Computational Methods in Stochastics

Lecture VI

Hamiltonian Monte Carlo

(HMC)

The Basic HMC Scheme

Hamiltonian dynamics (from mechanics) is used in combination with Metropolis sampling to construct an MCMC method.

The motivation for HMC is to sample the state space more efficiently, so that larger movements from the current state could be made in one step than what is possible in Metropolis-Hastings (M-H) sampling. This comes at the price of increased computation per time step. Despite this, HMC typically samples states much faster than M-H.

When reading this,

1. first have a look at this site.
2. Refer to it when reading.
3. After having stumbled this through, read it through and play around with the graphical models.

The Basic HMC Scheme

Hamiltonian function $H(q,p)$ is determined in terms of the probability distribution we want to sample from.

The **"position" variables** q are the ones we are interested in.

The **"momentum" variables** p are auxiliary that we need in order to move within the distribution and so to do the sampling. These momentum – or velocity – variables provide the more efficient sampling.

Simple updates of these variables alternate with Metropolis updates.

Gain: Proposed states can be distant from the current states and still have a high probability of acceptance.

Hamiltonian Dynamics

Hamiltonian (function) gives the total - here constant - **energy** of a dynamical system.

$$H(q,p) = U(q) + K(p) = E_{tot},$$

where $U(q)$ is the potential energy and $K(p)$ is the kinetic energy.

The dynamical system is completely described when we know how q and p change over time t. This is stated by equations of motion, here **Hamilton's equations**

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i},$$

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}.$$

$i = 1, \ldots, d$; (d is dimension)

Hamiltonian Dynamics

Combining vectors q and p into the vector $z = (q, p)$ we get

$$\frac{dz}{dt} = J\nabla H(z),$$

where the gradient $\nabla H = [\nabla H]_k = \partial H / \partial z_k$, and

$$J = \begin{bmatrix} 0_{d\times d} & I_{d\times d} \\ -I_{d\times d} & 0_{d\times d} \end{bmatrix}.$$

Potential and Kinetic Energy

In d dimensions, the kinetic energy

$$K(p) = \frac{p^T M^{-1} p}{2}.$$

M is a symmetric, positive-definite mass matrix, often iagonal, and a scalar multiple of the identity matrix.

Hamiltonian Dynamics

Potential and Kinetic Energy

In d dimensions, the kinetic energy

$$K(p) = \frac{p^T M^{-1} p}{2}.$$

M is a symmetric, positive-definite mass matrix, often diagonal, and a scalar multiple of the identity matrix.

Note that $p = Mv$, where $v = \dot{q}$, so

$$K(v) = \frac{v^T M v}{2}.$$

Hamiltonian Dynamics

This form of $K(p)$ corresponds to $-\log\big(p_P(p)\big) + \text{Const.}$ Here, $p_P(p) = N(0, \Sigma = M)$ (zero-mean Gaussian with covariance matrix M).

$$p_P(p) = \frac{1}{Z}\exp\left(\frac{-K(p)}{T}\right) = \frac{1}{Z}\exp\left(\frac{p^T M^{-1} p}{2}\right). \quad (T = 1\,.)$$

With these forms, Hamilton's equations become

$$\frac{dq_i}{dt} = [M^{-1}p]_i, \qquad \text{(Velocity.)}$$

$$\frac{dp_i}{dt} = -\frac{\partial U}{\partial q_i}. \qquad \text{(Force.)}$$

Hamiltonian Dynamics

Example: One-dimensional harmonic oscillator

$$H(q,p) = U(q) + K(p), \qquad U(q) = \frac{q^2}{2}, K(p) = \frac{p^2}{2m} = \frac{p^2}{2}.$$

Here, we choose $m = 1$.

So, distributions for both q and p are $N(0,1)$.

The dynamics:

$$\frac{dq_i}{dt} = p,$$

$$\frac{dp_i}{dt} = -q.$$

(Hamilton's equations)

Solution (a and r are constants determined by initial conditions):

$$q(t) = r\cos(a+t), \qquad p(t) = -r\sin(a+t)$$

Hamiltonian Dynamics

Properties of Hamiltonian dynamics

Reversibility

The mapping from the state at time t, $(q(t), p(t))$, to the state at time $t + s$, $(q(t + s), p(t + s))$, is one-to-one and so has an inverse, which is obtained by negating the time derivatives in Hamilton's equations.

Conservation of the Hamiltonian (= Conservation of Energy)

The dynamics keeps the Hamiltonian invariant.

$$\frac{dH}{dt} = \sum_{i=1}^{d} \left[\frac{\partial H}{\partial q_i} \frac{dq_i}{dt} + \frac{\partial H}{\partial p_i} \frac{dp_i}{dt} \right] = \sum_{i=1}^{d} \left[\frac{\partial H}{\partial q_i} \frac{\partial H}{\partial p_i} - \frac{\partial H}{\partial p_i} \frac{\partial H}{\partial q_i} \right] = 0$$

H is a *constant of motion*.

Hamiltonian Dynamics

Volume Preservation

Any **phase-space volume** evolving in time may change its shape but the volume does not change

$$\nabla \cdot \left(\frac{d\boldsymbol{q}}{dt}, \frac{d\boldsymbol{p}}{dt}\right) = \sum_{i=1}^{d} \left[\frac{\partial}{\partial q_i}\frac{dq_i}{dt} + \frac{\partial}{\partial p_i}\frac{dp_i}{dt}\right] = \sum_{i=1}^{d} \left[\frac{\partial}{\partial q_i}\frac{\partial H}{\partial p_i} - \frac{\partial}{\partial p_i}\frac{\partial H}{\partial q_i}\right] =$$

$$= \sum_{i=1}^{d} \left[\frac{\partial^2 H}{\partial q_i \partial p_i} - \frac{\partial^2 H}{\partial p_i \partial q_i}\right] = 0.$$

(A vector field with zero divergence preserves volume.)

