Hamiltonian Markov chain Monte Carlo

Jo Eidsvik

Mathematical Sciences, NTNU

Markov chain Monte Carlo

MCMC uses dependent samples to do Monte Carlo approximations.
Idea: Construct a Markov chain that converges to the target distribution π (density or probability mass function). Then take sample averages.
The first (transient) part (called burn-in) is discarded in the sample averages because it is would be biased from the initial state.
There are two main algorithms.

- **Gibbs sampling.**
- **Metropolis-Hastings sampling.**

Metropolis-Hastings algorithm

(Assuming discrete state-space S, but works generally.)

- Initialize x^0.
- Iterate the following $b = 1, \ldots, B$:
  1. Sample $U \sim U(0, 1)$
  2. Propose a new potential state x^*, from proposal distribution $P(x^* = j | x^{b-1} = i) = Q_{ij}$, $j \in S$.
  3. Compute the acceptance probability: $\alpha_{ij} = \min\left(1, \frac{\pi_j}{\pi_i}\frac{Q_{ji}}{Q_{ij}}\right)$
  4. Set $x^b = x^* = j$ if $U < \alpha_{ij}$. Else set $x^b = i$.

We are free to choose Q_{ij}. This gives enormous flexibility! The art is to select this wisely!

Random walk MH

Target density $\pi(\boldsymbol{x})$.
MH algorithm

- Initiate $\boldsymbol{x}^0 = (x_1^0, x_2^0)$, $b = 0$.
- Iterate while $b < B$,
  1. Sample $\boldsymbol{x}^* \sim N(\boldsymbol{x}^b, \sigma^2 \boldsymbol{I})$
  2. Accept $\boldsymbol{x}^{b+1} = \boldsymbol{x}^*$ with probability $\alpha = \min\left\{1, \frac{\pi(\boldsymbol{x}^*)}{\pi(\boldsymbol{x}^b)}\right\}$
     Else set $\boldsymbol{x}^{b+1} = \boldsymbol{x}^b$.
  3. $b = b + 1$

The symmetric proposal density cancels in the acceptance rate. Very easy to implement. Asymptotically optimal (under some assumptions) to have acceptance rate about 0.23.

Langevin MH

Langevin: $q(\boldsymbol{x}|\boldsymbol{x}^b) = N(\boldsymbol{x}^b + \frac{\sigma^2}{2}\nabla \log \pi(\boldsymbol{x}^b), \sigma^2 \boldsymbol{I})$.
Langevin MH algorithm

- Initiate $\boldsymbol{x}^0 = (x_1^0, x_2^0)$, $b = 0$.
- Iterate while $b < B$,
  1. Sample $\boldsymbol{x}^* \sim q(\boldsymbol{x}|\boldsymbol{x}^b)$
  2. Accept $\boldsymbol{x}^{b+1} = \boldsymbol{x}^*$ with probability $\alpha = \min\left\{1, \frac{\pi(\boldsymbol{x}^*)q(\boldsymbol{x}^b|\boldsymbol{x}^*)}{\pi(\boldsymbol{x}^b)q(\boldsymbol{x}^*|\boldsymbol{x}^b)}\right\}$

     Else set $\boldsymbol{x}^{b+1} = \boldsymbol{x}^b$.
  3. $b = b + 1$

Only requires tuning of σ. Asymptotically optimal (under some assumptions) to have acceptance rate about 0.57.

Mixing of the MCMC sampler

MCMC gives *dependent* samplers from the density or probability mass function π.
If the dependence (autocorrelation) is very large, the Markov chain is said to mix slowly, and it takes a very long time to explore the sample space adequately.

