



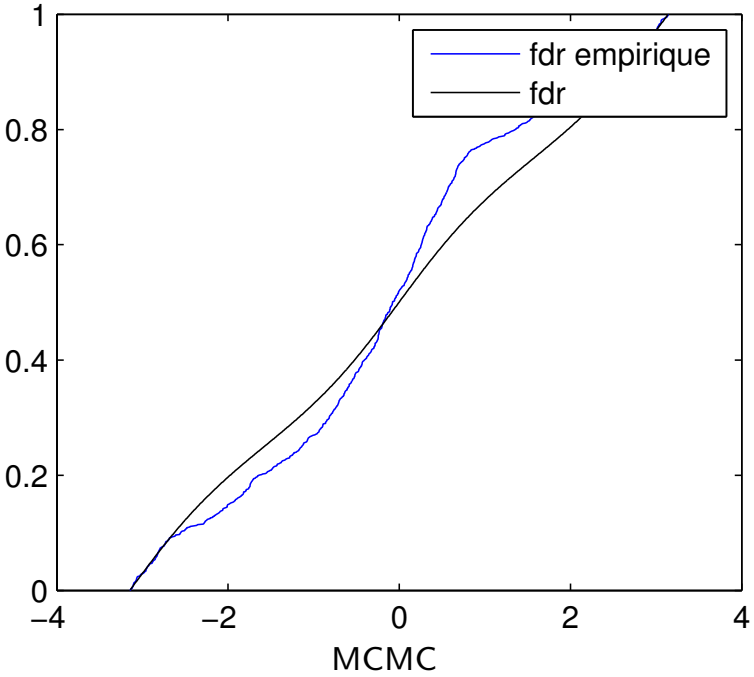
With $P(x) = \exp(-|x|)$

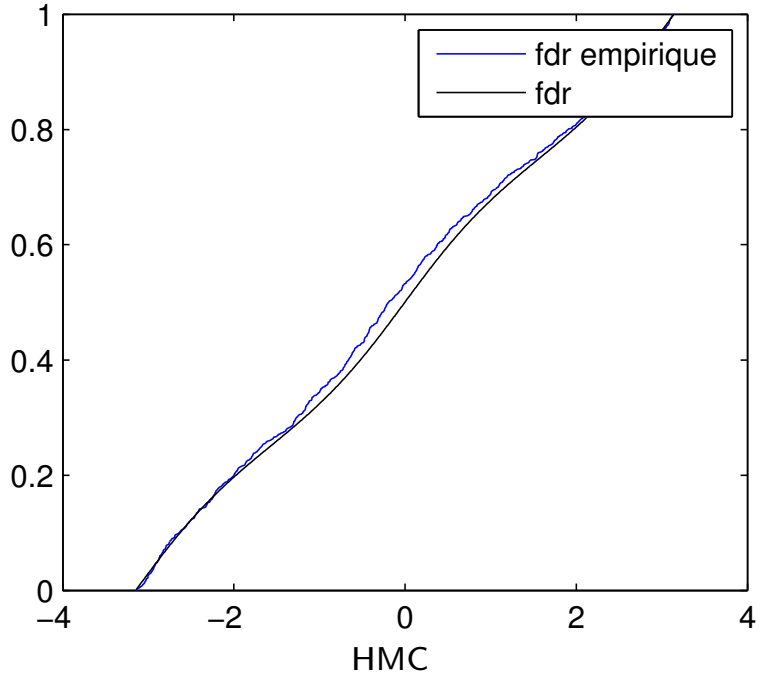

10
5
0
−5
−10
0
200
400
600
800
1000
MCMC


10
5
0
−5
−10
0
200
400
600
800
1000
HMC

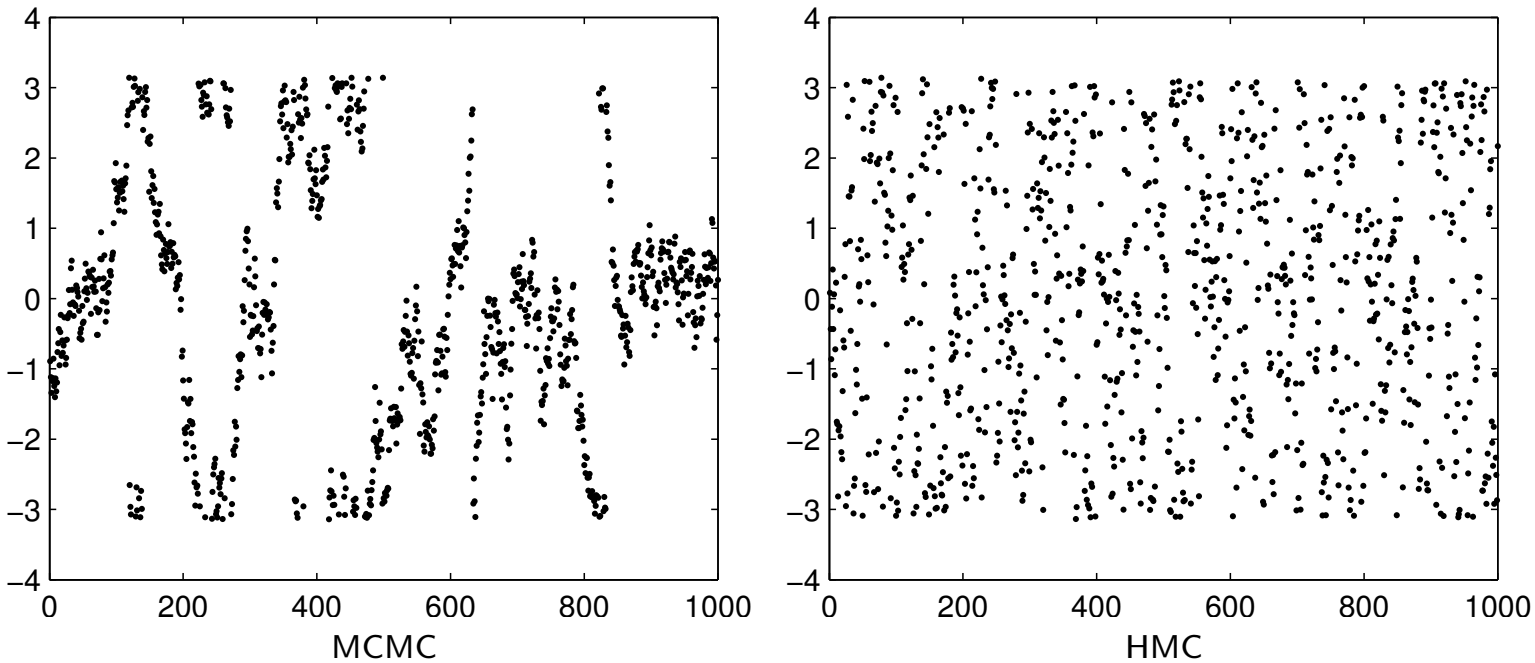
With $P(x) = \exp(-\sin(x)^2/2)$

Taking $P(x) = \exp(-\sin(x)^2/2)$, neither the normalizing constant nor the CDF can be computed in closed form.





With $p(x) = \exp(-\sin(x)^2/2)$



Conclusion

- MCMC and HMC allows to sample for a distribution for which we know only the density function up to a normalization constant.
- HMC will usually explore the space better but it requires gradients of the density function.
- This is because MCMC is similar to a random walk while HMC explore the space in a consistent direction, randomly resampled every now and then, while leaving the target distribution invariant.
- Obtaining the same properties with MCMC requires going from N iterations to N^2 between samples.