In the present context this means that **probability measure** is invariant in time. (My wording, so subject to misinterpretations ☺)

Hamiltonian Dynamics

Symplectictness

The **volume preservation** is the most important consequence of this more universal property. Hamiltonian dynamics is symplectic. In dynamics symplectictness and volume preservation are often treated as synonyms. Symplectictness can be defined via the Jacobian of the transformation defining the propagation in time (dynamics).

In Hamiltonian dynamics the **symplectic form** is defined as $\omega = dq \wedge dp$. Hamiltonian flow keeps $\int_S \omega$ invariant. (See e.g. **differentiable manifolds**, if you are interested.)

($\wedge$ is the exterior (vector) product ($\times$).)

Hamiltonian Dynamics

Discretization of Hamilton's Equations

Euler's Method

$$p_i(t+\varepsilon) = p_i(t) + \varepsilon \frac{dp_i}{dt}(t) = p_i(t) - \varepsilon \frac{\partial U}{\partial q_i}\big(q(t)\big),$$

$$q_i(t+\varepsilon) = q_i(t) + \varepsilon \frac{dq_i}{dt}(t) = q_i(t) + \varepsilon \frac{\partial K}{\partial p_i}(p(t))$$

We use $K(p) = \sum_{i=1}^{d} \frac{p_i^2}{2m_i} \Rightarrow \; q_i(t+\varepsilon) = q_i(t) + \varepsilon \frac{p_i(t)}{m_i}$.

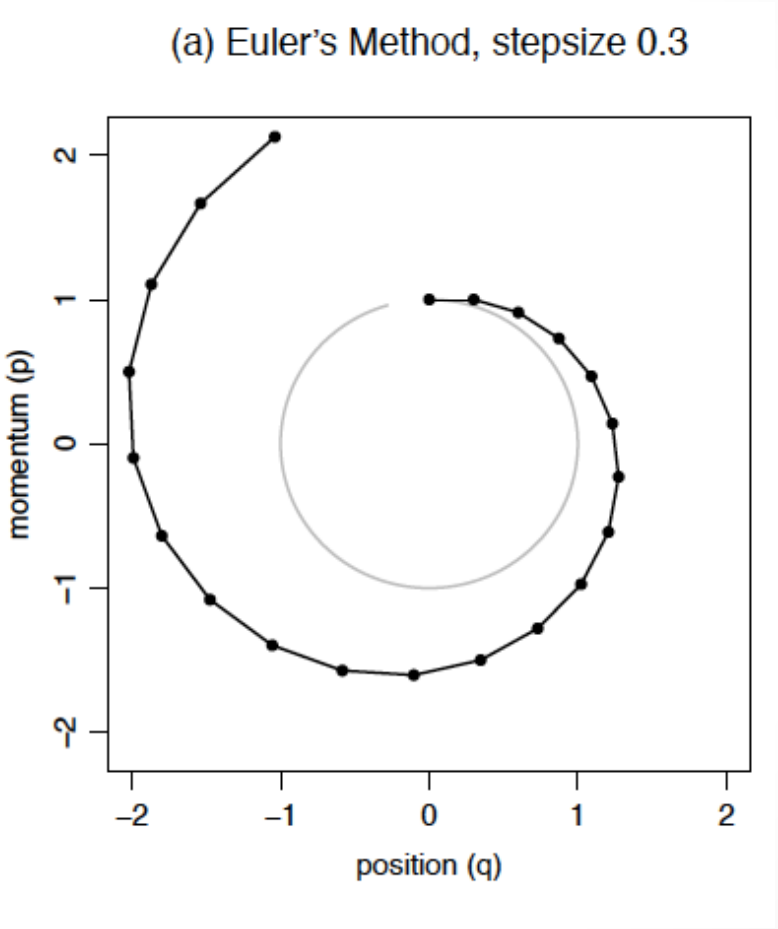
(ε is the step size)

The following examples of numerical integration is for the 1-D harmonic oscillator →

Hamiltonian Dynamics

Discretization of Hamilton's Equations

Euler's Method



Integrating in time using Euler involves **numerical error**; try to minimise it →

Hamiltonian Dynamics

A Modified Euler's Method

$$p_i(t+\varepsilon) = p_i(t) - \varepsilon \frac{\partial U}{\partial q_i}\big(q(t)\big),$$

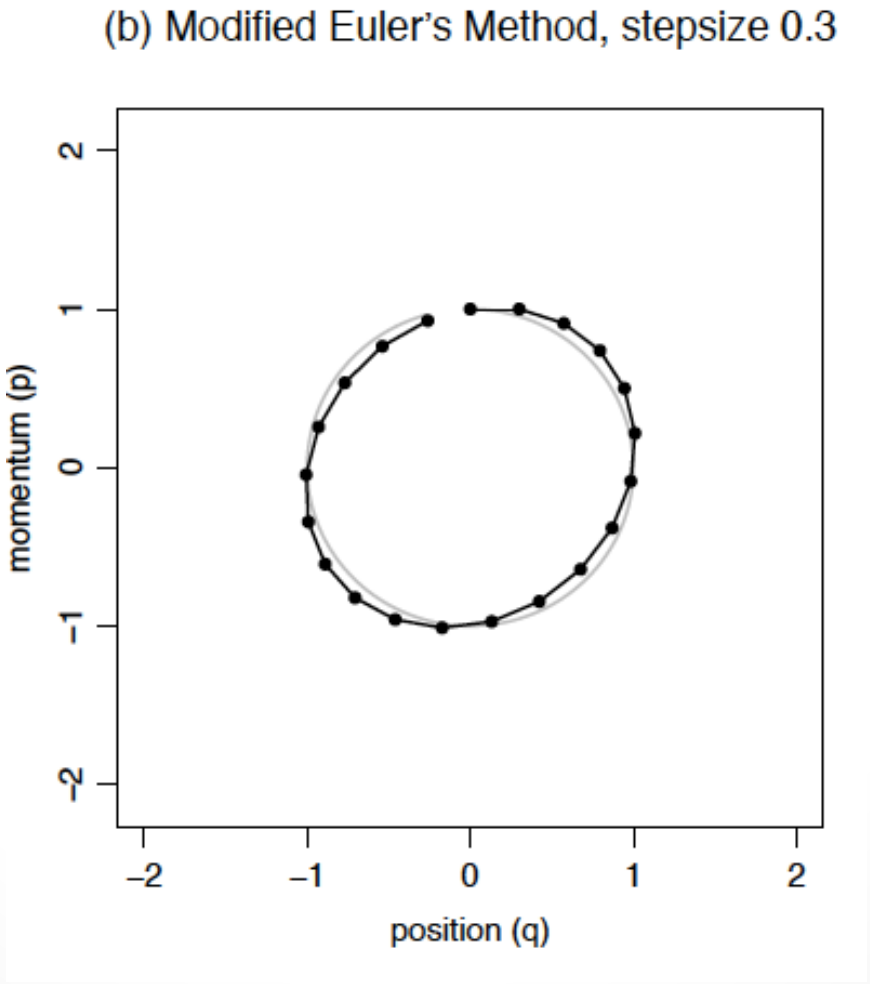
$$q_i(t+\varepsilon) = q_i(t) + \varepsilon \frac{p_i(t+\varepsilon)}{m_i}.$$