Autocorrelation in the MCMC sampler

Integrated autocorrelation:

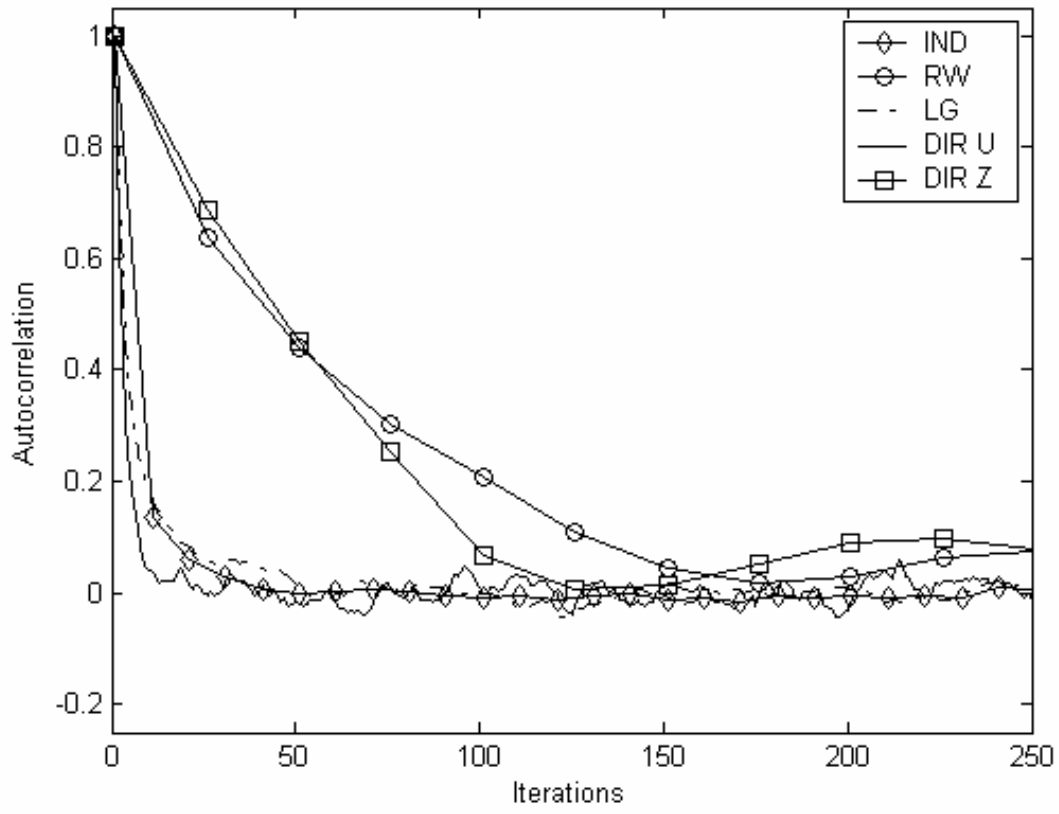
$$\text{IAC} = 1 + 2 \sum_{t=1}^{2T+1} \hat{\rho}_t, \quad \hat{\rho}_t = \hat{\text{Corr}}(x_s, x_{s+t}),$$

$$T = \max(\tau; \hat{\rho}_{2t} + \hat{\rho}_{2t+1} > 0 \text{ for all } t \leq \tau)$$

Effective sample size, N is number of dependent samples in chain:

$$\text{ESS} = \frac{N}{IAC}$$

Mixing vs evaluation of the MCMC sampler



Note: The number of evaluations (M) increases with the complexity of the target density per iteration (derivatives? or second derivatives?).

Looking for smart proposals that explore the sample space quickly

- Auxilliary variables.
- Use of derivatives of the target distribution.
- Hamiltonian MH.

Auxiliary variables in MH - hit and run

One iteration of hit-and-run.

- Draw a direction.
- Sample length along direction. (This defines proposal.)
- Accept or reject move with MH rate.

(Could be better than Gibbs sampling if directions are sampled wisely, and reasonable accept probability, which depends on the chance of getting back from $\boldsymbol{x}^*$ to $\boldsymbol{x}$.)