The computed trajectory deviates less from the exact trajectory.

→

Hamiltonian Dynamics

A Modified Euler's Method



We can do still better →

Hamiltonian Dynamics

The Leapfrog Method

Propagate in half steps.

$$p_i(t+\varepsilon/2) = p_i(t) - (\varepsilon/2)\frac{\partial U}{\partial q_i}\big(q(t)\big),$$

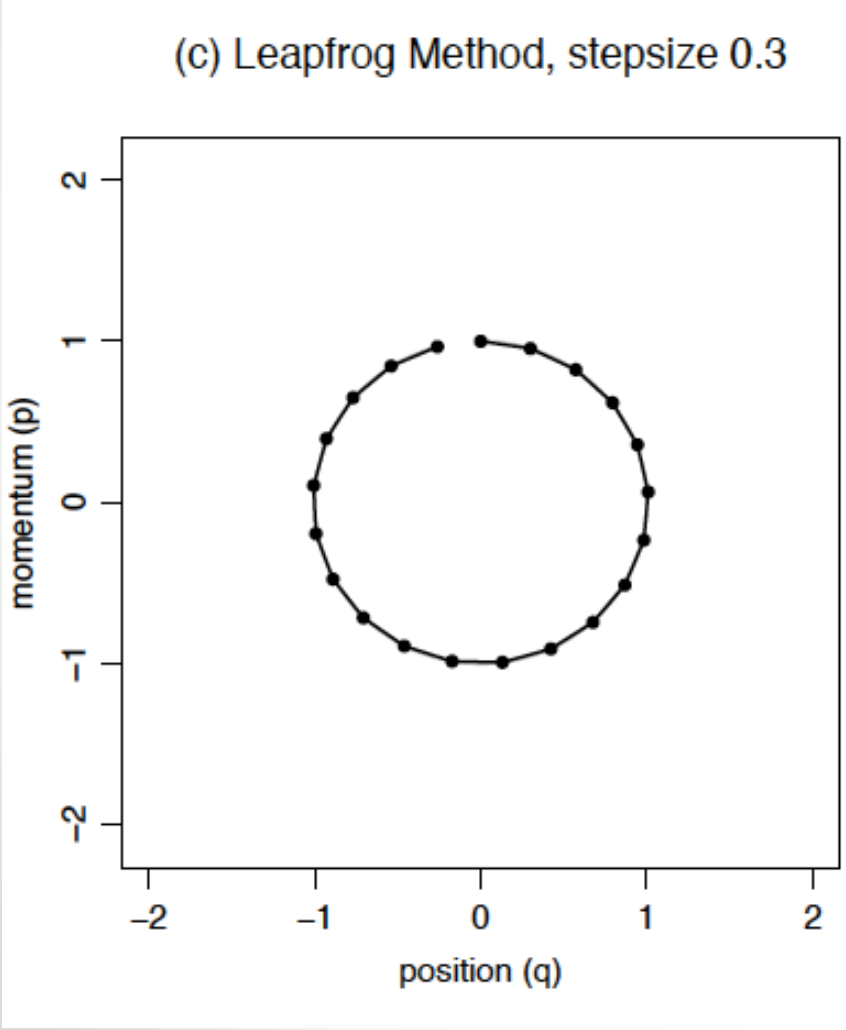
$$q_i(t+\varepsilon) = q_i(t) + \varepsilon\frac{p_i(t+\varepsilon/2)}{m_i}.$$

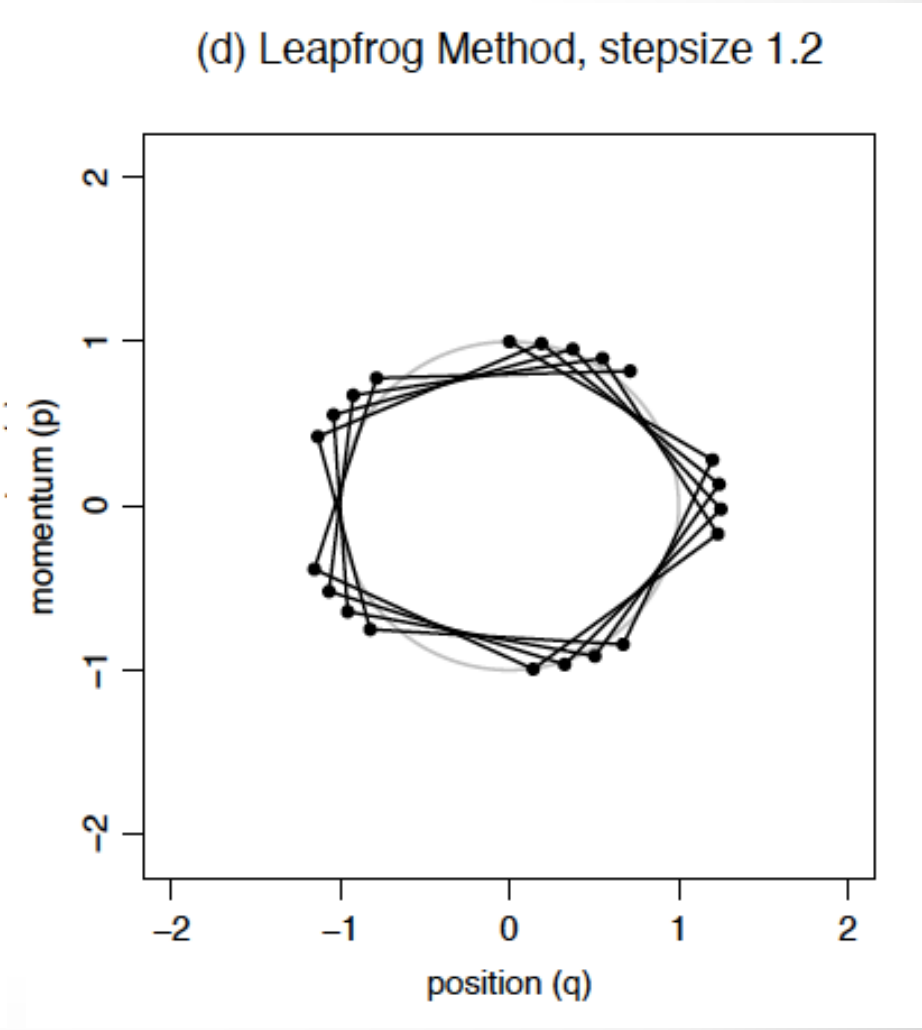
$$p_i(t+\varepsilon) = p_i(t+\varepsilon/2) - (\varepsilon/2)\frac{\partial U}{\partial q_i}\big(q(t+\varepsilon)\big).$$

Hamiltonian Dynamics

The Leapfrog Method

Even with increased time step the computation is stable.





MCMC from Hamiltonian Dynamics

The requirement: Translate **the density function** for the distribution to be sampled from **to a potential energy function** and introduce momentum variables to go with the original variables of interest, now seen as position variables.

The task: Simulate a Markov chain in which each iteration resamples the momentum and then performs a Metropolis update with a proposal found by using Hamiltonian dynamics.

How do we **map the probability distribution to a potential energy function**?

MCMC from Hamiltonian Dynamics

Probability and the Hamiltonian: Canonical Distributions

From statistical physics we know that in a canonical system (temperature and volume are constant) **the probability density function for the state x**, whose energy is $E(x)$, is the **canonical distribution (Gibbs ensemble)**. The probability

$$p_X(x) = \frac{1}{Z}\exp\left(\frac{-E(x)}{kT}\right)$$

Here, T is the temperature and k is the Boltzmann constant, which we just set to 1, since absolute energy values are of no consequence in what we do.

$$\Rightarrow p_X\,(x) = \frac{1}{Z}\exp\left(\frac{-E(x)}{T}\right)$$

MCMC from Hamiltonian Dynamics

Z is the **partition function**. It is the sum over *all* states in the system. It corresponds to the normalisation constant in the distributions we encounter in statistics/stochastics. Incalculable.

In our liberal spirit we also set $T = 1$ and write $E(x) = -\log P(x) - \log Z$ and choose for Z a convenient value.

Since the Hamiltonian $H(q,p) = U(q) + K(p)$ is an energy function for the *joint state* of position q and momentum p, it defines the *joint distribution*

$$p_{q,p}(q,p) = \frac{1}{Z}\exp\left(\frac{-U(q)}{T}\right)\exp\left(\frac{-K(p)}{T}\right)$$

***q* represent the variables of interest and *p* provide the dynamics.**

$H(q,p) = E_{\text{total}} = \text{const.} \Rightarrow p_{q,p}(q,p) = \text{const.}$ when computation is exact.

MCMC from Hamiltonian Dynamics

In ***Bayesian statistics***, the distribution of interest is the posterior distribution for the model. **The posterior distribution** can be expressed as a canonical distribution ($T = 1$) using a potential energy function defined as:

$$U(q) = -\log[\pi(q)L(q|D)]\,,$$

where $\pi(q)$ is the prior density and $L(q|D)$ is the likelihood function given data D. (log = ln.)

So, to **construct the potential function** to go with the distribution, use $\pi(q)L(q|D) = \exp[-U(q)]$.

MCMC from Hamiltonian Dynamics

Betancourt writes the (joint) canonical density as $\pi(q,p) = e^{-H(q,p)}$.

$$\Leftrightarrow H(q,p) = -\log \pi(q,p) = -\log[\pi(p|q)\pi(q)].$$

Accordingly, the decomposition of the Hamiltonian and the joint density correspond as

$$\begin{aligned} H(q,p) &= -\log[\pi(p|q)] - \log[\pi(q)] \\ &\equiv \quad K(q,p) \qquad\quad + \; V(q). \end{aligned}$$

MCMC from Hamiltonian Dynamics

The Hamiltonian Monte Carlo Algorithm

HMC can be used to **sample only from continuous distributions** on $\mathbb{R}^d$ for which

- the density function can be evaluated (up to an unknown normalising constant)
- the partial derivative of the density function (or the gradient of $U(q)$) can be computed: the derivatives must exist except for on a set of points with probability zero, where some arbitrary value can be returned

HMC samples from the canonical distribution $p_{q,p}(q,p)$.
q has the distribution of interest, as specified by $U(q)$.
The distribution of p can be chosen freely via $K(p)$. **Common practise is to use quadratic $K(p)$; consequently p has a zero-mean multivariate Gaussian distribution.**

MCMC from Hamiltonian Dynamics

p_i defined as independent with component i having variance m_i (and setting $T = 1$):

$$K(p) = \sum_{i=1}^{d} \frac{p_i^2}{2m_i}$$

MCMC from Hamiltonian Dynamics

The Two Steps of the HMC Algorithm

The First Step

Draw new values for p_i, independently of the current values of q_i. For $K(p) = \sum_{i=1}^{d} \frac{p_i^2}{2m_i}$ the d variables $p_i \sim N(0, \sigma_i^2)$. This step leaves $p_{q,p}(q,p)$ invariant.