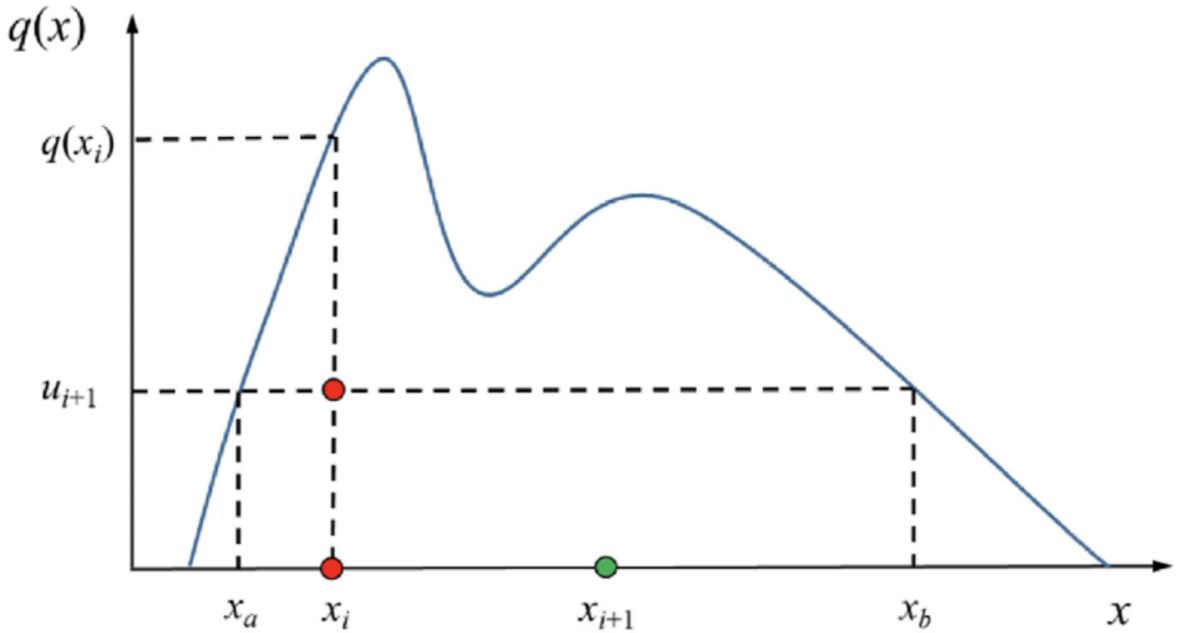
Auxiliary variables in MH - slice sampling

One iteration of slice sampling.

- Draw a level at the density axis $u \sim U(0, \pi(\boldsymbol{x}^b))$.
- Sample a variable that has a least this density level, $\boldsymbol{x} \sim U(\boldsymbol{x} : \pi(\boldsymbol{x}^{b+1}) > u)$.

(Can induce large steps in sample space, much larger than Gibbs sampling, if possible to sample at a level of the density.

Illustration slice sampling



Uniform sampling - slice sampling

$$\pi(\boldsymbol{x}, u) = U[(u, \boldsymbol{x}); u < \pi(\boldsymbol{x}), \boldsymbol{x} \in S_x]$$

$$\pi(\boldsymbol{x}) = \int_0^{\pi(\boldsymbol{x})} 1 du = \pi(\boldsymbol{x})$$