The Second Step ⟶

MCMC from Hamiltonian Dynamics

The Second Step

Propose a new state by performing a Metropolis update with Hamiltonian dynamics:

Start with the current state (q, p).
Simulate Hamiltonian dynamics for L steps using the leapfrog method (or some other reversible volume-preserving method) with a step size ε. (When $L = 1$, HMC is also called L(angevin)MC.)
At the end of L steps, negate p_i ($p_i \to -p_i$).
Now you have **the proposed state** (q^*, p^*).
Accept this proposed state (as the next state of the Markov chain) **with probability** (Metropolis)

$$P = \min\left[1, \exp\left(-H(q^*, p^*) + H(p, q)\right)\right]$$
$$= \min[1, \exp(-U(q^*) + U(q) - K(p^*) + K(p))].$$

MCMC from Hamiltonian Dynamics

If the proposed state is rejected, the next state is the current state. Be sure to count the occurrences of these states also, when computing expectations etc.

The negation of the momentum is done to ensure that the Metropolis proposal is symmetric (Neal).
In fact, there is a more fundamental reason to momentum reversal (Betancourt): If only states going forward can be proposed, i.e. $p^* > p$, the Metropolis-Hastings acceptance probability becomes ill-posed (see Betancourt p. 39).

Return to *The First Step*.

MCMC from Hamiltonian Dynamics

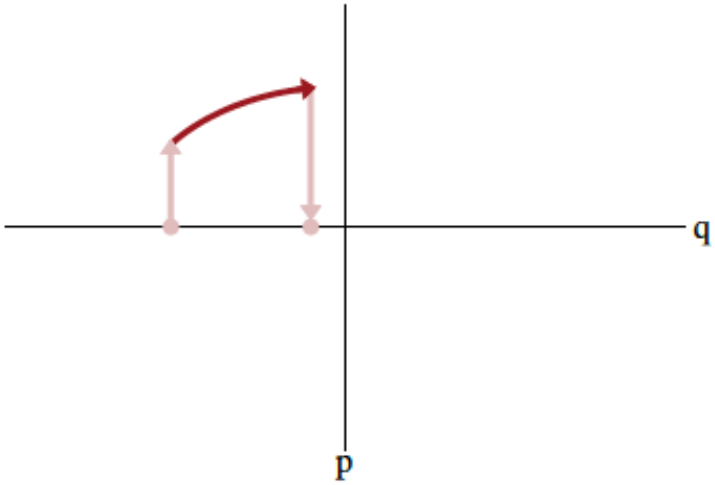
Viewing HMC as sampling from the joint distribution of q and p, the Metropolis step using a proposal found by Hamiltonian dynamics – i.e. the second step - leaves the probability density for (q, p) unchanged; in fact *almost* unchanged due to truncation in the Euler's method and finite numerical precision.

Movement to (q, p) points with a different probability density is accomplished only by the first step, the replacement of p by a new value. This replacement can change the probability density for (q, p) by a large amount.

MCMC from Hamiltonian Dynamics

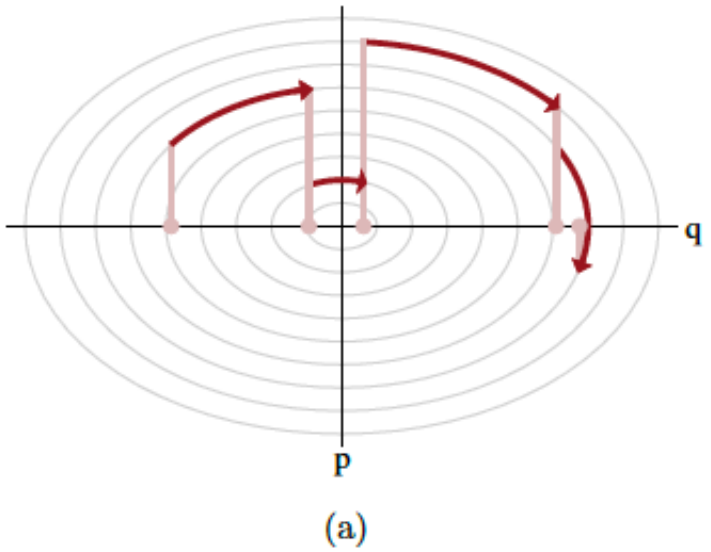
To rephrase, a value for q with a very different probability density and equivalently potential energy $U(q)$ can be produced by Hamiltonian dynamics. Still, resampling of p is necessary for obtaining the proper distribution for q, since without resampling the Hamiltonian $H(q,p) = U(q) + K(p)$ would be (nearly) constant and $U(q)$ could never exceed the initial value of $H(q,p)$.

First Step:
Random lift from the target parameter space onto phase space.



Second Step:
Deterministic Hamiltonian trajectory through phase space and a projection down to the target parameter space.