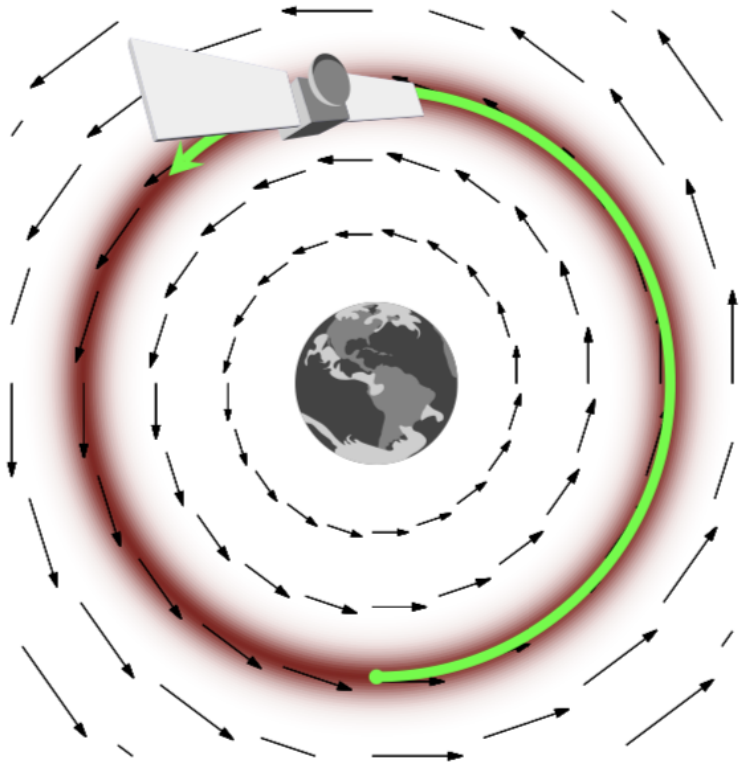
(Can also show invariance of resulting Markov chain.)

Hamiltonian MH

Combines the following:

- Uses an auxiliary variable (momentum) to improve mixing.
- Uses derivatives (Hamilton's differential equation).
- Naturally inspired from physics (energy).

Movement for Hamiltonian MCMC



Goal is to explore the key parts of the distribution effectively!
Large moves, efficient mixing of Markov chain.

Hamiltonian MH

$\boldsymbol{z}$ is momentum.
Hamiltonian function

$$H(\boldsymbol{x}, \boldsymbol{z}) = V(\boldsymbol{x}) + K(\boldsymbol{x}, \boldsymbol{z})$$

Joint distribution: $\pi(\boldsymbol{x}, \boldsymbol{z}) = \exp(-H(\boldsymbol{x}, \boldsymbol{z}))$.
$\pi(\boldsymbol{z}|\boldsymbol{x}) \propto \exp(-K(\boldsymbol{x}, \boldsymbol{z}))$ is kinetic energy, $V = -\log \pi(\boldsymbol{x})$ is potential energy.
One can choose $K(\boldsymbol{x}, \boldsymbol{z})$, usually Gaussian;

$$K(\boldsymbol{x}, \boldsymbol{z}) = \log |M(\boldsymbol{x})| + \frac{1}{2} \boldsymbol{z}' M(\boldsymbol{x})^{-1} \boldsymbol{z}$$

(Independent of $\boldsymbol{x}$ if $M(\boldsymbol{x}) = M$.)
This is symmetric for $\boldsymbol{z}$ and $-\boldsymbol{z}$.

Hamiltonian Diff eq

Langevin equation is a Stochastic differential equation:

$$\frac{d\boldsymbol{x}}{dt} = \nabla \log \pi(\boldsymbol{x}) + \sqrt{2}B_t$$

(Solution to this is $\boldsymbol{x} \sim \pi(\boldsymbol{x})$.)
Because exact solutions are not possible, one uses a MH sampler and accept-reject random proposals.

Hamilton's equation defines a deterministic differential equation:

$$\frac{d\boldsymbol{x}}{dt} = \frac{dH}{d\boldsymbol{z}} = \frac{dK}{d\boldsymbol{z}}, \quad \frac{d\boldsymbol{z}}{dt} = -\frac{dH}{d\boldsymbol{x}} = -\frac{dK}{d\boldsymbol{x}} - \frac{dV}{d\boldsymbol{x}}$$

With solution $\boldsymbol{x} \sim \pi(\boldsymbol{x})$. (marginally, when ignoring the auxiliary momentum variables.)

Hamiltonian exact

$$x \sim N(0,1), \quad z|x \sim N(0,1).$$

$$V(x) = x^2/2, \quad K(z|x) = z^2/2,$$

$$\frac{dx}{dt} = z, \quad \frac{dz}{dt} = -x$$

$$x(t) = r\cos(a+t), \quad z(t) = -r\sin(a+t)$$

For most systems Hamilton's equations cannot be solved like this. Instead discrete-time numerical integrators are used.