MCMC from Hamiltonian Dynamics



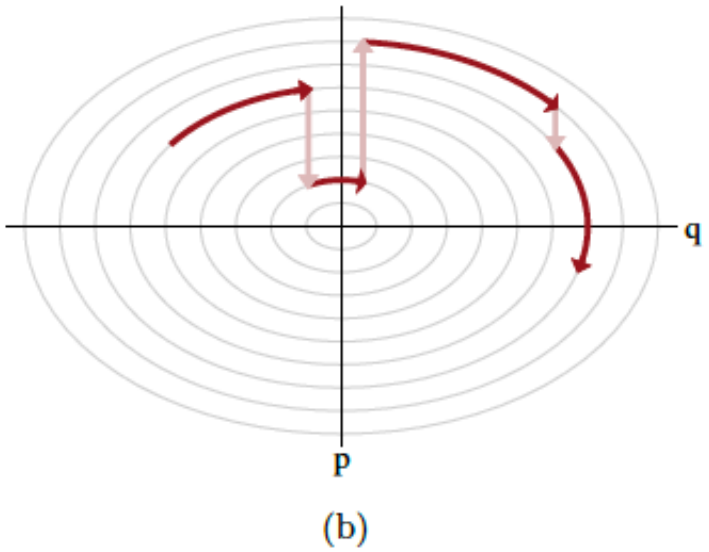


FIG 22. *(a) Each Hamiltonian Markov transition lifts the initial state onto a random level set of the Hamiltonian, which can then be explored with a Hamiltonian trajectory before projecting back down to the target parameter space. (b) If we consider the projection and random lift steps as a single momentum 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).*

Note: Energy level trajectories of this form are for the harmonic oscillator (see page 8).

MCMC from Hamiltonian Dynamics

Proof of the Invariance of the Canonical Distribution

Mentally partition the (q, p) space into regions A_k, each of the same small volume V. Define O as the operation of L leapfrog steps plus a negation of the momenta such that $O: A_k \rightarrow B_k$. (B_k is the image of A_k.)

Leapfrog steps are reversible, so B_k also partition the (q, p) space. Since leapfrog steps and negation preserve volume, each B_k has volume V.

Detailed balance holds if $\forall i, j,\ P(A_i)T(B_j|A_i) = P(B_i)T(A_i|B_j)$.

Here, P is probability under the canonical distribution, and $T(X|Y)$ is the conditional probability of proposing and then accepting a move to region X if the current state is in region Y.

MCMC from Hamiltonian Dynamics

When $i \neq j, T(A_i|B_j) = T(B_j|A_i) = 0$ and detailed balance holds.

When $i = j$:
In the limit as regions A_k and B_k become smaller, the Hamiltonian H_X within each region X becomes effectively constant. → The canonical probability density and the transition probabilities become effectively constant within each region. → The detailed balance condition when $i = j$ reads:

$$\frac{V}{Z}\exp(-H_{A_k})\min[1, \exp(-H_{B_k} + H_{A_k})] =$$

$$\frac{V}{Z}\exp(-H_{B_k})\min[1, \exp(-H_{A_k} + H_{B_k})]$$

This is seen to be true, so **detailed balance holds**.

MCMC from Hamiltonian Dynamics

We know from the stuff before that **if the detailed balance holds, the Markov chain renders the distribution invariant**.

So, **the HMC algorithm leaves the canonical distribution invariant**.

Ergodicity

Typically HMC is ergodic → **all states can be reached**, i.e. no traps. This may be compromised by periodic trajectories in the Leapfrog, when $L\varepsilon \approx 2\pi$.

MCMC from Hamiltonian Dynamics

Benefits of HMC

Consider sampling from a distribution for two variables that is bivariate Gaussian, with means of zero, and correlation 0.95. Regard these as "position" variables.

Introduce two corresponding "momentum" variables, defined to have a Gaussian distribution with means of zero, standard deviations of one, and zero correlation.

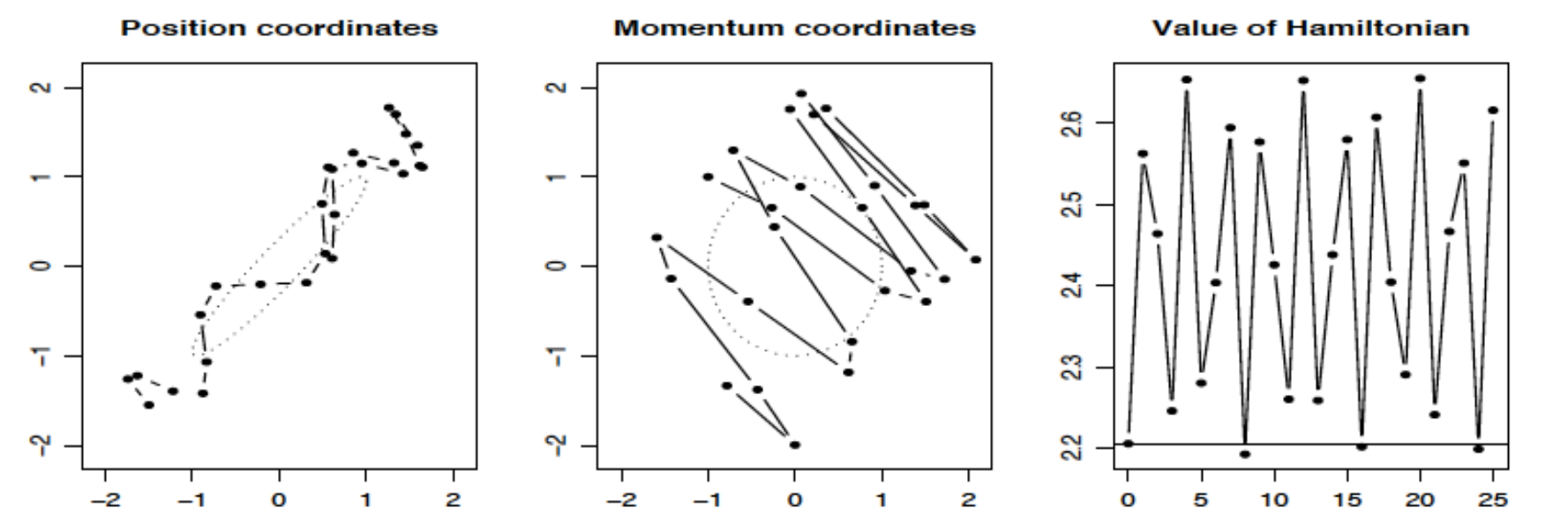
The Hamiltonian will then be

$H(q,p) = q^T \Sigma^{-1} q/2 + p^T p/2$, with covariance $\Sigma = \begin{bmatrix} 1 & 0.95 \\ 0.95 & 1 \end{bmatrix}$.