Leapfrog proposal MH

Leapfrog is the most popular method to propagate Hamiltonian dynamics to a proposal. (In the class of symplectic integrators)

Start by $\boldsymbol{x} = \boldsymbol{x}_0$, $\boldsymbol{z} = \boldsymbol{z}_0$.
For $i = 0, 1, \ldots, T/\epsilon$,

- $\boldsymbol{z}_{i+1/2} = \boldsymbol{z}_i - \frac{\epsilon}{2} \frac{dV}{d\boldsymbol{x}}(\boldsymbol{x}_i)$
- $\boldsymbol{x}_{i+1} = \boldsymbol{x}_i + \epsilon \boldsymbol{z}_{i+1/2}$
- $\boldsymbol{z}_{i+1} = \boldsymbol{z}_{i+1/2} - \frac{\epsilon}{2} \frac{dV}{d\boldsymbol{x}}(\boldsymbol{x}_{i+1})$

Tuning parameters are ϵ and T.
(There are adaptive heuristic methods for setting T and ϵ; NUTS (No-Uturn sampling.))

Illustration Hamiltonian eq

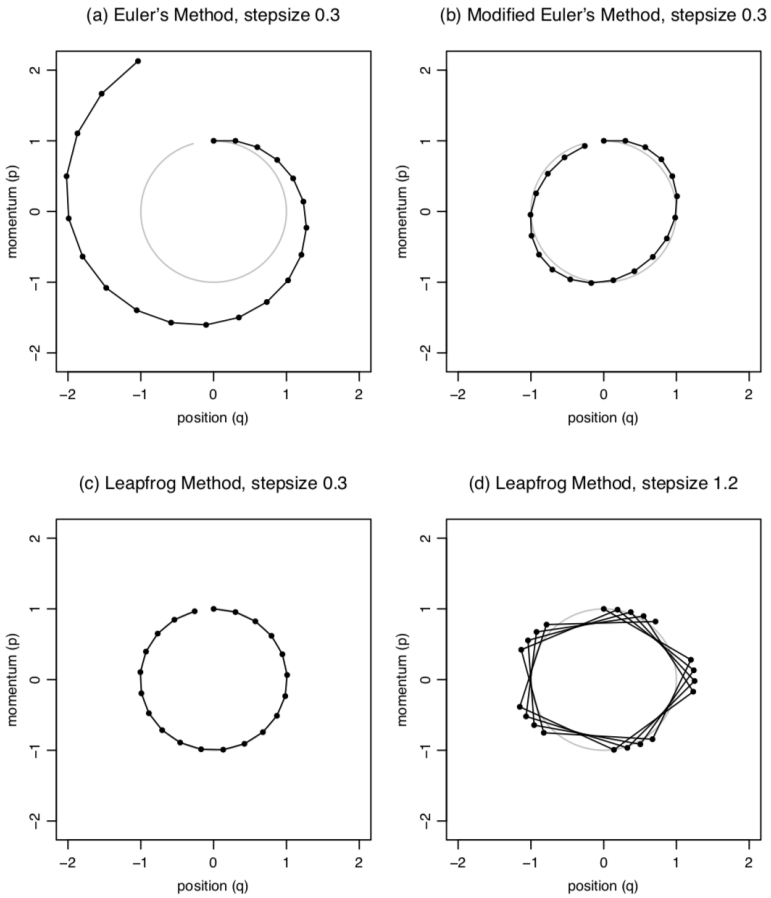


Figure 1: Results using three methods for approximating Hamiltonian dynamics, when $H(q,p) = q^2/2 + p^2/2$. The initial state was $q = 0$, $p = 1$. The stepsize was $\varepsilon = 0.3$ for (a), (b), and (c), and $\varepsilon = 1.2$ for (d). Twenty steps of the simulated trajectory are shown for each method, along with the true trajectory (in gray).