(In multiple dimensions the inverse Σ^{-1} is computed by e.g. Gauss-Jordan.)

MCMC from Hamiltonian Dynamics

Trajectories of a simulation based on this Hamiltonian. $L = 25$, $\varepsilon = 0.25$. The initial state



q moves from lower left to upper right corned and reverses – nothing like a random walk; efficient sampling.

This comes from the projection of p in diagonal direction changing slowly (gradient in this direction is small) → the direction of p stays the same for many leapfrog steps.

MCMC from Hamiltonian Dynamics

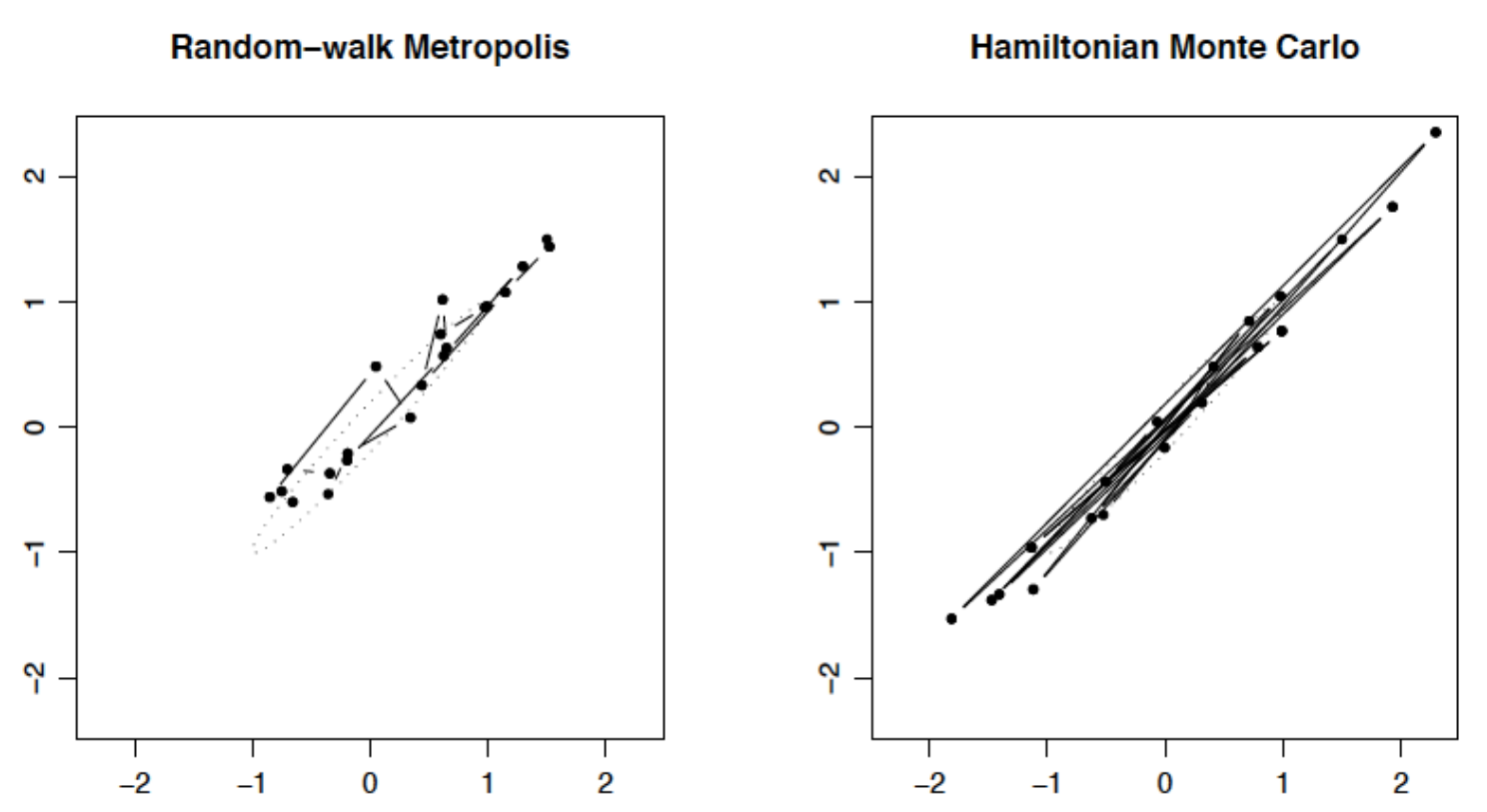
Smaller-scale oscillations result from high correlation between the variables. These oscillations set an upper limit to the step size. For this example, at a critical step size $\varepsilon = 0.45$ the trajectory becomes unstable $\rightarrow$ the value of the Hamiltonian grows without bound.

For this example, at the end of L steps the difference in H is $2.61 - 2.2 = 0.41$, so the probability of accepting the endpoint as the next state is $\exp(-0.41) = 0.66$.