(From Neal, 2013)

Illustration path : $N_2(0, \mathbf{\Sigma})$

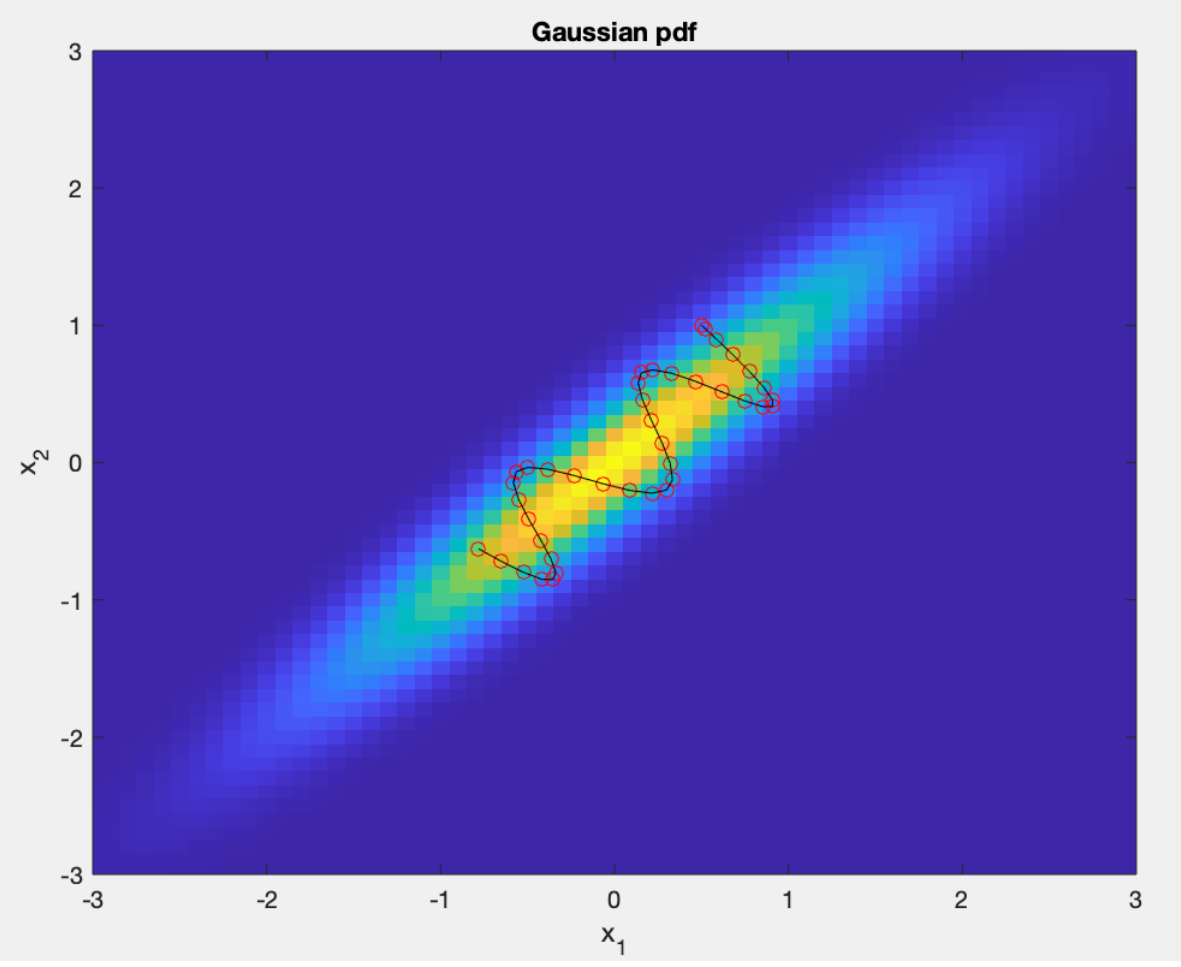


Illustration of Hamiltonian

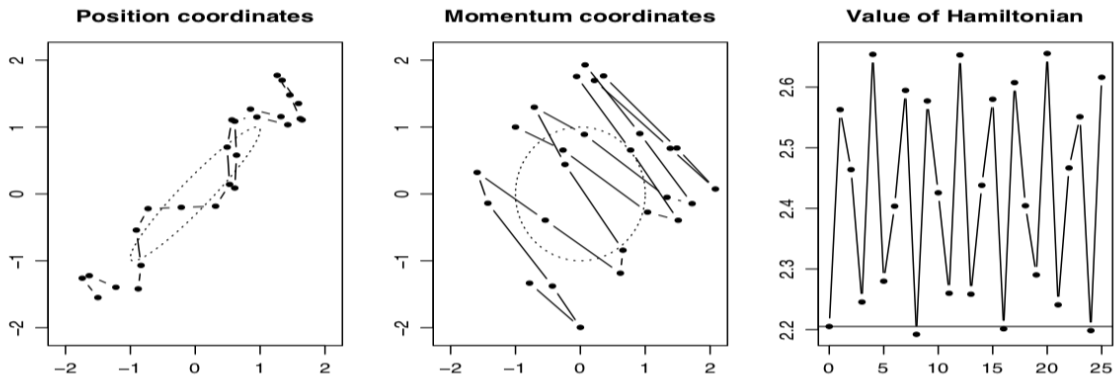


Figure 3: A trajectory for a 2D Gaussian distribution, simulated using 25 leapfrog steps with a stepsize of 0.25. The ellipses plotted are one standard deviation from the means. The initial state had $q = [-1.50, -1.55]^T$ and $p = [-1, 1]^T$.

(Neal 2013)
Hamiltonian does not change during leapfrog steps. (But state and momentum changes.)

Mathematics of path for : $N_2(0, \boldsymbol{\Sigma})$

Start by $\boldsymbol{x} = \boldsymbol{x}_0$, $\boldsymbol{z} = \boldsymbol{z}_0$.
For $i = 0, 1, \ldots, T/\epsilon$,

- $\boldsymbol{z}_{i+1/2} = \boldsymbol{z}_i - \frac{\epsilon}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_i$
- $\boldsymbol{x}_{i+1} = \boldsymbol{x}_i + \epsilon \boldsymbol{z}_{i+1/2}$
- $\boldsymbol{z}_{i+1} = \boldsymbol{z}_{i+1/2} - \frac{\epsilon}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_i$

Hamiltonian MH algorithm

Start by $\boldsymbol{x} = \boldsymbol{x}_0$, $\boldsymbol{z} = \boldsymbol{z}_0$, $b = 0$,
Iterate the following:

- Set $\boldsymbol{x} = \boldsymbol{x}^b$.
- Propose $\boldsymbol{z}|\boldsymbol{x} \sim \exp(-K(\boldsymbol{x}, \boldsymbol{z}))$.
- Run leapfrog Hamiltonian dynamics for T/ϵ steps to get proposal $(\boldsymbol{x}^*, \boldsymbol{z}^*)$.
- Calculate acceptance probability $\alpha = \min\left[1, \frac{\pi(\boldsymbol{x}^*)}{\pi(\boldsymbol{x}^b)}\right]$
- Accept $\boldsymbol{x}^{b+1} = \boldsymbol{x}^*$ if $U < \alpha$, else set $\boldsymbol{x}^{b+1} = \boldsymbol{x}^b$.

(Proposal cancels because of symmetry in leapfrog dynamics and $\boldsymbol{z} = -\boldsymbol{z}$.)