MCMC from Hamiltonian Dynamics

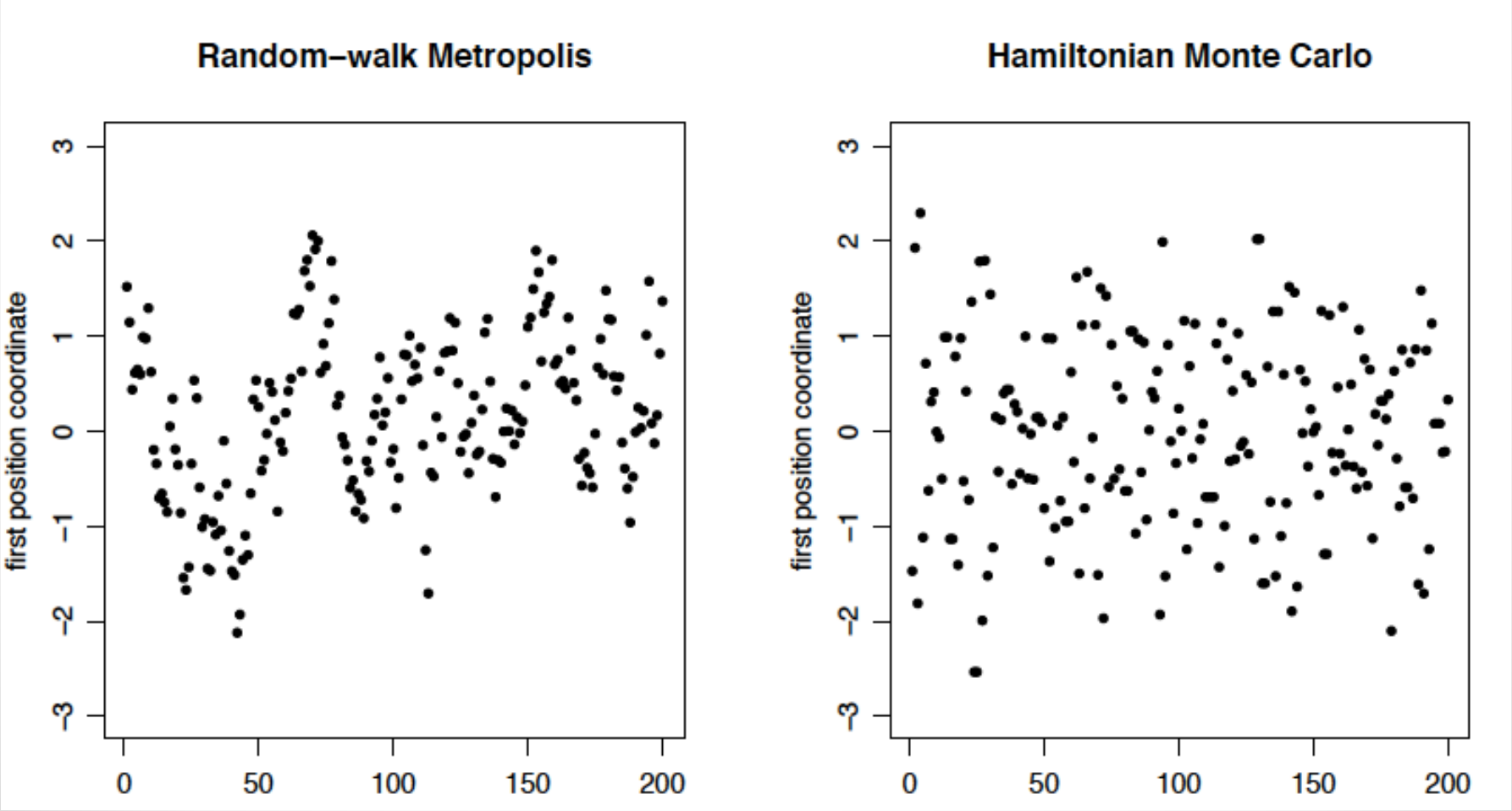
Comparing q's of 2d random-walk and HMC. Correlation is 0.98.

RW of 20 iterations with 20 updates or leapfrog steps per iteration Metropolis and HMC of 20 leapfrog steps per trajectory.



MCMC from Hamiltonian Dynamics

Two hundred iterations, starting with the 20 iterations shown above.



MCMC from Hamiltonian Dynamics

The benefit of HMC stated naively.
RW MCMC: The variance in the position after n iterations of RW MCMC is proportional to $n \rightarrow$ The standard deviation of the amount moved (the distance) in q-space $\propto \sqrt{n}$.
HMC: The distance moved after n will tend to be proportional to n.

The advantage of HMC compared to movement by a random walk will be a factor roughly equal to the ratio of the standard deviations in the least confined direction and most confined direction.

There are ways to enhance HMC like using multiple step sizes.

MCMC from Hamiltonian Dynamics

Pseudo algorithm for *a single iteration* of HMC in 1d (R.M. Neal, p 14):

1. Initiate q. $q^* = q_0$.
2. Sample $p_0 \sim N(0,1)$
3. Make a half step for momentum
   $p^* := p_0 - (\varepsilon/2) \cdot dU(q^*)/dq$
4. Alternate full steps for position and momentum
   for (i := 1, L)
      Make a full step for the position
      $q^* := q^* + \varepsilon \cdot p^*$
      Make a full step for the momentum, except at the end
      If $i \neq L$, $p^* := p^* - \varepsilon \cdot dU(q^*)/dq$
5. Make a half step for momentum at the end
   $p^* := p^* - (\varepsilon/2) \cdot dU(q^*)/dq$
6. Negate momentum at the end of trajectory
   $p^* := -p^*$
7. Evaluate potential and kinetic energies at start and end of trajectory
   $U_0 = U(q_0)$; $K_0 = p_0^2/2$; $U^* = U(q^*)$; $K^* = (p^*)^2/2$
8. Accept or reject the proposed state
   $q^* = q^*$, if $u < \exp(U_0 - U^* + K_0 - K^*)$; $q^* = q_0$, if $u \geq \exp(U_0 - U^* + K_0 - K^*)$
   $(u \sim U(0,1))$