Illustration MH proposal path

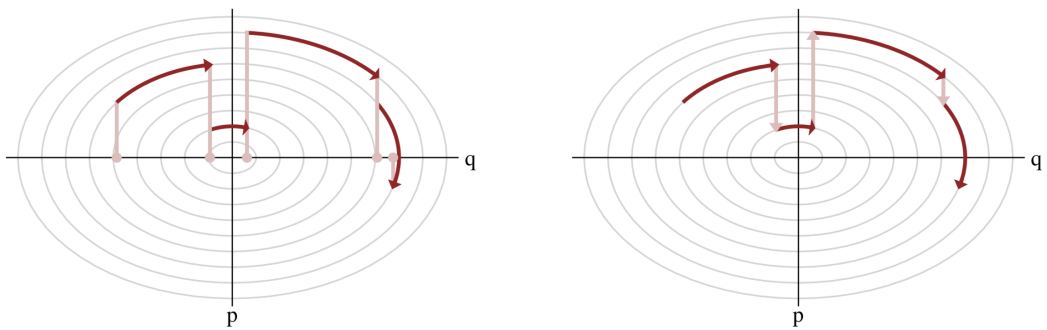
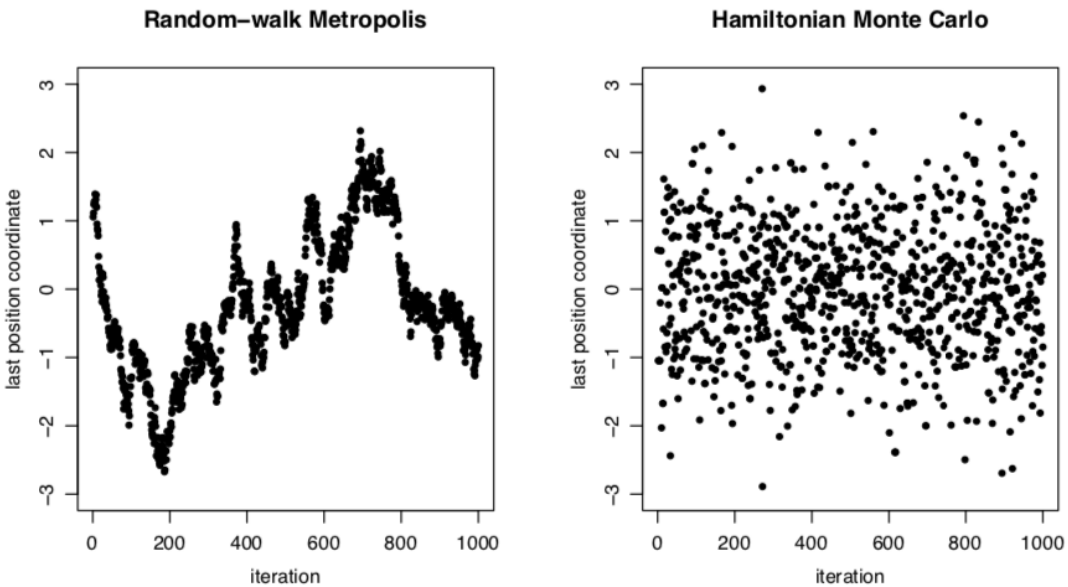


FIG 22. *(a) Each Hamiltonian Markov transition lifts the initial state onto a random level set of the Hamiltonian, which can then be explored with a Hamiltonian trajectory before projecting back down to the target parameter space. (b) If we consider the projection and random lift steps as a single momentum resampling step, then the Hamiltonian Markov chain alternates between deterministic trajectories along these level sets (dark red) and a random walk across the level sets (light red).*

(From Betancourt, 2013)

Illustration mixing



(From Neal, 2013)

Pros and cons of Hamiltonian MC

- Makes large jumps in the target density. Good mixing.
- Natural proposal - motivated by Diff Eq, geometry and physics
- Code exists (NUTS, Stan) for efficient implementation.
- Depends on time required to get derivatives.
- Not obvious how to tune (improve) proposal for momentum $\boldsymbol{z}$.
- Still struggles with difficult targets (multimodal densities, ridge densities).

Effective sample size is usually large for Hamiltonian MH.

$$\text{ESS} = \frac{N}{IAC}$$

(Could scale this with evaluation cost per iteration, which is larger when derivatives are required.)